Technische Universität München

Fakultät für Medizin

Improving the diagnostic process of multiple sclerosis by medical image processing and machine learning

Haike Zhang

Vollständiger Abdruck der von der Fakultät für Medizin

der Technischen Universität München zur Erlangung des akademischen Grades

einer Doktorin der Medizin (Dr. med.) genehmigten Dissertation.

Vorsitzender: Prof. Dr. Marcus Makowski

Prüfer:innen der Dissertation:

1. Priv.-Doz. Dr. Benedikt Wiestler

2. apl. Prof. Dr. Mark Mühlau

Die Dissertation wurde am 10.03.2022 bei der Technischen Universität München

eingereicht und durch die Fakultät für Medizin am 07.06.2022 angenommen.

Abstract

The use of magnetic resonance imaging plays a crucial role in the initial diagnosis and monitoring of patients with multiple sclerosis. In recent years, a growing number of computational tools were developed to meet the challenge of image data analysis in general and for this disease. In clinical routine, images from patients with (suspected) multiple sclerosis are visually screened by neuroradiologists for signs of inflammation, which present as T2w-hyperintense white matter lesions. Images are also compared to previous images to evaluate the disease course, looking for the occurrence of new or significantly enlarging lesions as surrogate markers of ongoing disease activity. This process is cumbersome and challenging due to the high lesion variety in terms of size, shape, and location. A tool to relieve this process is desirable. Patients in the possible preliminary stage of multiple sclerosis, the clinically isolated syndrome, frequently convert to multiple sclerosis while details about this pathogenesis are not well understood yet. Identifying parameters that can provide a reliable prediction of future conversion from clinically isolated syndrome allows the selection of patients who benefit from early therapy. We aimed to develop two computational tools to improve the diagnostic process and the image data utility for patients with multiple sclerosis.

In the first project, we wanted to improve the lesion detection in follow-up magnetic resonance imaging of patients with multiple sclerosis regarding accuracy and time. Therefore, we paired up the follow-up scan with the initial scan of 106 patients with multiple sclerosis. We built a computer script that aligns and subtracts the intensity values of these image pairs, creating subtraction maps for the sequences double inversion recovery (DIR) and fluid-attenuated inversion recovery (FLAIR). Two neuroradiologists assessed the existence of new or enlarged lesions for each patient in three different ways: by standard visual comparison, by using FLAIR subtraction maps, and by using DIR subtraction maps. All information from all readouts and all readers was combined to define a reference standard. Using DIR subtraction maps resulted in a higher lesion-detection accuracy than by standard visual comparison (0.96 vs. 0.86, $p = 0.013$) or by using FLAIR maps (0.82, $p < 0.001$). Using DIR maps increased the sensitivity (0.95 vs. 0.82) and provided a better negative predictive value (0.88 vs. 0.67) for detecting new or enlarged lesions in comparison to the standard readout method. Also, significantly more lesions were found in DIR maps (mean 6.26 vs. 3.68), while evaluation time per patient reduced tremendously compared to the standard readout time (median 2 min vs 8 min). Our results suggest that the DIR sequence on its own can provide reliable lesion detection in follow-up images of patients with multiple sclerosis when subtraction maps are used. Our protocol is ready to be used in everyday clinical practice.

The second project aimed to predict the conversion of patients in the clinically isolated syndrome stage to multiple sclerosis by analyzing lesion image features in their initial magnetic resonance imaging scan. For 84 patients with clinically isolated syndrome, lesions in the baseline scan were segmented based on three-dimensional FLAIR and three-

dimensional T1-weighted sequences. For each patient, two sets of brain lesion masks were generated to assess the influence of different segmentation methods on the prediction: one by computer-assisted manual segmentation and one by an automated segmentation algorithm. The intensity and shape parameters of the lesions were calculated from these masks and functioned as input for a random forest model. Oblique random forest models were trained with three different inputs: shape features, intensity features and including features from both categories. Prediction accuracies were validated through three-fold cross-validation. Conversion to multiple sclerosis for the patients in our cohort was defined according to the 2010 McDonald criteria at the time of follow-up of three years. The model based on shape features acquired from the manual segmentation showed the best prediction accuracy and outperformed the gold standard based on dissemination in space (0.85 vs. 0.79, $p = 0.03$). Shape parameters played a major role in a promising prediction, while intensity parameters could not improve prediction performance (accuracy 0.85 vs. 0.62). Especially those shape features that describe the ovality of the lesions, contributed the most to the prediction: *mean lesion volume*, *minimal sphericity,* and *minimal surface-volume ratio*.

We developed two computational tools to improve the radiological diagnostic process of multiple sclerosis and proved their advantage over the present clinical workflow. New and enlarged lesions in follow-up magnetic resonance imaging examinations were recognized faster and more reliably using DIR subtraction maps than via standard procedure. We showed that our random forest model relying on lesion-shape features in the initial magnetic resonance imaging examination can predict the conversion from clinically isolated syndrome to multiple sclerosis more accurately than the gold standard.

Zusammenfassung

Die Magnetresonanztomographie spielt eine entscheidende Rolle bei der Erstdiagnose und Verlaufskontrolle von Patient:innen mit Multipler Sklerose. In den letzten Jahren wurde eine wachsende Zahl von Computerwerkzeugen entwickelt, um die Herausforderung der Bilddatenanalyse im Allgemeinen und für diese Krankheit zu bewältigen. In der klinischen Routine werden Magnetresonanztomographieaufnahmen von Patient:innen mit (vermuteter) Multipler Sklerose von Neuroradiolog:innen visuell auf Anzeichen einer Entzündung untersucht, die sich als T2-gewichtete hyperintense Läsionen der weißen Substanz zeigen. Bilder werden auch mit früheren Aufnahmen verglichen und nach neuen oder deutlich vergrößerten Läsionen als Anzeichen für eine anhaltende Krankheitsaktivität abgesucht. Dieser Vorgang ist aufgrund der großen Vielfalt an Läsionen in Bezug auf ihre Größe, Form und Lage mühsam und schwierig. Patient:innen im möglichen Vorstadium der Multiplen Sklerose, dem klinisch isolierten Syndrom, konvertieren häufig in das Stadium der Multiple Sklerose, wobei die Details dieser Pathogenese noch nicht genau verstanden sind. Die Identifizierung von Parametern, die eine zuverlässige Vorhersage bezüglich einer Konversion ermöglichen, würde die Auswahl von Patient:innen mit klinisch isoliertem Syndrom erlauben, die von einer frühzeitigen Therapie profitieren. Unser Ziel war es, zwei Computerwerkzeuge zu entwickeln, die den Diagnoseprozess und die Nutzung von Bilddaten für Patient:innen mit Multipler Sklerose verbessern.

Das erste Projekt hatte das Ziel das Auffinden von Läsionen in der Magnetresonanztomographie von Patient:innen mit Multipler Sklerose hinsichtlich Genauigkeit und dafür nötigen Zeitaufwand zu verbessern. Zu diesem Zweck wurden die Nachuntersuchung mit der jeweiligen Erstuntersuchung von 106 Patienten mit Multipler Sklerose gepaart. Ein Computerskript wurde entwickelt, das die Intensitätswerte dieser seriellen Magnetresonanztomographie-Untersuchungen zueinander kongruent ausrichtet und subtrahiert, sodass Subtraktionskarten für die Sequenzen Double Inversion Recovery (DIR) und Fluid Attenuated Inversion Recovery (FLAIR) erstellt wurden. Zwei Neuroradiologen suchten auf drei verschiedene Weisen nach neuen oder vergrößerten Läsionen für jede Patient:in: mittels klinischen Standards, dem visuellen Abgleich, mittels FLAIR-Subtraktionskarte und mittels DIR-Subtraktionskarte. Alle Informationen von allen Analysedurchgängen und allen Analysten wurden kombiniert, um einen Referenzstandard zu definieren. Die Verwendung von DIR-Subtraktionskarten führte zu einer höheren Genauigkeit bei der Läsionsdetektion verglichen mit der Standardmethode (0.96 vs. 0.86, $p = 0{,}013$) oder der Verwendung von FLAIR-Karten (0.82, $p < 0{,}001$). Die Verwendung von DIR-Karten erhöhte die Sensitivität (0.95 vs. 0.82) und lieferte einen besseren negativen Vorhersagewert (0.88 vs. 0.67) für die Erkennung neuer oder vergrößerter Läsionen im Vergleich zur Standard-Analysemethode. Außerdem wurden in den DIR-Karten signifikant mehr Läsionen gefunden (Mittelwert 6.26 vs. 3.68), und gleichzeitig verkürzte sich bei dieser Methode die Auswertungszeit pro Patient:in auf ein Drittel der Standardanalysezeit (Median 2 min vs. 8 min). Unsere Ergebnisse deuten darauf hin, dass die DIR-Sequenz allein eine zuverlässige Erkennung

von Läsionen in Folgebildern von MS-Patient:innen ermöglichen kann, wenn Subtraktionskarten verwendet werden. Unser Protokoll kann in der täglichen klinischen Praxis eingesetzt werden.

Das zweite Projekt zielte darauf ab, die Konversion von Patient:innen im Stadium des klinisch isolierten Syndroms in eine Multiple Sklerose vorherzusagen, indem Bildmerkmale der Läsionen im ersten Magnetresonanztomographie-Scan analysiert wurden. Bei 84 Patient:innen mit klinisch isoliertem Syndrom wurden die Läsionen im Ausgangsscan anhand von dreidimensionalen FLAIR- und dreidimensionalen T1-gewichtete Sequenzen segmentiert. Die Läsionen wurden für jede Patient:in durch zwei verschiedene Methoden segmentiert, um den Einfluss verschiedener Segmentierungsmethoden auf die Vorhersage zu bewerten: eine durch computergestützte manuelle Segmentierung und eine durch einen automatischen Segmentierungsalgorithmus. Intensitäts- und Formparameter der Läsionen wurden aus diesen Segmentierungsasken berechnet und dienten als Input für ein Oblique-Random-Forest-Modell. Dieses wurde mit drei verschiedenen Eingaben trainiert: Formmerkmale, Intensitätsmerkmale und Einbeziehung sowohl von Formmerkmalen als auch Intensitätsmerkmalen. Die Vorhersagegenauigkeit wurde anschließend durch eine dreifache Kreuzvalidierung geprüft. Die Konversion zu Multipler Sklerose wurde für die Patient:innen in unserer Kohorte nach den McDonald-Kriterien von 2010 in einer Nachbeobachtungszeit von drei Jahren definiert. Das Computermodell, das auf Formmerkmalen aus der manuellen Segmentierung basiert, zeigte die beste Vorhersagegenauigkeit und übertraf den Goldstandard, der auf der räumlichen Dissemination basiert (0.85 vs. 0.79, $p = 0.03$). Formparameter spielten eine wichtige Rolle für eine vielversprechende Vorhersage, während Intensitätsparameter die Vorhersageleistung nicht verbessern konnten. Insbesondere die Formmerkmale, die die Ovalität der Läsionen beschreiben, trugen am meisten zur Vorhersage bei: mittleres Läsionsvolumen, minimale Sphärizität und minimales Oberflächen-Volumen-Verhältnis.

Wir haben zwei computergestützte Verfahren zur Verbesserung der radiologischen Diagnostik von Multipler Sklerose entwickelt und ihre Vorteile gegenüber dem derzeitigen klinischen Arbeitsablauf nachgewiesen. Neue und vergrößerte Läsionen in Magnetresonanztomographie-Folgeuntersuchungen konnten mit Hilfe von DIR-Subtraktionskarten schneller und zuverlässiger erkannt werden als mit dem Standardverfahren. Wir konnten zeigen, dass unser Random-Forest-Modell, das sich auf die Formmerkmale der Läsionen in der ersten Magnetresonanztomographie-Untersuchung stützt, die Konversion eines klinisch isolierten Syndroms in eine Multiple Sklerose genauer vorhersagen kann als der Goldstandard.

Table of contents

Index of Figures

Index of Tables

List of abbreviations

3D	Three-dimensional
AI	Artificial intelligence
A	Surface area
CDMS	Clinically definite multiple sclerosis
CIS	Clinically isolated syndrome
CNS	Central nervous system
DICOM	Digital Imaging and Communications in Medicine
DIR	Double inversion recovery
DIS	Dissemination in space
DIT	Dissemination in time
DMT	Disease-modifying therapy
DOR	Diagnostic odds ratio
EDSS	Expanded disability status scale
FLAIR	Fluid-attenuated inversion recovery
LPA	Lesion prediction algorithm
LGA	Lesion growth algorithm
ML	Machine Learning
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
NIFTI	Neuroimaging Informatics Technology Initiative
NPV	Negative predictive value
PACS	Picture archiving and communication system
PPV	Positive predictive value
RDMS	Radiologically definite multiple sclerosis
RF	Random forest
S	Sphericity
SPM	Statistical Parametric Mapping
STD	Standard deviation
V	Lesion volume

1 Introduction

In 1868, Parisian neurologist Jean-Martin Charcot was the first to identify multiple sclerosis (MS) in one of his patients. Approximately a century later, the development of imaging techniques in the 1970s allowed more insight into the central nervous system (CNS) and improved the diagnosis of this disease. Today, magnetic resonance imaging (MRI) is recognized as the most sensitive and specific paraclinical test for MS (Polman et al., 2011; Swanton et al., 2007). Magnetic resonance findings support the clinical diagnosis and represent a sensitive and objective method to monitor disease activity over time (Thompson et al., 2017).

Regardless of the speciality, images account for a large proportion of data in the healthcare industry. However, the mere production of a large amount of data is not advantageous unless correct analysis that exploits their intrinsic value is executed. In the present clinical routine, MRI analysis is done manually by the radiologist screening the three-dimensional (3D) scan layer by layer, which is time-consuming and tiring. There are currently only limited tools for assisting radiologists in extracting information from medical images. The emergence of artificial intelligence (AI) via breakthroughs in technology and computer science since the 1960s paved the way for strategies to improve the analysis of extensive data collections and provide a suitable solution to processing and harnessing large amounts of data in a meaningful way.

Computational image analysis can compensate for the weaknesses of human image analysis and can serve as a solution for utilizing hidden information otherwise lost. As a branch of AI, machine learning (ML) can search for patterns that eludes the human eye, be quicker and never tire; therefore, providing suitable solutions to the problem of accurate image analysis in the medical field.

This thesis developed computational tools that assist radiologists in their diagnostic workflow regarding time-efficiency and accuracy for patients with MS. The basic understanding of MS, the role of MRI for this disease, existing computational tools, and AI methods are crucial for this purpose. The following chapters provide background information and are followed by the objectives of this thesis.

1.1 Multiple sclerosis

Multiple sclerosis is primarily an inflammatory disease of the CNS, characterized by multifocal demyelination and subsequent axonal degeneration. It is a challenging and disabling condition with a mean global prevalence of 33 per 100,000 (Multiple Sclerosis International Federation, 2014). Multiple sclerosis is one of the most common neurological diseases in young adults and the most common cause of non-traumatic neurological disability among this group (Koch-Henriksen & Sørensen, 2010).

Unknown triggers cause a regulatory defect in lymphocytes, starting a cascade of immune responses, leading to the breakdown of the CNS's blood-brain barrier (Viglietta, Baecher-Allan, Weiner, & Hafler, 2004). Due to this focal inflammation, perivenular autoreactive lymphocytes infiltrate and damage the myelin sheath of axons, leading to axonal loss, edema, and gliosis. These focal damages induce the reactive proliferation of astrocytes and, thus, the formation of multiple sclerotic plaques. These injuries mainly occur along veins (Tallantyre et al., 2008; Tan et al., 2000). The progression of MS happens due to neurodegeneration and is maintained by ongoing inflammation.

Although MS's exact etiology is unknown, underlying genetic susceptibility and environmental exposure, such as low vitamin D levels, cigarette smoking, viral infection, and obesity, are discussed as driving the condition (Thompson, Baranzini, Geurts, Hemmer, & Ciccarelli, 2018). Multiple sclerosis prevalence increases when family members are affected by this disease (Compston & Coles, 2008). The epidemiology pattern with disproportionately high frequencies in regions populated with mostly northern Europeans might suggest a genetic predisposition to MS. However, the change in risk by migration among people of the same ancestry indicates a role of environmental factors in this disease's genesis (Gale & Martyn, 1995; Kurtzke, 1993).

Due to different locations of inflammation and different forms of MS, the symptoms are highly heterogeneous. The most common syndromes and symptoms in the early stages are sensory disturbance, fatigue, and unilateral optic neuritis, resulting in blurred vision and painful eye movements. During the progression of MS, more symptoms can affect the patient, such as muscle spasms, urination incontinence and dysphagia. A systemic classification tool for the disability grade in MS patients is the *expanded disability status scale* (EDSS) of Kurtzke (Kurtzke, 1983). This scale is the most widely used score to assess the clinical disease progression and the effectiveness of therapy (Meyer-Moock, Feng, Maeurer, Dippel, & Kohlmann, 2014). The score classifies the patient's possible walking distance and eight functional systems, such as motor function, sensory function, bladder and bowel regulation, and the function of the brainstem, the cerebellum and vision (Kurtzke, 1983). The score ranges from 0, indicating no neurological deficits, to 10, meaning death due to MS (Kurtzke, 1983).

1.2 Diagnosis of multiple sclerosis

A definite MS diagnosis cannot be provided by any existing test, even including biopsy (Rovira et al., 2015). Therefore, diagnostic criteria were adopted, subject to constant modification as new evidence emerges, to interpret symptoms for their probability of MS being their etiology and excluding alternative diagnoses that can mimic MS.

The diagnostic principles are based on three steps. First, the suspected diagnosis of an inflammatory demyelinating disease of the CNS is formed by anamnesis of the symptoms and clinical examination. Second, differential diagnoses are excluded via laboratory results and MRI. Third, MS is confirmed by the criteria *dissemination in time* (DIT) and *dissemination in space* (DIS) by either imaging or clinical evidence.

Over time, imaging assessment has been increasingly incorporated in the information channel to diagnose MS. Magnetic resonance imaging is the best non-invasive diagnostic tool for imaging soft tissue in anatomic detail. The tool relies on strong magnetic fields and radio waves that measure the relative water content in the area of interest. In the field of MS, MRI is based on substantiating the suspicion of ongoing or lapsed inflammation in the brain and spinal cord, which are displayed in the form of lesions. These lesions are spots of demyelination and correspond to pathological findings in autopsy (Stewart, Hall, Berry, & Paty, 1984). Magnetic resonance imaging is the most sensitive technique detecting these lesions (Grossman & McGowan, 1998) since they represent an increase in tissue water due to the breakdown of the blood-brain barrier and subsequent macrophage migration and infiltration. Therefore, after the gadolinium application, these lesions are visible in contrast-enhanced MRI as bright areas.

Since 2001, a set of guidelines called the McDonald criteria have incorporated MRI findings to facilitate MS diagnoses (Mcdonald et al., 2001). This guideline has been updated several times. The McDonald criteria highlight the importance of lesion location and apply two key concepts: first, the proof of DIS, which means the presence of at least one lesion in at least two of the following regions: infratentorial, juxtacortical, periventricular, and spinal cord; second, the DIT, which is proven either by the simultaneous presence of a gadolinium-enhancing lesion and a non-enhancing lesion or the detection of a new lesion in a follow-up image (Thompson et al., 2017). According to these guidelines, radiological findings may even supplement clinical proof of DIT and DIS (Thompson et al., 2017).

In addition to being an essential tool for the diagnosis of MS, MRI plays a crucial role in the monitoring of MS disease progression: Images provide objective data that enable a precise comparison between scans of two time points. The monitoring of disease activity by identifying new or enlarged MS lesions between scan time-points has implications for the diagnostic and therapeutic procedure (e.g. selection of medication, neuroimaging frequency, and clinical follow-up).

According to the present guidelines, several MRI sequences are recommended for the standardized brain MRI protocol (Rovira et al., 2015). For the baseline evaluation, mandatory sequences include inter alia fluid-attenuated inversion recovery (FLAIR), T2-weighted (T2w) sequences and contrast-enhanced T1-weighted (T1w) sequences (Rovira et al., 2015). Characteristic abnormalities in MRI for MS patients include T2w-hyperintense or T1-hypointense lesions in the white matter. Sequences of FLAIR improve the detection of white matter and gray matter lesions by suppressing cerebrospinal fluid signal and blood-flow effects.

The double inversion recovery (DIR) sequence has an emerging role as a diagnostic tool for MS patients. The DIR sequence at 3 tesla provides the highest overall sensitivity for detecting MS lesions, compared with the T2w turbo spin echo and the FLAIR sequence (Khangure & Khangure, 2011; Wattjes et al., 2007). The DIR sequence is also a powerful tool for detecting cortical and infratentorial lesions in MS (Geurts et al., 2005; Vural, Keklikoğlu, Temel, Deniz, & Ercan, 2013; Wattjes et al., 2007). This sequence offers images with a lower signal-to-noise ratio than FLAIR images. Therefore, DIR images appear noisier, but the high contrast-to-noise ratio makes them usable in a subtraction map approach.

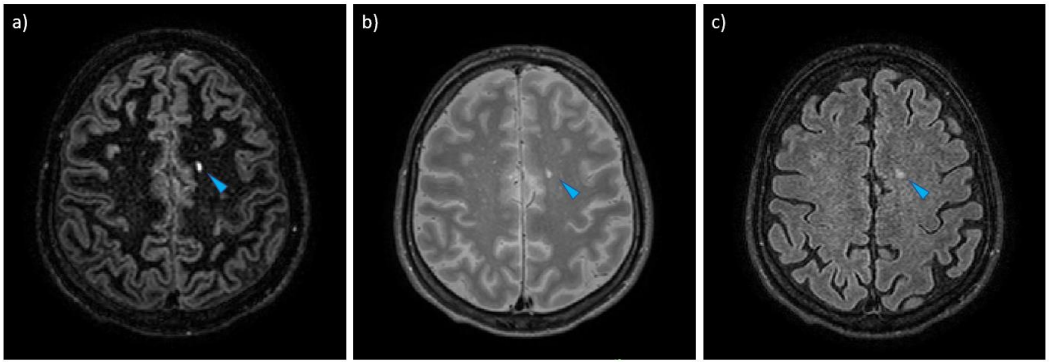

Figure 1 MRI sequences used for patients with MS a) DIR b) FLAIR c) T2w
This example displays the same brain-scan slice from a patient of the second project. The blue arrows point at a subcortical MS lesion.

1.3 Neuroradiologists' workflow for patients with multiple sclerosis

A core task of radiological diagnostics for patients with MS is analyzing new and enlarged lesions representing active inflammation. In clinical routine, this lesion quantification is done manually by neuroradiologists looking at T2w and FLAIR sequences to compute lesion count and total lesion volume. However, manual lesion detection and segmentation are difficult due to high variability in size, shape, and location and due to artifacts, which occur even under ideal conditions, making the processing time-consuming and suffering from interobserver variability.

Semi-automated and fully automated quantitative lesion segmentation methods that deliver reliable and accurate results compared with manual segmentation have been reported (Ashton et al., 2003; Egger et al., 2017; Filippi et al., 1995; P. Schmidt et al., 2012; Zijdenbos, Forghani, & Evans, 2002). These methods are beneficial regarding time effectiveness and the reduction of interrater variability while delivering objective measurements regarding lesion volume and lesion count. Thus, they enable standardized MRI quantification.

Semi-automated methods rely on threshold-based algorithms or region-growing algorithms. The first method is based on the automatic recognition of brightness thresholds. The user manually marks an area in the two-dimensional reformation, and the lesion border is delineated in that slice according to the highest difference in brightness between neighboring voxels. For region-growing algorithms, a beginning voxel, called seed, is selected manually by the user using a single mouse click, preferably in the middle of the lesion. Subsequently, neighbor voxels are analyzed and assigned iteratively to the lesion when meeting certain conditions. This algorithm ends when no further voxels are assigned to the lesion (Ashton et al., 2003; P. Schmidt et al., 2012).

There are several methods for the fully automatic segmentation of MS lesions. One approach was developed and evaluated in-house: the lesion segmentation tool (LST) for SPM (P. Schmidt et al., 2012). This algorithm uses a combination of a 3D gradient-echo T1-weighted and a FLAIR scan at 3 tesla to detect FLAIR-hyperintensities in patients with MS.

The assessment of lesion activity in follow-up images can be even more cumbersome since a comparison is needed between the current and previous scans. The images to be compared are not oriented and sliced in the same way, which complicates this procedure. For every lesion in the current image, a correlation must be searched for in the previous scan, making the evaluation especially time-consuming for patients with a high lesion load. Several studies have investigated the use of subtraction maps to address this problem. For this approach, serial scans of MS patients are co-registered, and the intensity values are subtracted voxel by voxel to produce images in which the radiological-stable situation and the background noise are cancelled out while alterations are directly visualized.

The foundation of using subtraction images to detect lesion change was laid by Hajnal et al. (1995), who investigated the improved detection of subtle changes with registered

subtraction images in patients with MS, among other brain diseases. In that study, differences due to contrast enhancement, such as in patients with MS, were visible.

Further studies have investigated subtraction maps for patients with MS using T2w sequences (Moraal, Pohl, et al., 2009; Tan, Van Schijndel, et al., 2002), 3D FLAIR sequences (Moraal et al., 2010; Tan, van Schijndel, Pouwels, Adèr, & Barkhof, 2002), and DIR sequences (Moraal et al., 2010), proving the superiority of subtraction maps over the standard pairwise comparison. More complex approaches in this field are based on supervised (Elliott, Arnold, Collins, & Arbel, 2013; Sweeney, Shinohara, Shea, Reich, & Crainiceanu, 2013) and unsupervised (Battaglini et al., 2014; Ganiler et al., 2014) subtraction pipelines to interpret the results of subtraction maps.

The DIR sequence, with its high lesion-to-parenchyma contrast, is promising for subtraction maps to visualize active lesions prominently. Moraal et al. (2010) investigated DIR subtraction maps for lesion detection in MS patients and found an increased detection of new lesions, but included only a small number of patients and testing in a clinical setting was not executed. Therefore, our goal was to provide a study that investigates a cohort of a bigger size with the hypothesis that DIR subtraction maps may provide an advantage in the analysis of follow-up images over the present gold standard. We aimed to provide a study of this hypothesis with a large study cohort.

1.4 Treatment of multiple sclerosis and clinically isolated syndrome

There is medication available to slow MS progression and relieve symptoms. The disease is usually diagnosed in young adulthood, and patients may need treatment for an extended period. The selection of those patients who benefit from early therapy is essential information since MS agents may come with severe side effects. For the long-term management of MS patients, disease-modifying therapies (DMT) delay progression and reduce early disease activity that would likely contribute to future disability. (Hart & Bainbridge, 2016)

These DMTs are especially of interest for patients with clinically isolated syndrome (CIS), a possible preliminary stage of MS. The syndrome refers to a single episode of neurologic symptoms suspected of an inflammatory demyelinating event of the CNS. The clinically isolated syndrome is recognized as an initial presentation of MS without fulfilling the DIT criterion (Miller et al., 2005; Miller, Chard, & Ciccarelli, 2012). A second clinical exacerbation is evidence of DIT and would qualify as *clinically definite* MS (CDMS*).* Additionally, conversion from CIS can also be determined by radiological proof of DIT, which is the simultaneous presence of active and lapsed lesions or the appearance of new lesions in later scans. These cases are considered *radiologically definite* MS (RDMS). The number of CIS patients who suffer a second clinical attack can be reduced by admission of DMTs, as well as their MRI activity and MS conversion can be delayed by DMTs (Comi et al., 2001, 2009; Jacobs et al., 2000; Kappos et al., 2016). The effect of these drugs is particularly beneficial in the very early stages (Comi et al., 2001; Kappos et al., 2016).

Therefore, MRI plays a key role in the initial evaluation of patients with CIS. Magnetic resonance imaging abnormalities in patients with MS or its preliminary stage are considered predictive for disease progression and future disability (Barkhof et al., 1997; Brex et al., 2002; Fisniku et al., 2008). Although the risk of a second clinical attack is affected by clinical factors including oligoclonal bands in the central spinal fluid, and male gender and older age at onset, the conversion risk can be assessed most consistently by the presence of lesions (Dalton, Brex, Miszkiel, et al., 2002; Swanton et al., 2007; M. Tintoré et al., 2003; W. Y. Zhang & Hou, 2013), as well as the number of lesions in the initial scan (Fisniku et al., 2008; M. Tintoré et al., 2006). Furthermore, 70–80% of CIS patients with an abnormal scan developed CDMS in long-term follow-up studies of 15–20 years, whereas only 20–25% of CIS patients with normal imaging converted to CDMS (Brodsky et al., 2008; Fisniku et al., 2008). Periventricular lesions and lesions in the brainstem and spinal cord also correlate with disease progression (Di Filippo et al., 2010; Giorgio et al., 2013; Mostert et al., 2010)

Early detection of CIS and its early management are crucial for delaying disability progression via the early administration of MS agents. Moreover, the proportion of CIS patients who experience a favorable course should not be neglected, and their needless medication should be avoided. Therefore, the reliable prediction of the individual conversion risk is highly relevant from the first presentation.

1.5 Machine-learning methods

Given that modern (neuro-)radiology is primarily a data analysis task, implementing modern ML techniques is a suitable solution to cope with the high numbers of generated data. As a branch of AI, ML enables systems to learn automatically and improve from data without relying on a predetermined equation. Machine learning begins with a mathematical algorithm called a learner, which repeatedly modifies its operating ways by iterating over the training data, and thereby builds a ML model. With continuous exposure to new data, ML models adapt independently and learn from previous computations until a robust pattern is found. This process results in reliable and repeatable decisions. The trained model is tested on unseen data, called test data, to evaluate its real-world performance. (Mitchell, 2010)

The two main basic approaches in ML are supervised learning and unsupervised learning. Supervised learning algorithms are used to classify data or predict outcomes from historical data. These algorithms are trained by receiving a set of labeled inputs and outputs, meaning inputs have corresponding outputs (e.g. a medical record that contains several potential risk factors and the corresponding information of whether a patient developed a disease). The algorithm learns patterns and measures their accuracy from training data to predict values on unseen test data. Labeling the datasets for training can be a laborious process. One advantage of unsupervised learning algorithms is that no corresponding outputs are necessary because they are trained with unlabeled data to explore the structure within (e.g. similarities or anomalies that stand out) to gain insights into large volumes of new data. (Mitchell, 2010)

A random forest (RF) is a supervised ML algorithm used for prediction. It is a commonly used and popular algorithm because of its simplicity and utility for classification and regression tasks. An RF model builds multiple decision trees and then merges them to develop a stable prediction. A decision tree is built on two elements: nodes and branches. Each node represents a test on input features, and each branch represents the outcome of the test. Each individual tree is trained via a randomly chosen sample of subsets of the entire dataset, and at every node, certain input features are selected randomly for evaluation. This process leads to constructed heterogeneity and decorrelation. The final nodes are called leaf nodes and represent the final prediction, and therefore, the attribution of a category (e.g. disease or no disease) or a numerical class is made. (Breiman, 2001; Masetic & Subasi, 2016)

An additional useful quality of the RF is the possibility to measure the relative importance of each input feature on the prediction. In the medical field, clinical risk factors often function as input features, and onset of a disease as the prediction endpoint; thus, the feature importance can deliver a better understanding of the genesis and development of a disease. Additionally, the feature importance enables the user to decide on which features could be excluded from a model since they do not contribute enough to the prediction process. The limitation of features is important because of a general rule in ML: the more features put

into an algorithm, the more likely it will suffer from overfitting, and vice versa. (Breiman, 2001)

Overfitting is one of the biggest problems in machine learning. It occurs when a model with high capacity essentially memorizes the training data by fitting them too closely. The problem stems from the model not only learning the actual relationships in the training data, but also any present noise, thus, generalizing unsatisfyingly on unseen test data. Fortunately, the RF solves this problem by resampling the trees and constraining the number of nodes. However, a selection of input features with high contribution to the classification may improve the model's performance. (Breiman, 2001)

1.6 Machine learning in the field of multiple sclerosis

Applying ML in the field of MS is particularly interesting since clinicians aim to prevent or delay the disease progression from the preliminary stage of CIS and from early MS stages, but the details of risk factors and pathogenesis are still unclear.

As mentioned previously, MRI is the most sensitive paraclinical tool for MS and delivers objective data. Neuroimaging predictors of clinical outcomes in patients with CIS have demonstrated that number, location, and distribution of asymptomatic white matter lesions on baseline scans are associated with conversion to CDMS (Alroughani, Al Hashel, Lamdhade, & Ahmed, 2012; Brex et al., 2002; Fisniku et al., 2008; Giorgio et al., 2013; M. Tintoré et al., 2003; W. Y. Zhang & Hou, 2013). These predictors only incorporate data visible to the human eye and somehow tangible for the radiologist. Machine-learning methods can help to extract and process further intrinsic value from these data to include in the prediction criteria.

At the time of the second project (March to October 2017), only a few studies had dealt with the prediction of conversion from CIS to MS using ML methods. Wottschel et al. (2014) used support vector machines to predict the conversion by combining clinical features and information from baseline MRI scans, including inter alia, lesion count, lesion load, lesion intensities, and lesion distances. Other studies with a similar aim used advanced MR techniques, such as measuring myelin water fraction in white matter (Kitzler et al., 2018) or MR spectroscopy (Ion-Mărgineanu et al., 2017). A prediction of conversion by analyzing lesions' shape and intensity features from MRI baseline scans via a RF model has not been performed at that time.

1.7 Research aims and objectives

Evaluating lesions in MR images is an important part of the diagnosis and follow-up for patients with MS. This process is tedious and lengthy; therefore, tools to simplify and accelerate this process for neuroradiologists are desirable in everyday clinical practice. Furthermore, the evaluation of image data beyond human perception can improve each patient's individual assessment. This work was driven by the motivation to provide practical solutions that are ready and easy enough to be integrated into the clinical routine. The aim was to build computational tools that assist neuroradiologists in the diagnosis of patients with MS and CIS. This thesis consists of two parts.

The first project was implemented between September 2016 and July 2017. At that time, few studies of subtraction maps of patients with MS have investigated the diagnostic performance and time saving with a large cohort. We aimed to develop an algorithm that calculates longitudinal intensity subtraction maps from DIR sequences to improve the detection of new or enlarged lesions in follow-up images of patients with MS. The specific objectives were as follows:

- Analyse the accuracy and time of lesion detection using DIR subtraction maps and FLAIR subtraction maps
- Compare the results with the standard pairwise image analysis
- Investigate the tool’s utility for non-neuroradiologists

The second project was implemented between September 2017 and November 2018 and aimed to develop a machine-learning tool to predict CIS patients' possible conversion to MS using imaging features from their baseline scans. The specific objectives of the second project were as follows:

- Assess the predictive performance of shape and intensity features
- Compare the best predictive model with the gold standard
- Determine the features that are the most important for the prediction
- Identify the possible effect of the segmentation method on the prediction performance
- Study weak points for future application of ML methods in this field

In the following chapter, the first project about subtraction maps is described. Chapter 3 contains the second project about the RF prediction model. Chapter 4 outlines the main conclusions of both projects and delivers recommendations for further research.

2 Lesion detection using a DIR subtraction map

The content of this chapter was published as “A novel imaging technique for better detecting new lesions in multiple sclerosis” in the *Journal of Neurology* on 29th July 2017 (Eichinger et al., 2017).

2.1 Methods

2.1.1 Subjects

The observational cohort included 106 patients with a confirmed MS or CIS diagnosis from the MS database of the Department of Neurology of the *Klinikum Rechts der Isar*: 65.0% (n = 69) were female and 35.0% (n = 37) were male. The F:M ratio was 1.86:1. The age ranged from 17 to 66 years old, and the mean age at onset was 33 ± 11 years. All patients received at least two MRI scans between January 2014 and March 2016 every 6 to 12 months. The first and the last scans were at least 13 months apart and were used as a follow-up pair. In total, 212 scans were evaluated.

2.1.2 MRI acquisition

All MR images were acquired using a 3 tesla MR scanner (Achieva, Philips Healthcare, Best, the Netherlands). Every scan included the sequences 3D FLAIR, 3D DIR, and 3D T2-turbo spin echo. Additionally, at least one 3D gradient-echo T1w sequence using a magnetization-prepared 180-degree radiofrequency pulse and rapid gradient-echo sampling (pre- and/or post-contrast) were included. The imaging parameters are stated in the Appendix (Chapter 6.1.1 on page XXII).

2.1.3 Data processing

For this stated workflow, we used MATLAB 8.6. The citation of all software used for this project can be found in the Appendix on page XXVII. Our custom-developed scripts are in the Appendix on page XXIII and also public on Github: 'https://github.com/CompImg/DIR-sub'.

FLAIR and DIR sequences were converted manually from the Digital Imaging and Communications in Medicine (DICOM) file format in the Picture Archiving and Communication System (PACS) to the Neuroimaging Informatics Technology Initiative (NIfTI) file format. We used the custom-built script Statistical Parametric Mapping package (SPM 12) for MATLAB to perform a rigid registration to co-register the baseline DIR image and the follow-up DIR image of each patient. The follow-up image was set as *reference,* and the baseline image as *source*. Subsequently, the intensities of the registered baseline images were subtracted voxel by voxel from the intensities of the follow-up images, obtaining a subtraction map (Figure 2).

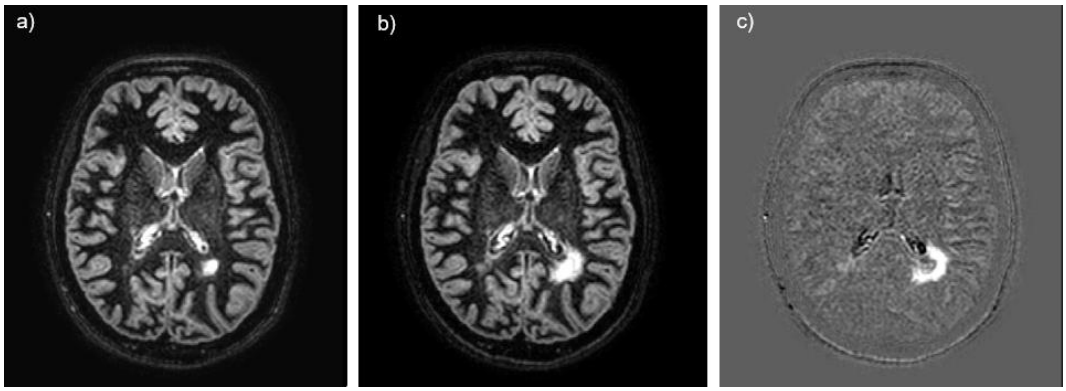

Figure 2 Figurative illustration of the subtraction pipeline for DIR images
The same axial images from a patient of the study cohort: a) baseline image set as source, b) follow-up set as reference, c) resulting subtraction image.

For the FLAIR images, subtraction maps were calculated amorously with additional post-processing steps: we used the *N4 bias correction* algorithm to perform biasfield correction. Then, we used the *Robust Brain Extraction (ROBEX)* algorithm for brain extraction. The citation of all algorithms used for this project can be found in the Appendix on page XVIII. The follow-up scan was set as *reference,* and the baseline scan was set as *source*. Later, histogram matching was performed on the baseline scan using MATLAB's *imhistmatch* function. The resulting image was subtracted in the same manner as described for DIR images from the follow-up image, obtaining the FLAIR subtraction map (Figure 3).

A similar post-processing procedure was performed for DIR subtraction maps. However, both neuroradiologists came to the same conclusion, via the visual inspection of examples of generated subtraction maps, that no benefit was gained compared with the subtraction maps obtained from non-post-processed DIR images. Thus, this post-processing step was omitted in the DIR subtraction map algorithm to avoid slowing the process.

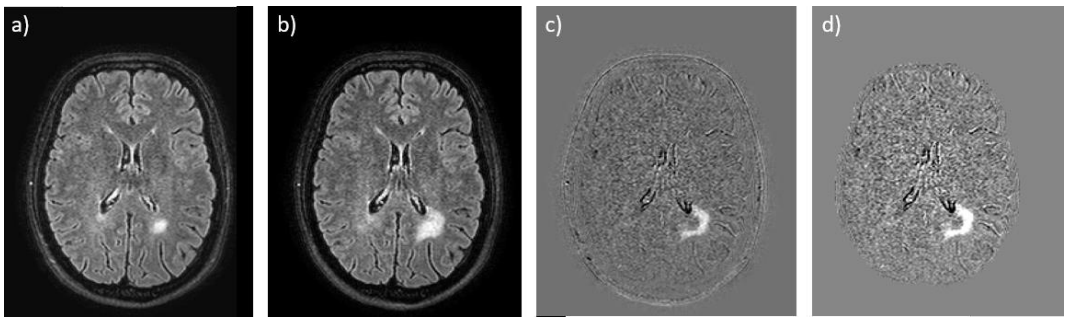

Figure 3 Figurative illustration of the subtraction pipeline for FLAIR images
The same axial brain slices of the same patient as in Figure 2: a) FLAIR baseline image, b) FLAIR follow-up image, c) subtraction map without post-processing steps, d) corrected subtraction map.

2.1.4 Image readout protocols

Two experienced neuroradiologists assessed the images (Paul Eichinger, five years' experience; Benedikt Wiestler, eight years' experience) on multi-display PACS workstations. Each neuroradiologist assessed all patients in three different readouts, with one week between each to minimize memory bias. Images were viewed and annotated for new lesions in consensus using ITK Snap 3.6.

The first readout imitated the standard clinical process. Therefore, images were compared pairwise visual screening of all the available sequences in a non-registered form featuring a split-screen modality. Each axial brain slice of the follow-up image was evaluated. In the case of a lesion, the corresponding lesion was searched for in the baseline picture to judge whether it was a new lesion or an enlargement of an existing one.

In the second readout, only DIR subtraction maps, the co-registered follow-up DIR image, and baseline DIR image were viewed on a single laptop screen. The axial slices of the DIR subtraction map were screened for hyperintensities. In case of detection, both DIR images were used to confirm the finding as a real new lesion as opposed to an artifact in the subtraction map. In Figure 4, two examples illustrate how new and enlarged lesions were depicted to the viewer in the second readout. For the third readout, the FLAIR subtraction maps, the co-registered FLAIR follow-up images, and the FLAIR baseline images were used analogously.

A neurologist (Hanni Wiestler, five years' clinical expertise in MS treatment) and a medical student (Haike Zhang, 4th year, no prior imaging experience, brief introduction) executed the same readout protocols on single laptop screens.

The two neuroradiologists reviewed the combined information from all four readers, and all readout protocols in a consensus read to define the reference standard. No other independent standard was used to define the lesions on the imaging data. The primary outcome measure was set as the existence of new or enlarged lesions. For each patient, all readers recorded the overall time for evaluation and lesion annotation. The number of new and enlarged lesions was counted and classified according to their location into *periventricular*, *juxtacortical/cortical*, *subcortical,* or *infratentorial*. Additionally, lesion size was semi-quantitatively classified into *small* or *large* based on their diameter.

The diameter was measured by the *distance measuring* tool provided by ITK Snap 3.6. As proposed in the 2016 MAGNIMS criteria (Filippi et al., 2016), only lesions with a diameter greater than or equal to 3 mm were considered. As proposed by Moraal et al. (2009b), lesions were regarded as enlarged in the case of an increase in diameter of more than 50%. A change in lesion shape suggestive of confluent lesions was considered enlargement, even if the diameter did not increase more than 50%.

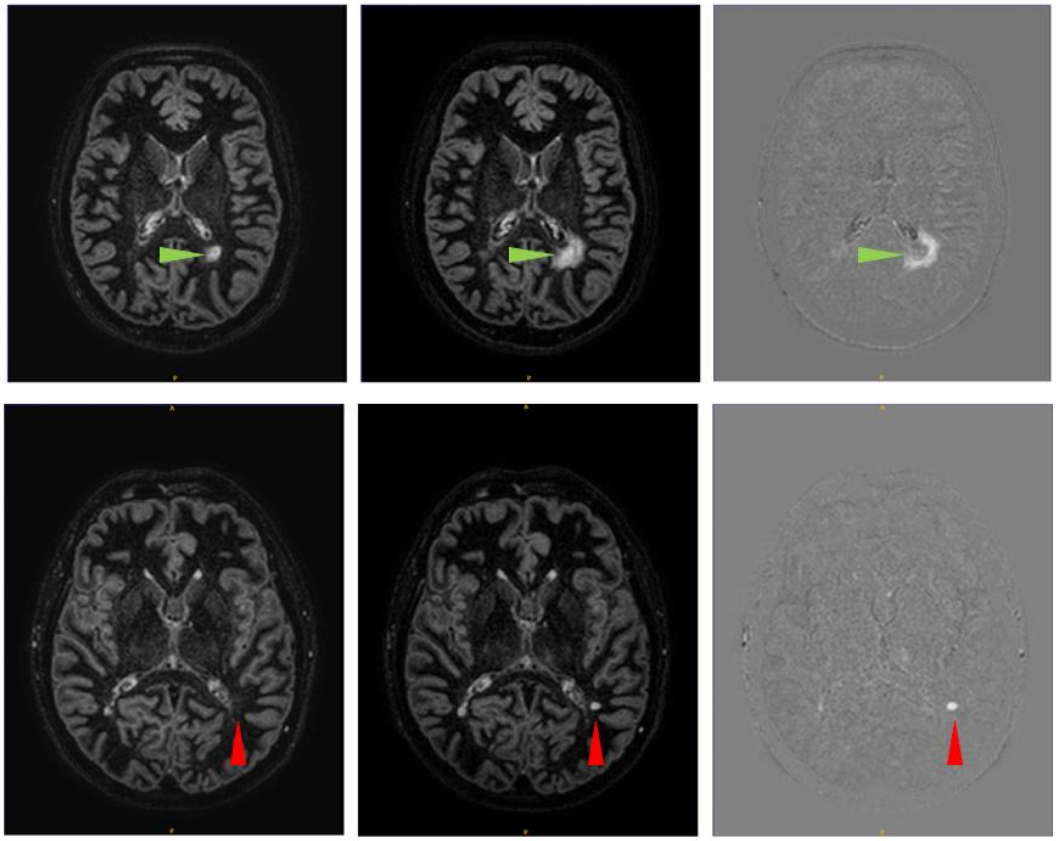

Figure 4 Examples of new and enlarged lesions in the DIR subtraction map
The upper row illustrates a patient from the study cohort with a lesion (green arrows) in the white matter of the left posterior brain quadrant that is enlarged compared with the baseline scan. The lower row displays another patient from the study cohort with a new lesion (red arrows) in the white matter of the left posterior quadrant of the brain.

The threshold for large lesions was set as a diameter longer or equal to 5 mm. This threshold was chosen since it divides the group of new or enlarged lesions found by the standard readout into two nearly equally large groups (192 lesions < 5 mm, 198 lesions ≥ 5 mm). This separation allows statistically meaningful analysis.

The quality of the subtraction images was semi-quantitatively graded into three categories: *2*: all hyperintensities in brain parenchyma on subtraction maps depict new lesions; *1*: source images were needed in at least some hyperintensities to decide whether these depict new lesions or are artificial; and *0*: no additional benefit of the subtraction images over-using the source images alone.

2.1.5 Statistical analysis

The statistical calculations were performed in MATLAB 8.6 and R 4.0.3. The significance level was set at $p = 0.05$. Standard diagnostic accuracy measures were calculated, including sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). The image analysis methods were compared with each other and the reference standard using an exact conditional McNemar's test (McNemar, 1947). Wilcoxon signed-rank tests for paired observations were used to compare the numbers of new lesions per patient, the time needed for each image analysis, and the image quality of the subtraction maps.

2.2 Results

2.2.1 Image processing

In a clinical setting, the average time for preparing the image for the MATLAB script took approximately 4 min. The preparation included transferring the image data from the MR scanner to the processing computer and the organization of the data to be put into MATLAB. The proposed subtraction algorithms' calculation took approximately 45 s for each DIR map, and approximately 3:15 min for FLAIR subtraction maps since post-processing, as mentioned in Chapter 2.1.3 (page 13), was needed.

2.2.2 Subtraction map quality

In both versions of the subtraction maps, newly occurring lesions revealed as hyperintense, whereas preexisting lesions were cancelled out (Figure 4, above). Table 1 lists the quality of the subtraction maps. The one DIR map with grade *0* was due to heavy motion artifacts in both DIR images (Figure 5). In comparison, the DIR map quality proved significantly higher than that of the FLAIR maps ($p = 0.004$, Wilcoxon signed-rank test).

Table 1 Quality of the subtraction maps

Grade	0	1	2
DIR maps	1	37	68
FLAIR maps	2	55	49

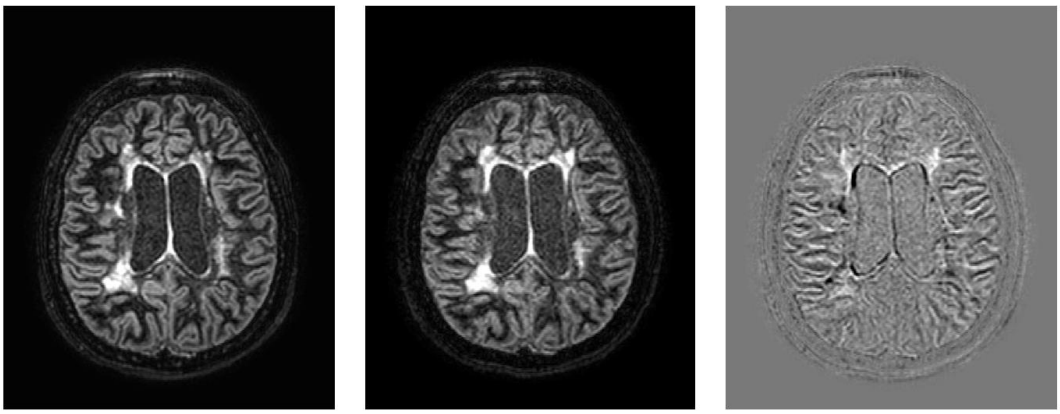

Figure 5 The one DIR map with grade *0*

2.2.3 Algorithm performance

In our cohort, 77 out of 106 patients (73%) displayed disease activity as per retrospective ground-truth assessment. Since the reference was set as the consensus read, as previously explained, all classification in the standard reference is considered correct. Table 2 lists the results of the three readout sessions. Seventy-three patients (95%) with new or enlarged lesions were detected using the DIR subtraction maps, whereas using the standard comparison only correctly detected 63 patients (82%).

The measures of diagnostic accuracy are stated in Table 3. The DIR readout significantly outperformed the standard readout: a sensitivity of 0.95 compared with 0.82 (McNemar test, p = 0.013). No false positives occurred in either of these two readouts, so specificity was 1.00 twice. The diagnostic accuracy of the DIR map readout of 0.96 was significantly higher than the standard readout of 0.87 (exact McNemar, p = 0.013).

The FLAIR map readout correctly classified 61 patients (79%) with new or enlarged lesions and with three false positives, resulting in a sensitivity of 0.79 and a specificity of 0.90. The sensitivity proved significantly worse than in the DIR maps (exact McNemar, p < 0.001), whereas the discrepancy from the standard readout did not prove significant (exact McNemar, p = 0.36).

Table 2 Results from the three readouts

a) Standard comparison	**New lesion**	**No new lesion**
New lesion found	63	0
No new lesion found	14	29
b) DIR map		
New lesion found	73	0
No new lesion found	4	29
c) FLAIR map		
New lesion found	61	3
No new lesion found	16	26

Source: Table based on Eichinger et al. (2017, Tab. 2b)

Table 3 Diagnostic accuracy measures for all readouts

	Standard readout	DIR map	FLAIR map
Accuracy	0.87 (0.81–0.93)	0.96 (0.92–1.00)	0.82 (0.75–0.89)
Sensitivity	0.82 (0.79–0.90)	0.95 (0.87–0.99)	0.79 (0.68–0.88)
Specificity	1.00 (0.88–1.00)	1.00 (0.88–1.00)	0.90 (0.73–0.98)
PPV	1.00	1.00	0.95 (0.87–0.99)
NPV	0.67 (0.56–0.77)	0.88 (0.74–0.95)	0.62 (0.51–0.72)

Abbreviations: PPV positive predictive value, NPV negative predictive value. The 95% confidence interval is shown in parenthesis.

Source: Table based on Eichinger et al. (2017, Tab. 2a)

2.2.4 Lesion counts

The mean number of total lesions per person and the mean numbers classified into their location are listed in Table 4. In the DIR readout, approximately 1.7 times as many new lesions were found, on average, per person compared with the standard readout (6.24 vs. 3.67). This discrepancy proved significant in the Wilcoxon signed-rank test ($p < 0.001$). Regarding the individual predefined areas, significantly more lesions were detected by using DIR subtraction maps than by the standard assessment for every location ($p < 0.001$ for each location, Wilcoxon signed-rank test). Lesion detection in one location in particular benefited from the use of DIR maps: The *juxtacortical/cortical* region reached a 2.1-fold increase in mean lesion detection compared with the visual readout (1.20 vs. 0.58). Regarding lesion size, the DIR subtraction maps offered a significant improvement, compared to standard readout, for detecting lesions that are assigned as big (2.90 vs. 1.87, $p < 0.001$) and small (3.35 vs. 1.81, $p < 0.001$).

FLAIR maps only provided an advantage over the standard assessment regarding location in subcortical lesions, but not significantly (Wilcoxon signed-rank test, $p = 0.053$). In all other lesion locations, using FLAIR maps led to worse results than by using the standard comparison. The difference is significant for juxtacortical ($p = 0.001$) and infratentorial lesions ($p = 0.002$), but not significant for periventricular lesions ($p = 0.099$). The difference between DIR- and FLAIR readouts regarding all lesion locations and total lesion count was highly significant in favor of the DIR readouts (Wilcoxon signed-rank tests, subcortical $p = 0.005$, all other locations $p < 0.001$).

Table 4 Mean number of new lesions per patient in the respective location

	Standard	DIR map	FLAIR map	DIR map vs stand-ard[1]	FLAIR map vs stand-ard[1]	DIR map vs FLAIR map[1]
Periventricular	1.47 ± 3.58	2.25 ± 4.46	1.15 ± 2.49	P < 0.001	P = 0.099	P < 0.001
Juxtacortical / cortical	0.58 ± 1.70	1.20 ± 2.89	0.28 ± 0.95	P < 0.001	P = 0.001	P < 0.001
Subcortical	1.35 ± 3.68	2.24 ± 5.44	1.55 ± 3.51	P < 0.001	P = 0.053	P = 0.005
Infratentorial	0.27 ± 3.51	0.55 ± 1.26	0.08 ± 0.36	P < 0.001	P = 0.002	P < 0.001
Total	3.68 ± 9.05	6.26 ± 12.67	3.06 ± 6.69	P < 0.001	P = 0.19	P < 0.001
Small	1.81 ± 4.44	3.35 ± 6.88	1.12 ± 2.32	P < 0.001	P = 0.004	P < 0.001
Big	1.87 ± 5.26	2.90 ± 6.24	1.93 ± 4.80	P < 0.001	P = 0.228	P < 0.001

Mean ± standard deviation
[1]Wilcoxon signed-rank test

Source: Table based on Eichinger et al. (2017, Tab. 3)

2.2.5 Readout time

The time per patient for each readout was documented in full minutes. Figure 6 illustrates the median time and interquartile range per readout. The DIR map readouts and FLAIR map readouts were significantly quicker than the standard readouts (2 min vs. 8 min, $p < 0.001$, Wilcoxon signed-rank test). Significance even remained in a comparison between the DIR readout time and one-third of the standard readout time ($p = 0.007$, Wilcoxon signed-rank test). For patients without new lesions, the median standard readout was 6 min, whereas readout using DIR maps took 1 min in median and 2 min at most, and the readout with FLAIR maps took 2 min in median.

Transferring the image data to the computer, which processes the subtraction algorithm, took 4:45 min per patient. Adding this time to the readout time for the DIR maps, this pipeline still provides a quicker overall process than the standard readout (6:45 min vs. 8 min, $p = 0.002$, Wilcoxon signed-rank test).

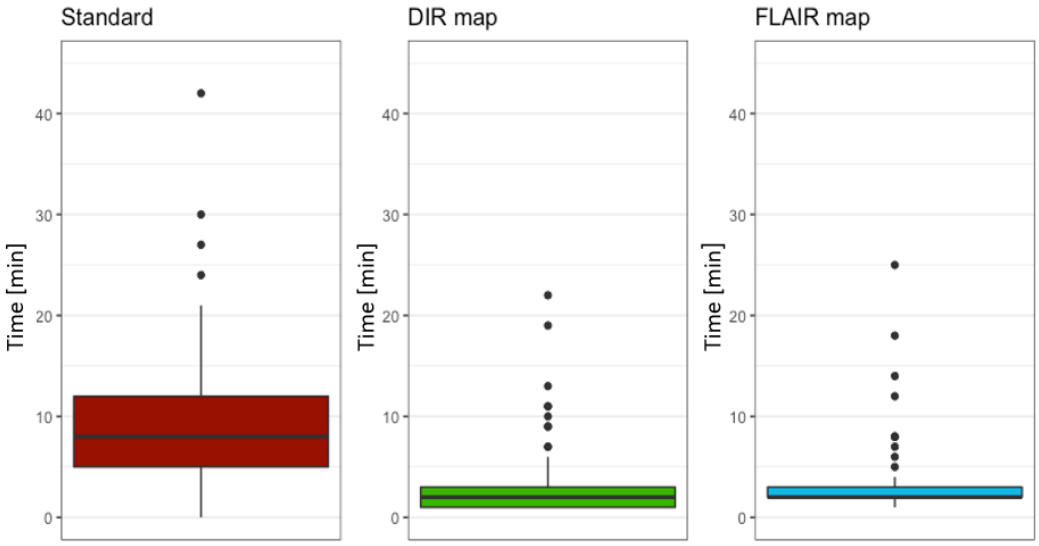

Figure 6 Median readout times and their interquartile range in comparison
The median readout time using the standard method is 8 min, whereas the median times for DIR and FLAIR map readouts is 2 min. The remaining exact numbers are in the Appendix on page XXII.

Source: Figure based on Eichinger et al. (2017, Fig. 5)

2.2.6 Application for non-neuroradiologists

To assess the usefulness of the subtraction maps for readers without prior training in the radiological screening of MS patients, a neurologist and a medical student performed the three readout protocols. The results of each readout are summarized in Table 5.

The neurologist and the medical student improved their detection sensitivity using DIR maps, but specificity decreased for the medical student from 0.90 to 0.79. In the FLAIR readout, both the neurologist and the medical student decreased in specificity to 0.52 and 0.59, respectively.

The DIR maps allowed both the neurologist and the medical student to speed up their evaluation. The median time per patient improved significantly, from 10 min to 3 min for the neurologist, and from 15 min to 5 min for the medical student (two-tailed t-test, both $p < 0.001$).

Table 5 Evaluation of the readouts by the neurologist and the medical student

neurologist	**standard**	**DIR map**	**FLAIR map**
sensitivity	0.82	0.95	-
specificity	0.86	1.00	0.52
median time per patient [min] ± standard deviation	10 ± 6.91	3 ± 5.27	3 ± 2.80
medical student			
sensitivity	0.59	0.80	-
specificity	0.90	0.79	0.59
median time per patient [min] ± standard deviation	15 ± 6.62	5 ± 2.06	5 ± 2.15

2.3 Discussion

This project investigated using DIR subtraction maps to improve detection of new or enlarged lesions in follow-up images of patients with MS in a clinical setting. In this context, accuracy and speed of the proposed method were assessed in comparison with the standard readout.

The DIR maps outperformed lesion detection by standard visual comparison and FLAIR maps in terms of accuracy and speed. Additionally, by using DIR maps, significantly more lesions per patient were found than by using the standard readout. Using subtraction maps also improved the sensitivity and speed of lesion detection for non-neuroradiologists.

This study delivered promising results and suggests using DIR subtraction maps as a useful application in a clinical setting. For the evaluation of these novel imaging techniques, the significant findings are discussed in detail in the following sections.

2.3.1 Higher detection accuracy using DIR subtraction maps

The sensitivity for detecting new or enlarged lesions increased markedly from 0.82 using the standard readout to 0.95 using the DIR maps. For neuroradiologists, who are experienced with DIR images, specificity was not compromised. Specificity was 1.00, since no false-positive cases occurred, although no additional MR sequences were analyzed in the DIR readout.

Our sensitivity is similar to the value (0.91) in the study of Battaglini et al. (2014), who applied an automated lesion-detection method on DIR subtraction maps of 19 patients. A unsupervised subtraction approach, incorporating multisequence information, was proposed by Ganiler et al. (2014) and tested on 20 patients. That pipeline provided slightly lower sensitivity (0.83) than our DIR map subtraction approach.

Our findings suggest that a reliable assessment of new or enlarged lesions' overall presence is possible using DIR images only. This finding supports the present considerations of using DIR as a single sequence imaging technique to detect alterations of MS lesions (Khangure & Khangure, 2011). A single MRI sequence would be favorable since image-acquisition time in follow-up monitoring would be decreased when serial contrast scans are unnecessary. At the same time, it must be considered that follow-up scans also deliver information about negative drug side effects, which might be poorly presented in the DIR sequence.

It should be noted also that the reference is set up by combining the results from all the different readouts from all four readers because a human reader is naturally necessary to interpret imaging. All values of diagnostic accuracy are related to this reference, which was acquired by a “consensus read” by the two neuroradiologists, therefore not being

completely independent. However, we put great effort into establishing a reliable reference via an independent consensus read. A considerable improvement to this approach could be to determine the reference from independent neuroradiologists.

2.3.2 Higher detected lesion count using DIR maps

In the DIR readout, significantly more lesions were found, on average per patient, than by the standard readout (median lesion number 6.24 vs. 3.67). Detection of new or enlarged lesions increased 1.7-fold using the DIR subtraction maps compared with the standard readout. The DIR maps exceeded visual comparison, particularly for detecting juxtacortical/cortical lesions, with more than twice as many new active lesions, on average, being found (1.2 vs. 0.58).

Our lesion count factor is in line with a previous study that proved a 1.7-fold higher lesion-detection rate using 2D subtraction maps (Moraal, Meier, et al., 2009) compared with the conventional pairwise readout.

Patients in our cohort presented with a high base rate of overall existence of new lesions, because we chose scan pairs with maximized time interval between to allow the maximization of new lesions. This interval was needed to obtain a substantial quotient of new lesions found by the different methods. However, this act leads to a less pronounced difference in NPVs for new lesions' overall existence in a clinical setting, in which follow-up scans are taken every six to 12 months.

Changes in lesion count and total lesion load are used in a clinical setting to assess disease progression and response to therapy. These are quantitative measures of focal differences in pathology between scans. A sensitive and robust method to determine these values quickly is useful to treat patients individually.

2.3.3 Quicker lesion detection using DIR maps

The DIR subtraction maps sped up image analysis significantly. Using the DIR maps took less than one-third of the time required for the visual comparison (median time: 2 min vs. 7 min). In patients without new lesions, the median readout time was 2 min, and 1 min at most using DIR maps, compared with 7 min in the standard-setting.

The readout time for these patients is a reasonable speed indicator for detecting new lesions' overall existence since the additional time for readouts of patients with new lesions was needed to mark and annotate lesions. When the time for data organization and algorithm performance was added, lesion evaluation took 7 min in total per patient.

Both studies by Moraal et al. (2009; 2010) which explore the advantage of DIR subtraction maps in patients with MS, closely to our objectives, did not investigate the time efficiency in his studies about DIR subtraction maps.

The time for transferring the data from PACS to other computers, organizing them, and performing the subtraction algorithm must be added to the subtraction maps' readout time to evaluate the time savings of subtraction maps compared with the standard procedure. The existence of new lesions can reliably be evaluated within 7 min after acquiring the scan. However, this approach requires basic familiarity with the working pipeline using the program MATLAB. To increase the subtraction algorithm's acceptance into the clinical routine, we recommend implementing the algorithm in the computer that communicates with the PACS system. This integration would also benefit the preparation time between scan acquirement and analysis.

Although it takes time and software familiarity to use this algorithm, it saves overall time compared with the standard visual comparison. Therefore, the reasonable effort necessary is a trade-off for substantially quicker image analysis. The proposed DIR subtraction map approach delivers rapid detection of newly developed or enlarged lesions in MS patients' follow-up scans.

2.3.4 No advantage of FLAIR maps in terms of time and accuracy

In the FLAIR readout, less lesions were found, on average, per patient than by the standard readout (median lesion number: 3.06 vs 3.67) and by using DIR maps (6.26). FLAIR maps only provided a non-significant advantage in detected lesion number over the standard readout in one location: the subcortical region. The sensitivity for detecting new lesions decreased markedly to 0.79 using the FLAIR maps, which is non-significantly worse than in the standard readout (0.82) and significantly worse than using DIR maps (0.95). In this study, FLAIR maps showed no advantage over standard comparison regarding lesion count and detection sensitivity.

When comparing our results to a previous study with similar aim by Tan and Van Schijndel et al. (2002), it must be pointed out that they found a 1.46-fold increase in lesion detection using FLAIR maps compared with the conventional pairwise readout. However, only a small cohort of 20 patients was investigated with a 6-month period between the serial scans and 3 of the patients did not develop any new or enlarged lesions.

FLAIR maps readouts took less than one-third of the time required for visual comparison (median time: 2 min vs 7 min, respectively). This finding is in accordance to another study which used 3D FLAIR subtraction maps to detect new MS lesions: M.A. Schmidt et al. (2018) found a 5-fold improvement in mean reading time compared with the standard side-by-side readout (35.6 s vs 163.7 s). The low mean reading time and increase in reading speed that exceeds our numbers could be explained by the low number of investigated

patients: The study only included 20 patients, and 10 of those had no new lesions, and only one patient had more than four lesions.

FLAIR maps provide a faster lesion detection than the standard readout and a comparable readout time as DIR map readouts. However, calculating, and post-processing for each FLAIR subtraction image (3:15 min) takes more time than for DIR maps (0:45 min). Taking this into account, the total amount of time for calculation and readout of FLAIR maps is longer than for DIR maps according to our approach.

2.3.5 DIR maps also useful for non-neuroradiologists

We assessed the usefulness of our proposed method for readers who are not trained neuroradiologists. The neurologists achieved the same sensitivity (0.82) in the standard readout as the neuroradiologists at the cost of a lower specificity (0.86 vs 1.00). This implies a less restrictive approach by the neurologist with the downside of producing more false-positive results. With DIR maps, the neurologist's performance increased to a level that exactly matches the results of the neuroradiologists (sensitivity 0.95, specificity 1.00).

The medical student expectedly delivered a lower detection sensitivity in the standard readout (0.51) and increased her performance using DIR maps to 0.80. However, specificity declined from 0.90 to 0.79. This finding indicates that our DIR subtraction map improves the visual lesion analysis for patients with MS, regardless of the reader's imaging analysis experience. However, readers with no prior experience in brain image analysis in general cannot deliver satisfying discrimination between MS lesions and artefacts even when our DIR subtraction map is used. Training in imaging is necessary for image analysis even with help of computational tools. The amount of prior experience for a sufficient analysis quality with help of subtraction images is a question for further research.

3 Prediction of conversion from CIS to MS

The content of this chapter was published in “Predicting conversion from clinically isolated syndrome to multiple sclerosis – An imaging-based machine learning approach*”* in *NeuroImage: Clinical* on 5th November 2018 (H. Zhang et al., 2018).

3.1 Methods

3.1.1 Subjects

We included 84 patients from the MS database of the Department of Neurology of the *Klinikum Rechts der Isar* for this retrospective observational study. All patients initially presented with CIS; thus, not fulfilling the 2010 McDonald criteria. All patients received a baseline MRI scan between 2009 and 2013. Subsequently, the patients were followed up for at least three years, and the endpoint of the study was set three years after the initial scan. Conversion to MS was defined according to the McDonald criteria 2010; therefore, including the radiological occurrence of a new lesion and clinical proof of DIT by a second clinical attack.

3.1.2 MRI acquisition

All MR images were acquired using a 3 tesla MR scanner (Achieva, Philips Healthcare, Best, the Netherlands). All MR scans contained a 3D FLAIR sequence and a 3D T1w sequence used for this study. The imaging parameters are stated in the Appendix on page XXII.

3.1.3 Image processing and lesion segmentation

The FLAIR and T1w images were processed in MATLAB 9.1. A custom-built script was used to co-register the FLAIR and T1w image of each patient by performing a rigid registration using the SPM 12 package for MATLAB. The T1w image was set as *reference*, whereas the FLAIR image was set as *source*. Then, two sets of segmentation masks were generated.

One set of lesion masks was acquired using the Lesion Segmentation Tool 2.0.1 (LST), which was designed for the SPM package. Lesions were segmented using the lesion growth algorithm (LGA), which calculates lesion belief masks from co-registered T1w and FLAIR images. Later, these maps are thresholded with a pre-chosen threshold (κ). The optimal initial threshold was determined for the in-house MR scanners at $\kappa = 0.3$. Following thresholding, an initial binary lesion map was attained.

The other segmentation mask set was acquired by performing a computer-assisted manual segmentation with BrainSeg3D, which is based on the software Seg3D. The citation of the software can be found in the Appendix on page XXVII. An experienced neuroradiologist of six years (Paul Eichinger) and a medical student in the fifth year (Haike Zhang) were blinded to the clinical information and then segmented all lesions in the baseline MRI scans independently. Therefore, the segmentation assistant with the brightness separation

function (Lesjak et al., 2018) was applied to axial reformations of the FLAIR images. A two-dimensional field containing the target lesion's full area was marked manually in the viewed brain slice. The precise borders of the lesion were delineated automatically and confirmed by visual inspection (Figure 7). Manual readjustment of the segmentation was possible and performed if necessary. Later, segmentation masks were chosen from the mask selection of both viewers by consensus.

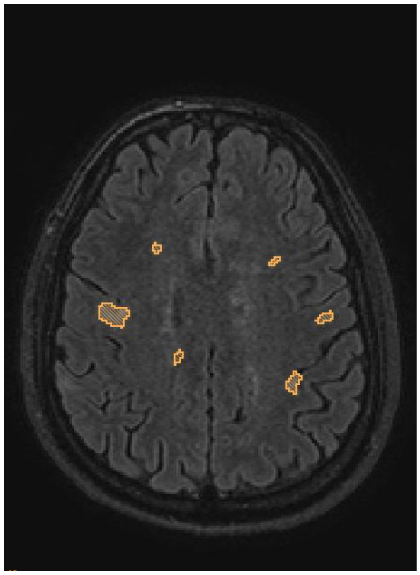

Figure 7 Example of an axial slice with overlaid lesion mask of a patient from the cohort

3.1.4 Random forest model

We calculated the parameters of interest in the programming language Python 3.8.1 using the packages *nibabel*, *NumPy*, *scikit-image*, and *SciPy*. The citation of the packages can be found in the Appendix on page XXVIII and XXIX.

The *total number of lesions* and the *total lesion volume* per patient were calculated using the co-registered FLAIR and T1w images for each patient. Then, for each lesion in each patient, *single lesion volume*, intensity features, and shape features were calculated.

Intensity features included *skewness*, *kurtosis,* and *entropy* of intensity histograms. Shape features included *surface area* (A), *sphericity* (S), and *surface-volume-ratio* (SVR). The *lesion volume* (V) was calculated as follows:

$$V = n^* V_n$$

V_n volume of one voxel, n number of voxels in the lesion

We used the *marching cubes algorithm* (Lorensen & Cline, 1987) that is implemented in *skimage* to approximate the lesion *surface area* (A). Sphericity, in general, is defined as the ratio of the surface area of a sphere of equal volume to the body's surface (Wadell, 1935). The following formula was used to calculate the *sphericity* (S):

$$S = \frac{\sqrt[3]{36\pi V^2}}{A}$$

A surface area, V volume

Subsequently, we calculated the descriptive statistics for the features of each patient describing volume, intensity, and shape because the RF algorithm requires a feature vector of the same length for each patient, but lesion numbers varied between the patients. The statistics included the minimum, maximum, mean, and standard deviation of each feature concerning all lesions in each patient, because only averaging across all patients' lesions would neglect the lesions' heterogenic information. The *total lesion volume* and *lesion count* were included as additional vector elements.

The calculated uniform feature vectors were used as input for the RF algorithm. Analysis was conducted in R 3.4.4 using the package *obliqueRF* to perform three classification models. The first model was based on intensity features, the second model on shape features, and the third model on both intensity and shape features.

All three models included the features of total *lesion count* and *lesion volume*. The hyperparameters *mtry* (number of variables tested in each node) and *ntree* (number of trees generated) were optimized on the *out-of-bag error* (Breiman, 2001).

3.1.5 Reference standard for prediction

A second predictive model was obtained based on the DIS criterion according to the 2010 McDonald criteria, analogous to Filippi et al. (2018). This model predicts the conversion in the baseline scan when DIS is present, whereas a lack of this criterion predicts non-conversion. This model functioned as a benchmark for the RF model.

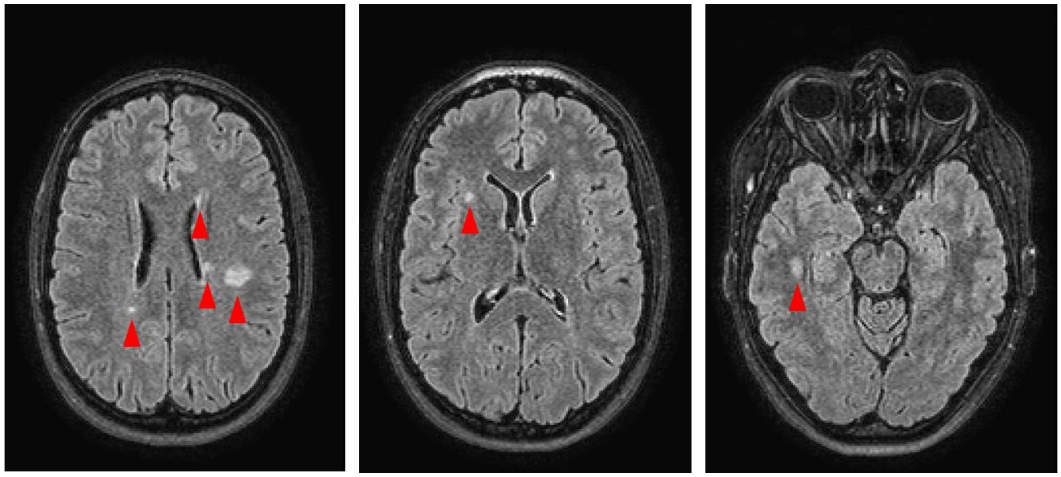

Figure 8 Example of a CIS patient from the cohort fulfilling DIS according to the 2010 McDonald criteria by presenting several lesions at a time without evidence of DIT
Lesions marked by red triangles.

3.1.6 Data analysis

Three-fold cross-validation using the *scikit-learn* package was performed to validate the model's performance. For this validation method, the study collective set was randomly split into three subsets, called folds, of approximately equal size. The first fold was treated as a validation fold. The other two folds were used to train the RF model and then tested on the validation fold. This procedure was performed three times, so every fold was used as a validation set once and every subject was within the validation fold once. The overall performance was determined by calculating the average performance of the three folds.

We applied a bootstrapping approach with 100 iterations to calculate *the feature importance* scores. The most important features for the classification were identified by inspecting each feature's relative contribution to the model. The *feature importance* counts how often a variable was considered relevant when chosen for a split at a node. The factor is calculated by a logistic regression model employed at every node. The importance value increases by one if a variable leads to a logistic regression model with $p < 0.05$ at a node, and decreases by one otherwise.

3.1.7 Statistical analysis

Statistical analysis was performed using MATLAB 9.1 and IBM SPSS Statistics 24. Demographic data of the patients were compared between the groups of converters and non-converters. For gender and age, a Pearson chi-square test and a two-tailed t-test were applied, respectively. The EDSS values were compared using Mann-Whitney U tests, and the comparison between the EDSS value at baseline and after three years within the groups was compared by a Wilcoxon signed-rank-sum test. A Mann-Whitney U-test compared the mean lesion volume of all the lesions in the two groups.

The results from the defined prediction models were expressed as confusion matrices. Accuracy, sensitivity, specificity, PPV, and NPV were selected as statistical measures. Since the conversion rate of 79% is high, balanced accuracy was also calculated to improve the assessment of the model performance with the posterior probability interval for $\alpha = 0.05$ using the MATLAB tools provided by Brodersen et al. (2010) and the diagnostic odds ratio (DOR) (Glas et al., 2013).

The confidence intervals for accuracy, sensitivity, and specificity were calculated as Clopper-Pearson confidence intervals. The confidence intervals for PPVs and NPVs were calculated as standard logit confidence intervals (Mercaldo, Lau, & Zhou, 2007). The confidence interval for the DOR was calculated according to Glas et al. (2003). The RF classifier's performance was compared with the prediction based on the DIS criterion by using an exact McNemar’s test.

The correlation between the three most important shape features for the RF model was calculated as Pearson correlation coefficient (Pearson’s r). The correlation between *minimum sphericity* and *minimum SVR* was calculated by Spearman rank correlation (Spearman’s rho).

The following figures (9–14) were created using the *ggplot2* package in R 3.4.4. The confusion matrices (Table 7) were created using the *crate* function in R.3.4.4.

3.2 Results

3.2.1 Subjects

Our observational cohort included 84 CIS patients: 69.0% (n = 58) were female and 31.0% (n = 26) were male. The F:M ratio was 2.2:1. The distribution of gender and age for our cohort is illustrated in Figure 9. 81 patients received no therapy before their baseline scans were acquired. One person received interferon-beta, one received steroids, and one received plasmapheresis before the baseline scan.

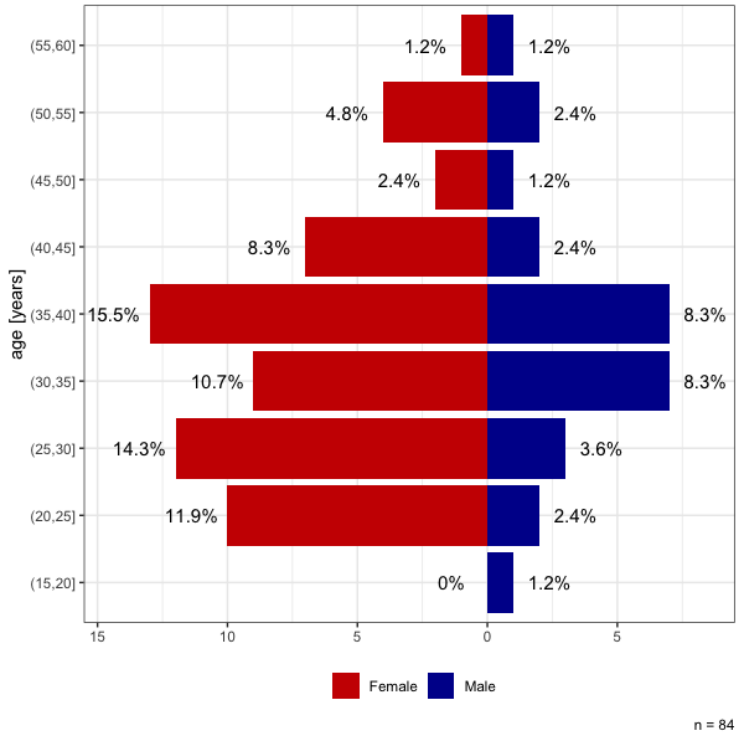

Figure 9 Age and gender distribution in our cohort

After three years, 66 patients (78.5%) converted to MS, hereafter referred as converters, and 18 stayed in the CIS course (21.5%), hereafter referred to as non-converters. Out of the converters, 33 persons (50.0%) suffered a second clinical attack defining CDMS, whereas the other 33 converters fulfilled the radiological McDonald criteria of DIT, defining RDMS (Figure 10).

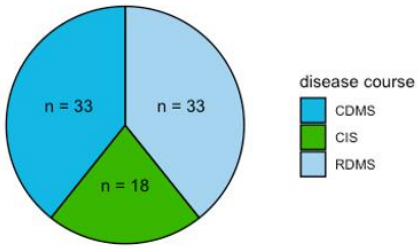

Figure 10 Distribution of disease course after follow-up in our cohort

Although more patients in our cohort were female, there was no significant difference between converters and non-converters regarding gender. The same applies to age and the EDSS at baseline (Table 6). Figure 11 depicts the comparison of the EDSS between the converters and non-converters, and Figure 12 depicts the comparison in age. 51 patients had a change in the EDSS during observation: in 25 patients, the EDSS improved, whereas in 26 patients, the EDSS deteriorated. There was no significant change in the EDSS within the groups during the follow-up (p = 0.87 and p = 0.29; Wilcoxon signed-rank-sum test). A significant difference in *mean lesion volume* between the converters and non-converters was found (Table 6).

Table 6 Patients' characteristics of converters compared with non-converters

	Non-converter	Converter	Tests
Gender	7 men 11 women	18 men 47 women	Pearson chi-square, p = 0.411
Age at onset	mean = 44.44 STD = 11.21	mean = 41.89 STD = 8.808	2-tailed t-Test, p = 0.308
EDSS at baseline	median = 1 range 0–2.5	median = 1 range 0–6	Mann-Whitney U-test, p = 0.560
EDSS after three years	median = 0 range 0–2.5	median = 1 range 0–6.5	Mann-Whitney U-test, p = 0.0800
Mean lesion volume [mm³]	mean = 71 range 22–314	mean = 135 range 22–671	Mann-Whitney U-test, p = 0.0013

STD: standard deviation

Source: Table based on Zhang et al. (2018, Tab. 1)

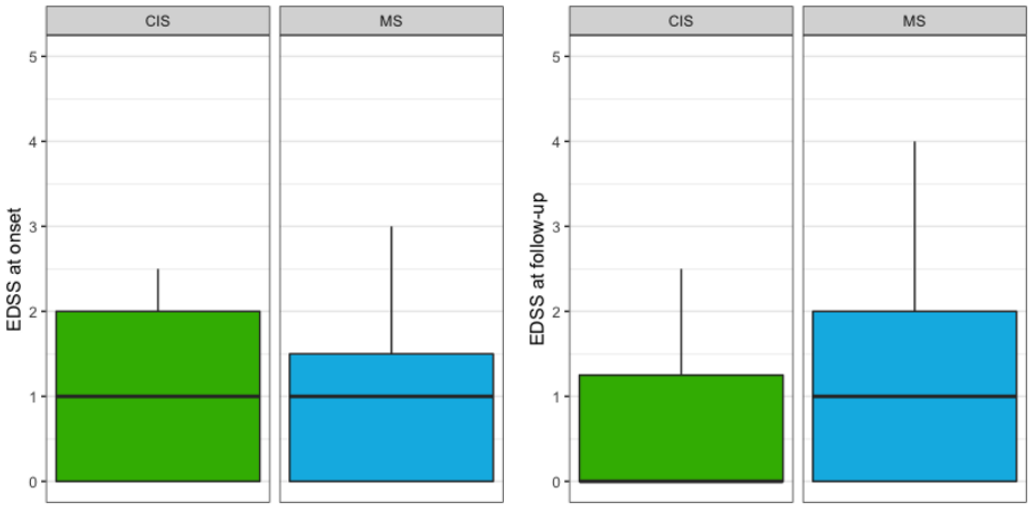

Figure 11 Comparison of the EDSS spread between the non-converter (CIS) and converter group (MS) at baseline and at follow-up

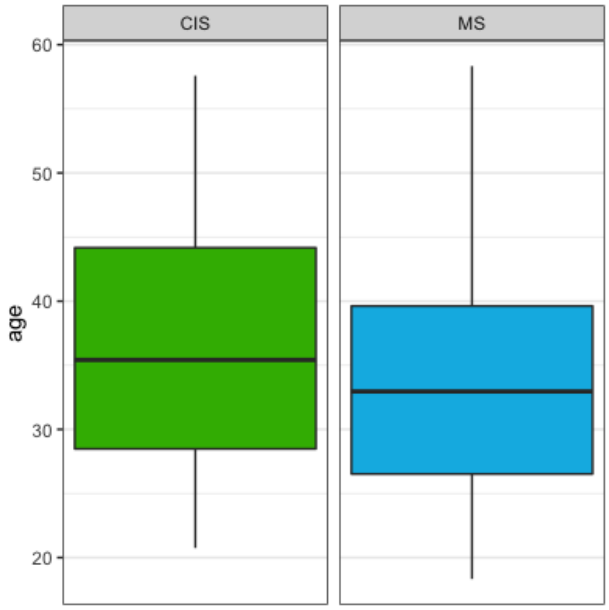

Figure 12 Comparison of the age spread between non-converters (CIS) course and converters (MS)

3.2.2 Classification outcome and accuracy

The confusion matrices (Table 7) display the numbers of converters and non-converters who were predicted correctly or wrongly by each model. The gold standard prediction model using DIS from the 2010 McDonald criteria was used as the benchmark.

The RF algorithm was performed three times: with shape features calculated from computer-assisted (BrainSeg3D) segmentation mask; with shape features calculated from fully automatically (LST for SPM) generated segmentation masks; and with intensity features from computer-assisted (BrainSeg3D) segmentation masks. An algorithm with intensity features from fully automatically generated segmentation masks was not performed due to unsatisfying results of the model.

Table 7 Confusion matrices for predictions from McDonald criteria 2010

a) McDonald 2010 (DIS)	**Non-conversion**	**Conversion**
Predicted non-conversion	4	4
Predicted conversion	14	62
b) Intensity-based model		
Predicted non-conversion	11	25
Predicted conversion	7	41
c) Shape-based model		
Predicted non-conversion	9	4
Predicted conversion	9	62
d) Shape-based model (LST)		
Predicted non-conversion	6	3
Predicted conversion	12	63

DIS dissemination in space. *LST* lesion segmentation tool

Source: Table based on Zhang et al. (2018, Tab. 2)

The shape-based model with features derived from the LST segmentation masks achieved the best prediction performance of the three developed classifiers, with a balanced accuracy of 0.72 and a DOR of 15.50. This model achieved a significantly better predictive accuracy than the McDonald criteria 2010 (McNemar's test, $p = 0.03$).

We compared both segmentation methods by calculating the correlation between the values of the three most important features. These features were explored according to their relevance to the final choice of the classifier, which is referred to in the following section. We found a high correlation between each of these three features: mean lesion volume (Pearson's $r = 0.79$, $p < 0.0001$), minimum sphericity (Pearson's $r = 0.42$, $p < 0.0001$), and minimum SVR (Pearson's $r = 0.88$, $p < 0.0001$).

Table 8 Statistical measures calculated from the confusion matrices in Table 7

	Mc Donald 2010 (DIS)	Intensity-based model	Shape-based model	Shape-based model (LST)
Accuracy	0.79 (0.68–0.87)	0.62 (0.51–0.72)	0.85 (0.75–0.91)	0.82 (0.72–0.90)
Sensitivity	0.94 (0.85–0.98)	0.62 (0.49–0.74)	0.94 (0.85–0.98)	0.95 (0.87–0.99)
Specificity	0.22 (0.06–0.48)	0.61 (0.36–0.83)	0.50 (0.26–0.74)	0.33 (0.13–0.59)
PPV	0.81 (0.77–0.85)	0.85 (0.76–0.92)	0.87 (0.81–0.91)	0.84 (0.79–0.87)
NPV	0.50 (0.22–0.78)	0.31 (0.21–0.42)	0.69 (0.44–0.87)	0.67 (0.36–0.88)
Balanced Accuracy	0.58 (0.50–0.70)	0.62 (0.49–0.72)	0.72 (0.60–0.82)	0.64 (0.54–0.76)
DOR	4.43 (0.99–19.89)	2.58 (0.88–7.51)	15.50 (3.93–60.98)	10.50 (2.30–47.87)

DOR diagnostic odds ratio. *PPV* positive predictive value. *NPV* negative predictive value. *LST* lesion segmentation tool. 95%-confidence intervals in parenthesis, except for balanced accuracy the posterior probability interval for the level 0.05 is given.

Source: Table based on Zhang et al. (2018, Tab. 3)

3.2.3 Three most relevant features for classification

To explore the relative influence of shape features for the classifier, the shape features' importance scores were calculated using a bootstrapping approach (Figure 13). *Mean lesion volume*, *minimum sphericity*, and *minimum SVR* had the highest importance for the RF model's final vote. *Minimum sphericity* and *minimum SVR* displayed an expected significant positive correlation (Spearman's rho = 0.53, $p < 0.001$). *Lesion count* was not of high importance.

Regarding the distribution of lesion features between the classes, the model found that the converter group's lesions had a higher *mean lesion volume* and smaller *minimum sphericity* and smaller *minimum SVR* (Figure 14). Figurately, these lesions appeared, on average, larger and less round. Illustrative examples for this feature distribution are in Figure 15.

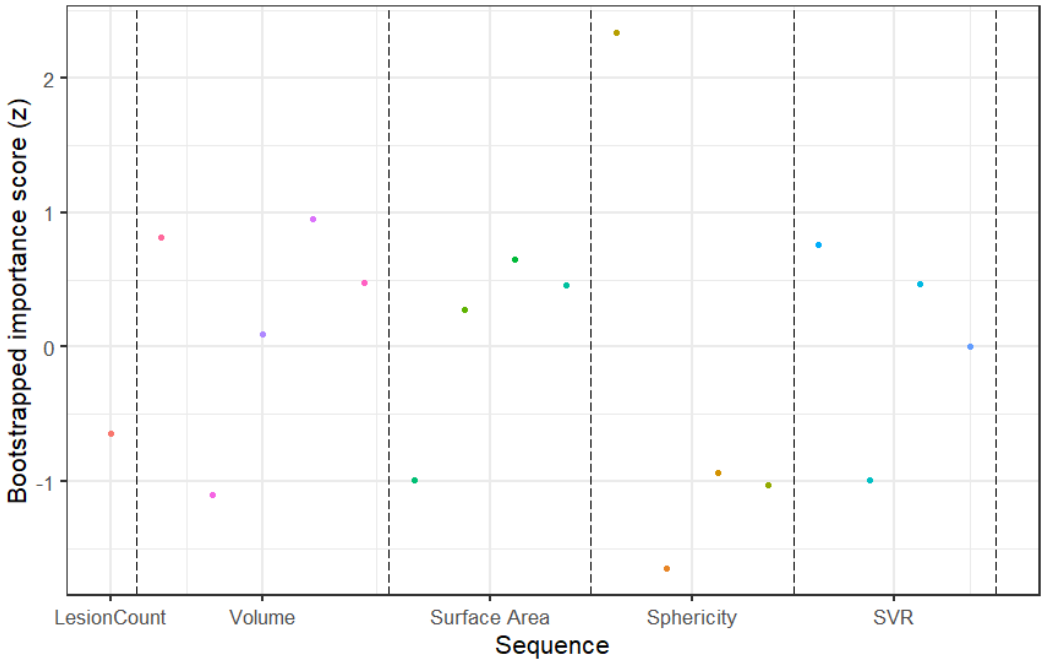

Figure 13 Bootstrapped feature importance
Each dot represents one shape feature. Dots represent from left to right: minimum, maximum, mean, and standard deviation. For *volume* the additional fifth dot at first place (from left to right) represents total lesion volume. The higher the value, the more important the feature is.
Reference: The slight differences to the analogous figure in the paper Zhang et al. (2018, Fig. 1a) are explained in the Chapter 3.3.6 on page 44.

Source: Figure based on Zhang et al. (2018, Fig. 1a)

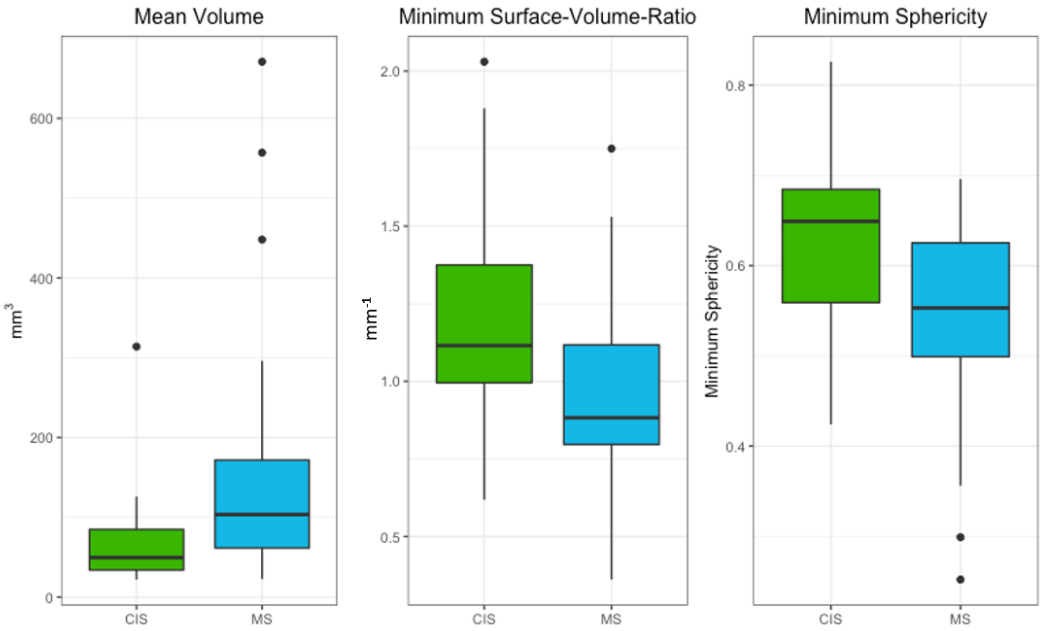

Figure 14 Comparison of three most important lesion features for the shape-based RF model between converters (CIS) and non-converters (MS)
The exact numbers are in the Appendix on page XXIII.

Source: Figure based on Zhang et al. (2018, Fig. 1b–d)

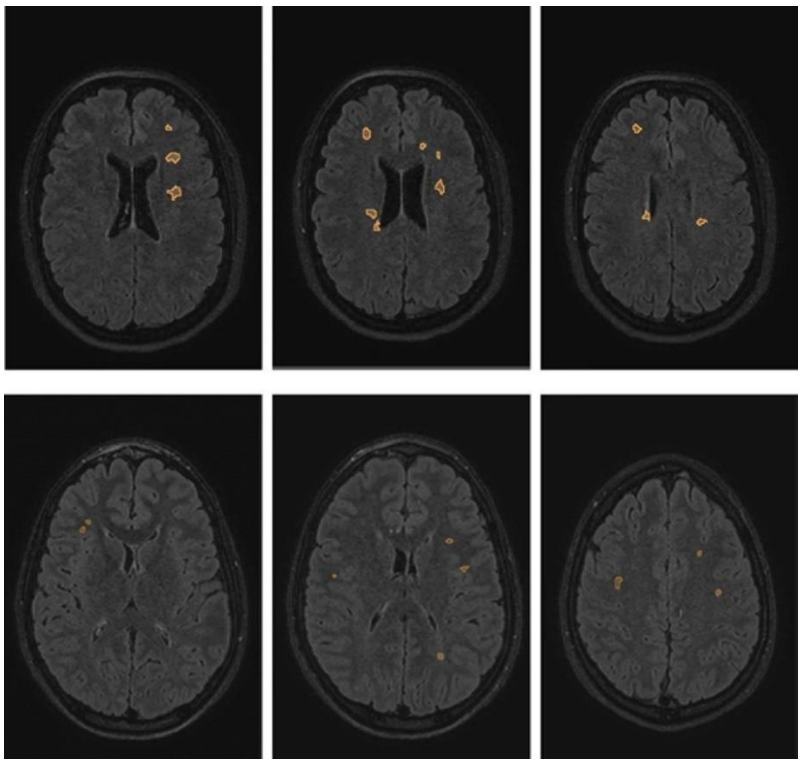

Figure 15 Illustrative example images with overlaid lesion masks of two patients from our cohort
The upper row displays a scan of a converter with prominently larger and less-round lesions, whereas the lower row displays a scan of a non-converter with smaller and rounder lesions.

Source: Zhang et al. (2018, Fig. 2)

3.3 Discussion

In this second project, we developed an RF model to predict the possible conversion of CIS patients to MS. The classifier was based on lesion features in the baseline MRI scan. Shape features demonstrated high discriminative potential, whereas intensity features did not provide a satisfying contribution to the classification. We segmented the lesions by two methods, and both proved useful in constructing an accurate prediction model. Our best performing classifier relies on shape features and achieved a better prediction accuracy than the gold standard, the DIS criterion from the 2010 McDonald criteria. The shape features that contributed the most to the classification were *mean lesion volume*, *minimum sphericity*, and *minimum SVR*. Comparing these features between the groups revealed that converters presented with larger and less-round lesions in their baseline scan.

3.3.1 Predictive accuracy of the random forest model

Our RF classifier predicted the conversion to MS more accurately than using the DIS criterion from the 2010 McDonald criteria. The shape-based RF model, with features derived from manual segmentation, achieved a prediction sensitivity of 0.94 and a specificity of 0.50, whereas the DIS criterion achieved 0.94 and 0.22, respectively. The measures of *balanced accuracy* and *DOR* are designed for unbalanced cohorts and were calculated to consider the imbalanced conversion rate. The RF classifier achieved a balanced accuracy of 0.72 and a DOR of 15.50, as opposed to 0.58 and 4.43 for the DIS criterion.

The prediction performance for the 2010 McDonald criteria closely matches that from a study that investigated a large cohort, which found a sensitivity of 0.92 and a specificity of 0.33 (Filippi et al., 2018). Other studies that applied the 2010 McDonald criteria for the prediction of conversion reported a sensitivity of 0.86 and specificity of 0.65 for a Spanish cohort of 67 patients (Gómez-Moreno, Díaz-Sánchez, & Ramos-González, 2012), and a sensitivity of 0.71 and a specificity of 0.63 for a Korean cohort of 170 patients (Hyun et al., 2017). The difference may be explained by the different ethnicities of these cohorts, as our cohort mainly consisted mainly of Caucasian patients. Filippi (2018) had the largest cohort, with 368 patients, and the most similar results.

In this project, we focused mainly on imaging features that describe lesion properties. In an earlier study with a similar aim, basic clinical features were combined with lesion localization features and basic intensity features (Wottschel et al., 2015). This classifier predicted the conversion to CDMS with a sensitivity of 0.77 and a specificity of 0.66, one year after the baseline scan (Wottschel et al., 2015). In contrast to our classifier, support vector machines were used as a machine-learning method (Wottschel et al., 2015).

Other studies have used advanced MR techniques to include imaging biomarkers in their prediction model. Kitzler et al. (2018) measured myelin water fraction in white matter and

found that myelin loss was crucial for conversion prediction. Ion-Mărgineanu et al. (2017) also combined clinical data with lesion load. Additionally, they used MR spectroscopy to extract magnetic resonance metabolic features, which allowed differentiation between MS forms.

Using the whole range of technical possibilities can provide new insights into the radiological assessment of MS patients. However, these methods are rarely used in the clinical setting. In contrast, our RF models rely on FLAIR and T1w images, which are part of the standard imaging protocol for MS patients.

3.3.2 Discriminative contribution of shape features

The best prediction was achieved by our RF which is based solely on shape features. The calculation of the individual features' importance scores revealed that lesion load is not of high importance for the classification.

Several studies have found that the lesion number in CIS patients' baseline images is associated with an increased risk of conversion to CDMS (Alroughani et al., 2012; Brex et al., 2001; Dalton, Brex, Miszkiel, et al., 2002). However, our finding regarding the low importance of lesion load is in line with a meta-analysis that concludes that the presentation of an abnormal T2w image alone was associated with the risk of conversion, regardless of the lesion load (W. Y. Zhang & Hou, 2013).

Intensity features also could not contribute to a successful prediction model. The model based on intensity features delivered a worse prediction accuracy than all other investigated models. Although intensity features have long been used in image analysis, they are, unlike shape features, variable and poorly comparable across different scanners. However, the robustness of shape features is limited by the issue of spatial resolution of a scan: The voxel size of the scan determines the minimum depictable diameter of a lesion and can therefore influence the shape of small lesions. To keep our high spatial resolution 3D sequences unaffected from these effects, we followed the minimum size threshold as proposed in the MAGNIMS and McDonald criteria (Filippi et al., 2016; Polman et al., 2011). The extraction of reliable shape features from small lesions in 2D images may be impossible due to this described limitation.

3.3.3 Lesion shape useful in differentiating MS lesions from its mimics

In this study, we identified the most important features for the classifier's final decision: *mean lesion volume*, *minimum sphericity*, and *minimum SVR*. Figurately, lesions found in patients of the converter group were, on average, larger, and the lesion that was least spherical in the patients' scans contributed the most to the prediction model.

Our findings regarding difference in size and shape may reflect the specific pathophysiological origin of MS lesions. These lesions correspond to axonal and neuronal injuries which mainly occur along veins (Tallantyre et al., 2008; Tan et al., 2000). A vein positioned centrally in white matter lesions of the brain is called a central vein sign and distinguishes lesions caused by MS from white matter hyperintensities caused by other neurological phenotypes, such as chronic inflammatory vasculopathy (Cortese et al., 2018; Maggi et al., 2020). The central vein sign is recommended as a potential MS-specific imaging marker to reduce misdiagnosis risk (Margareta A. Clarke et al., 2020; Maggi et al., 2020; Sinnecker et al., 2019). A similar conclusion was found by M. A. Clarke et al. (2020), who investigated the prediction performance of central vein signs in baseline MRI scans for CIS patients to convert to MS and found that 3 lesions with central vein sign delivered a prediction sensitivity of 0.70 and a specificity of 0.86.

The inflammatory origin of lesions of converters might explain the more elongated and less-spherical shape of their lesions. The typical *Dawson finger configurations* of MS lesions reflect the elongated lesion shape and explain how (neuro-)radiologists incorporate shape information in their lesion classification. Elongated and less spherical lesion properties may translate as these configurations for the RF algorithm.

In contrast, lesions in non-converters are probably caused by a pathomechanism that differs substantially from the chronic demyelinating events with persistent inflammation in MS, and therefore, present a different shape than lesions found in converters. Newton et al. (2017) demonstrated that MS lesions present as more elongated and with a more complex surface morphology than non-specific white matter lesions. This stands in line with our finding of bigger and less round lesions being predictive of future conversion to MS. Further research, regarding the predictive value of CIS patients' lesion shape could be promising.

3.3.4 Comparability of segmentation types regarding classification outcome

One of the strengths of our study is that two methods of lesion segmentation were executed, and both methods yielded shape features to construct an accurate classifier. The computer-assisted segmentation classifier achieved a sensitivity of 0.94, a specificity of 0.50, and a balanced accuracy of 0.72. The classifier using fully automatically generated segmentation masks achieved 0.95, 0.33, and 0.64, respectively. The three most important features were calculated from both segmentation methods and displayed a high correlation.

Classifiers based on shape features calculated from both techniques can significantly distinguish between converters and non-converters. Although high variability of MS lesion appearance is a problem for satisfying automated segmentation performance, our findings indicate that the existing differences to manually segmented masks play a subordinate role for further processing. This independence of the segmentation method enables the time-

efficient transfer of our proposed technique to larger cohorts, for which quick lesion segmentation is required.

Our comparable prediction result between fully automatically generated segmentation and computer-assisted segmentation is limited, and our segmentation methods and subsequent prediction model should be tested on datasets acquired from other neuroradiological departments that use MRI scanners of a different brand to test the robustness of the segmentation method for the classification. Additionally, it would be interesting to test our classifier with inputs segmented by different automatic algorithms.

3.3.5 High conversion rate due to the inclusion of radiological criteria

After three years, 79% of patients (n = 66) in the investigated cohort (n = 84) converted to MS. Multiple sclerosis was determined according to the 2010 McDonald criteria, which allow the diagnosis of MS by radiological criteria in the absence of clinical proof. This determination of diagnosis differs from other studies in which only the occurrence of a second clinical attack is considered as conversion to MS; thus, counting only those cases as conversion, which meet the definition of CDMS.

Therefore, we reported a remarkably higher conversion rate of 79% for our cohort than previous studies: Lo et al. (2009) described a conversion of 46.9% out of 64 patients within a mean of 9.5 months; Ruet et al. (2014) described 35.2% out of 505 patients within a median follow-up of 44.6 months; Wottschel et al. (2015b) described 44% out of 74 patients within three years; and Filippi et al. (2018) described 51% out of 368 patients within a median follow-up of 50 months. However, the rate of CDMS in our cohort is 39%, which is comparable to these studies.

Ruet et al. (2014) reported a 77.6% conversion rate in a study with 505 patients. In 42.4% of patients, the conversion was based solely on imaging findings, and 35.2% of patients had a second clinical attack, delivering a similar rate to our study. Recent studies that included RDMS in their conversion criteria have also found comparable conversion rates of 84.7% (Gaetani et al., 2017) and 74% (Gómez-Moreno et al., 2012).

Studies reporting long-term CDMS conversion rates also display a higher percentage: 80% (Fisniku et al., 2008) and 61% (Chard et al., 2011) both within two decades. Additionally, Chard et al. (2011) found that only 11–15% of patients with RDMS exhibited no second clinical attack within two decades. Hence, we regarded our choice as sensible to define MS according to the 2010 McDonald criteria for a prediction model used to evaluate CIS patients' long-term prognosis.

It cannot be excluded that some patients of the non-converter group may develop MS after our follow-up period. However, our conversion rates are consistent with other studies, as the comparisons above indicate. Furthermore, according to the predominance of

converters in our cohort, hyperplane weighting bias toward the larger group can occur. The statistical measures that balanced accuracy and DOR were calculated to compensate for this effect. It is desirable to test the RF algorithm on a larger cohort with better-balanced groups.

3.3.6 Machine learning limitations

Machine learning models such as the RF method are often complicated and predisposed to overfitting. Due to the small cohort size, our training data was not abundant. Thus, independent training with unused data was not possible, and we employed three-fold cross-validation to validate the classifier's prediction. This validation method comes with a positive bias in the absolute accuracy, possibly leading to a lower prediction accuracy for unseen data. Validation in an independent and larger cohort, even from another hospital with data acquired from a different MRI scanner, would be preferable. The k-fold cross-validation estimates the prediction accuracy more realistically than our method (James, Witten, Hastie, & Tibshirani, n.d.), and the training and validation of our classifier with this method would be desirable.

Another weakness of supervised learning, which also applies to our RF models, is that the features, which the classifier will be trained on, must be selected beforehand. The size of our dataset also precluded extensive testing of large feature vectors. Since we aimed to assess intracranial lesion characteristics' contribution to classification performance, our feature analysis was limited to the predefined choice of those shape and intensity features described in the previous chapters. Other clinical parameters with predictive value, such as intrathecal synthesis of oligoclonal bands (Mar Tintoré et al., 2008), age (Ruet et al., 2014), inflammatory cerebrospinal fluid (Ruet et al., 2014), gender, genetic preposition (Kelly et al., 1993; Tossberg et al., 2013), ventricular enlargement (C. M. Dalton et al., 2002), and gray matter atrophy (Di Filippo et al., 2010; Zivadinov et al., 2013) were not included in our prediction algorithm even though they are associated with a higher risk of conversion. It remains unclear how the combination of features of radiological and clinical nature could contribute to the prediction accuracy.

Furthermore, only intracranial lesions were segmented and used for our purpose. The improvement in prediction accuracy when spinal lesions are integrated into prediction models remains to be investigated in future studies. The same applies to other promising subdivisions of ML (e.g. neuronal networks).

The values depicted in the importance score plot (Figure 13) on page 38 slightly differ from those depicted in the analogous plot in the paper (Zhang et al., 2018, Fig. 1a). This is because the structure of the resampling procedure is not determined in advance. This leads to slight shifts for the dots in the figure, but the core statement and distribution of the feature importance score remain.

4 Conclusion

This present thesis aimed to develop and assess computational tools that relieve human drawbacks in image analysis for patients with MS. We developed two tools that are ready to use in clinical routine in two separate projects.

In the first project, our proposed lesion analysis method improved the visualization of change in lesion load and lesion size in follow-up MR images by erasing radiological-stable status. The proposed method sped up the analysis process and improved the detection accuracy significantly. Since MS lesions are small and can be widely distributed throughout the CNS, a tool to relieve this cumbersome analysis is helpful. Lesion count and information about the occurrence of new or enlarged lesions are crucial factors in disease monitoring and for determining further therapy strategy, amongst other factors. For the second project, our proposed RF model predicted conversion of patients with CIS to MS based on radiological lesion data in their baseline MRI scan more accurately than by the McDonald criteria of 2010. Lesion shape parameters proved to have a high discriminative potential in classifying converters and non-converters on a three-year time scale. A more accurate prediction at the early disease stage can help to identify patients who benefit from early treatment and reduce unnecessary treatments, with their side effects, for patients who do not. The segmentation method did not significantly affect the prediction result, raising the possibility of replacing manual segmentation for this purpose in clinical routine. Our cohort size did not allow a deep-learning model, which offers far more possibilities for investigating a wider range of features with no previous selection. Such an investigation, with a multicentre background, including more clinical and paraclinical features for a more differentiated prediction of the expected disease progression, would be interesting.

This thesis is an example of how computational methods improve the interpretation of imaging data whose full value eludes human vision. Imaging is progressively taking a bigger account in the diagnostics and further therapy strategy of patients. To ensure that the utility and diagnostic value is growing with the amount of data, such clinically relevant algorithms must be integrated into radiologists' workflow. This requires infrastructure, implementation of guidelines, and the system's ability to update. A preferable aim for the near future, regarding imaging for patients with MS, is a continuous diagnostic pipeline: starting from processing MRI sequences and arranging clinical data to an automatically constructed individual report and prognosis for each patient. Ultimately, more work will be needed to achieve such a pipeline.

5 Bibliography

Alroughani, R., Al Hashel, J., Lamdhade, S., & Ahmed, S. F. (2012). Predictors of Conversion to Multiple Sclerosis in Patients with Clinical Isolated Syndrome Using the 2010 Revised McDonald Criteria. *ISRN Neurology*. https://doi.org/10.5402/2012/792192

Ashton, E. A., Takahashi, C., Berg, M. J., Goodman, A., Totterman, S., & Ekholm, S. (2003). Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.10258

Barkhof, F., Filippi, M., Miller, D. H., Scheltens, P., Campi, A., Polman, C. H., … Valk, J. (1997). Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain*, *120*(11), 2059–2069. https://doi.org/10.1093/brain/120.11.2059

Battaglini, M., Rossi, F., Grove, R. A., Stromillo, M. L., Whitcher, B., Matthews, P. M., & De Stefano, N. (2014). Automated identification of brain new lesions in multiple sclerosis using subtraction images. *Journal of Magnetic Resonance Imaging*, *39*(6), 1543–1549. https://doi.org/10.1002/jmri.24293

Breiman, L. (2001). Random forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324

Brex, P. A., Ciccarelli, O., O'Riordan, J. I., Sailer, M., Thompson, A. J., & Miller, D. H. (2002). A Longitudinal Study of Abnormalities on MRI and Disability from Multiple Sclerosis. *New England Journal of Medicine*, *346*(3), 158–164. https://doi.org/10.1056/NEJMoa011341

Brex, P. A., Miszkiel, K. A., O'Riordan, J. I., Plant, G. T., Moseley, I. F., Thompson, A. J., & Miller, D. H. (2001). Assessing the risk of early multiple sclerosis in patients with clinically isolated syndromes: the role of a follow up MRI. *J Neurol Neurosurg Psychiatry*, *70*(3), 390–393. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11181865

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2010.764

Brodsky, M., Nazarian, S., Orengo-Nania, S., Hutton, G. J., Buckley, E. G., Massey, E. W., … Smith, C. H. (2008). Multiple sclerosis risk after optic neuritis: Final optic neuritis treatment trial follow-up. *Archives of Neurology*. https://doi.org/10.1001/archneur.65.6.727

Chard, D. T., Dalton, C. M., Swanton, J., Fisniku, L. K., Miszkiel, K. A., Thompson, A. J.,

… Miller, D. H. (2011). MRI only conversion to multiple sclerosis following a clinically isolated syndrome. *Journal of Neurology, Neurosurgery & Psychiatry*, *82*(2), 176–179. https://doi.org/10.1136/jnnp.2010.208660

Clarke, M. A., Pareto, D., Pessini-Ferreira, L., Arrambide, G., Alberich, M., Crescenzo, F., … Rovira, À. (2020). Value of 3T susceptibility-weighted imaging in the diagnosis of multiple sclerosis. *American Journal of Neuroradiology*, *41*(6), 1001–1008. https://doi.org/10.3174/AJNR.A6547

Clarke, Margareta A., Samaraweera, A. P. R., Falah, Y., Pitiot, A., Allen, C. M., Dineen, R. A., … Evangelou, N. (2020). Single Test to ARrive at Multiple Sclerosis (STAR-MS) diagnosis: A prospective pilot study assessing the accuracy of the central vein sign in predicting multiple sclerosis in cases of diagnostic uncertainty. *Multiple Sclerosis Journal*. https://doi.org/10.1177/1352458519882282

Comi, G., Filippi, M., Barkhof, F., Durelli, L., Edan, G., Fernández, O., … Hommes, O. R. (2001). Effect of early interferon treatment on conversion to definite multiple sclerosis: A randomised study. *Lancet*, *357*(9268), 1576–1582. https://doi.org/10.1016/S0140-6736(00)04725-5

Comi, G., Martinelli, V., Rodegher, M., Moiola, L., Bajenaru, O., Carra, A., … Filippi, M. (2009). Effect of glatiramer acetate on conversion to clinically definite multiple sclerosis in patients with clinically isolated syndrome (PreCISe study): a randomised, double-blind, placebo-controlled trial. *The Lancet*, *374*(9700), 1503–1511. https://doi.org/10.1016/S0140-6736(09)61259-9

Compston, A., & Coles, A. (2008). Multiple sclerosis. *Lancet*, *372*(9648), 1502–1517. https://doi.org/10.1016/S0140-6736(08)61620-7

Cortese, R., Magnollay, L., Tur, C., Abdel-Aziz, K., Jacob, A., De Angelis, F., … Ciccarelli, O. (2018). Value of the central vein sign at 3T to differentiate MS from seropositive NMOSD. *Neurology*, 10.1212/WNL.0000000000005256. https://doi.org/10.1212/WNL.0000000000005256

Dalton, C. M., Brex, P. A., Jenkins, R., Fox, N. C., Miszkiel, K. A., Crum, W. R., … Miller, D. H. (2002). Progressive ventricular enlargement in patients with clinically isolated syndromes is associated with the early development of multiple sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *73*, 141–147. https://doi.org/10.1136/jnnp.73.2.141

Dalton, C. M., Brex, P. A., Miszkiel, K. A., Hickman, S. J., MacManus, D. G., Plant, G. T., … Miller, D. H. (2002). Application of the new Mcdonald criteria to patients with clinically isolated syndromes suggestive of multiple sclerosis. *Annals of Neurology*, *52*(1), 47–53. https://doi.org/10.1002/ana.10240

Di Filippo, M., Anderson, V. M., Altmann, D. R., Swanton, J. K., Plant, G. T., Thompson, A. J., & Miller, D. H. (2010). Brain atrophy and lesion load measures over 1 year relate

to clinical status after 6 years in patients with clinically isolated syndromes. *Journal of Neurology, Neurosurgery and Psychiatry*. https://doi.org/10.1136/jnnp.2009.171769

Egger, C., Opfer, R., Wang, C., Kepp, T., Sormani, M. P., Spies, L., … Schippling, S. (2017). MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? *NeuroImage: Clinical*. https://doi.org/10.1016/j.nicl.2016.11.020

Eichinger, P., Wiestler, H., Zhang, H., Biberacher, V., Kirschke, J. S., Zimmer, C., … Wiestler, B. (2017). A novel imaging technique for better detecting new lesions in multiple sclerosis. *Journal of Neurology*, *264*(9), 1909–1918. https://doi.org/10.1007/s00415-017-8576-y

Elliott, C., Arnold, D. L., Collins, D. L., & Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Transactions on Medical Imaging*, *32*(8), 1490–1503. https://doi.org/10.1109/TMI.2013.2258403

Filippi, M., Horsfield, M. A., Bressi, S., Martinelli, V., Baratti, C., Reganati, P., … Comi, G. (1995). Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis: A comparison of techniques. *Brain*. https://doi.org/10.1093/brain/118.6.1593

Filippi, M., Preziosa, P., Meani, A., Ciccarelli, O., Mesaros, S., Rovira, A., … Rocca, M. A. (2018). Prediction of a multiple sclerosis diagnosis in patients with clinically isolated syndrome using the 2016 MAGNIMS and 2010 McDonald criteria: a retrospective study. *The Lancet Neurology*. https://doi.org/10.1016/S1474-4422(17)30469-6

Filippi, M., Rocca, M. A., Ciccarelli, O., Stefano, D., Evangelou, N., Kappos, L., … Barkhof, F. (2016). MRI Criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol*, *15*(3), 292–303. https://doi.org/10.1016/S1474-4422(15)00393-2.MRI

Fisniku, L. K., Brex, P. A., Altmann, D. R., Miszkiel, K. A., Benton, C. E., Lanyon, R., … Miller, D. H. (2008). Disability and T2 MRI lesions: A 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain*, *131*(3), 808–817. https://doi.org/10.1093/brain/awm329

Gaetani, L., Fanelli, F., Riccucci, I., Eusebi, P., Sarchielli, P., Pozzilli, C., … Di Filippo, M. (2017). High risk of early conversion to multiple sclerosis in clinically isolated syndromes with dissemination in space at baseline. *Journal of the Neurological Sciences*, *379*, 236–240. https://doi.org/10.1016/j.jns.2017.06.008

Gale, C. R., & Martyn, C. N. (1995). Migrant studies in multiple sclerosis. *Progress in Neurobiology*. https://doi.org/10.1016/0301-0082(95)80008-V

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., … Lladó, X. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology*, *56*(5), 363–374.

https://doi.org/10.1007/s00234-014-1343-1

Geurts, J. J. G., Pouwels, P. J. W., Uitdehaag, B. M. J., Polman, C. H., Barkhof, F., & Castelijns, J. A. (2005). Intracortical lesions in multiple sclerosis: Improved detection with 3D double inversion-recovery MR imaging. *Radiology*, *236*(1), 254–260. https://doi.org/10.1148/radiol.2361040450

Giorgio, A., Battaglini, M., Rocca, M. A., De Leucio, A., Absinta, M., Van Schijndel, R., … De Stefano, N. (2013). Location of brain lesions predicts conversion of clinically isolated syndromes to multiple sclerosis. *Neurology*, *80*(3), 234–241. https://doi.org/10.1212/WNL.0b013e31827debeb

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/S0895-4356(03)00177-X

Gómez-Moreno, M., Díaz-Sánchez, M., & Ramos-González, A. (2012). Application of the 2010 McDonald criteria for the diagnosis of multiple sclerosis in a Spanish cohort of patients with clinically isolated syndromes. *Multiple Sclerosis Journal*. https://doi.org/10.1177/1352458511417828

Grossman, R. I., & McGowan, J. C. (1998). Perspectives on multiple sclerosis. *American Journal of Neuroradiology*.

Hajnal, J. V., Saeed, N., Oatridge, A., Williams, E. J., Young, I. R., & Bydder, G. M. (1995). Detection of subtle brain changes using sub voxel registration and subtraction of serial mr images. *Journal of Computer Assisted Tomography*, *19*(5), 677–691. https://doi.org/10.1097/00004728-199509000-00001

Hart, F. M., & Bainbridge, J. (2016). Current and emerging treatment of multiple sclerosis. *Am J Manag Care*.

Hyun, J.-W., Huh, S.-Y., Kim, W., Park, M. S., Ahn, S.-W., Cho, J.-Y., … Kim, H. J. (2017). Evaluation of 2016 MAGNIMS MRI criteria for dissemination in space in patients with a clinically isolated syndrome. *Multiple Sclerosis Journal*, *8*(4), 135245851770674. https://doi.org/10.1177/1352458517706744

Ion-Mărgineanu, A., Kocevar, G., Stamile, C., Sima, D. M., Durand-Dubief, F., Van Huffel, S., & Sappey-Marinier, D. (2017). Machine Learning Approach for Classifying Multiple Sclerosis Courses by Combining Clinical Data with Lesion Loads and Magnetic Resonance Metabolic Features. *Frontiers in Neuroscience*, *11*(JUL), 398. https://doi.org/10.3389/fnins.2017.00398

Jacobs, L. D., Beck, R. W., Simon, J. H., Kinkel, R. P., Brownscheidle, C. M., Murray, T. J., … Sandrock, A. W. (2000). Intramuscular Interferon Beta-1A Therapy Initiated during a First Demyelinating Event in Multiple Sclerosis. *New England Journal of Medicine*, *343*(13), 898–904. https://doi.org/10.1056/NEJM200009283431301

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *Springer Texts in Statistics An Introduction to Statistical Learning*. Retrieved from http://www.springer.com/series/417

Kappos, L., Edan, G., Freedman, M. S., Montalbán, X., Hartung, H. P., Hemmer, B., … Wicklein, E. M. (2016). The 11-year long-term follow-up study from the randomized BENEFIT CIS trial. *Neurology*, *87*(10), 978. https://doi.org/10.1212/WNL.0000000000003078

Kelly, M. A., Cavan, D. A., Penny, M. A., Mijovic, C. H., Jenkins, D., Morrissey, S., … Francis, D. A. (1993). The influence of HLA-DR and -DQ alleles on progression to multiple sclerosis following a clinically isolated syndrome. *Human Immunology*, *37*(3), 185–191. https://doi.org/10.1016/0198-8859(93)90184-3

Khangure, S. R., & Khangure, M. S. (2011). MR Imaging in Multiple Sclerosis: The Accuracy of 3D Double Inversion Recovery at 3 Tesla and the Potential for Single Sequence Imaging. *The Neuroradiology Journal*, *24*(1), 92–99. https://doi.org/10.1177/197140091102400114

Kitzler, H. H., Wahl, H., Eisele, J. C., Kuhn, M., Schmitz-Peiffer, H., Kern, S., … Linn, J. (2018). Multi-component relaxation in clinically isolated syndrome: Lesion myelination may predict multiple sclerosis conversion. *NeuroImage: Clinical*, *20*(May), 61–70. https://doi.org/10.1016/j.nicl.2018.05.034

Koch-Henriksen, N., & Sørensen, P. S. (2010). The changing demographic pattern of multiple sclerosis epidemiology. *The Lancet Neurology*. https://doi.org/10.1016/S1474-4422(10)70064-8

Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology*. https://doi.org/10.1212/WNL.33.11.1444

Kurtzke, J. F. (1993). Epidemiologic evidence for multiple sclerosis as an infection. *Clinical Microbiology Reviews*. https://doi.org/10.1128/CMR.6.4.382

Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., & Špiclin, Ž. (2018). A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. *Neuroinformatics*, *16*(1), 51–63. https://doi.org/10.1007/s12021-017-9348-7

Lo, C.-P., Kao, H.-W., Chen, S.-Y., Hsueh, C.-J., Lin, W.-C., Hsu, W.-L., … Liu, G.-C. (2009). Prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis according to baseline MRI findings: comparison of revised McDonald criteria and Swanton modified criteria. *Journal of Neurology, Neurosurgery & Psychiatry*, *80*(10), 1107–1109. https://doi.org/10.1136/jnnp.2008.169045

Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, *21*(4), 163–169. https://doi.org/10.1145/37402.37422

Maggi, P., Absinta, M., Sati, P., Perrotta, G., Massacesi, L., Dachy, B., … Théaudin, M. (2020). The “central vein sign” in patients with diagnostic “red flags” for multiple sclerosis: A prospective multicenter 3T study. *Multiple Sclerosis Journal*, *26*(4), 421–432. https://doi.org/10.1177/1352458519876031

Masetic, Z., & Subasi, A. (2016). Congestive heart failure detection using random forest classifier. *Computer Methods and Programs in Biomedicine*. https://doi.org/10.1016/j.cmpb.2016.03.020

Mcdonald, W., Compston, a, Edan, G., Goodkin, D., Hp, H., Fd, L., & Hf, M. (2001). Recommended diagnostic criteria for multiple scle ... [Ann Neurol . 2001] - PubMed - ... Recommended diagnostic criteria for multiple sclerosis : guidelines from the International Panel on the diagnosis of multiple sclerosis . Publication Types , MeSH T. *Annals of Neurology*, *59*(April), 11456302. https://doi.org/10.1002/ana.1032

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika 1947 12:2*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mercaldo, N. D., Lau, K. F., & Zhou, X. H. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine*. https://doi.org/10.1002/sim.2677

Meyer-Moock, S., Feng, Y. S., Maeurer, M., Dippel, F. W., & Kohlmann, T. (2014). Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. *BMC Neurology*, *14*(1). https://doi.org/10.1186/1471-2377-14-58

Miller, D., Barkhof, F., Montalban, X., Thompson, A., & Filippi, M. (2005). Review Clinically isolated syndromes suggestive of multiple sclerosis, part I : natural history, pathogenesis, diagnosis , and prognosis. *Lancet Neurology*. https://doi.org/10.1016/S1474-4422(05)70071-5

Miller, D., Chard, D. T., & Ciccarelli, O. (2012). Clinically isolated syndromes. *The Lancet Neurology*, *11*(2), 157–169. https://doi.org/10.1016/S1474-4422(11)70274-5

Mitchell, T. M. (2010). CHAPTER 1 GENERATIVE AND DISCRIMINATIVE CLASSIFIERS : NAIVE BAYES AND LOGISTIC REGRESSION Learning Classifiers based on Bayes Rule. *Machine Learning*, *1*(Pt 1-2), 1–17. https://doi.org/10.1093/bioinformatics/btq112

Moraal, B., Meier, D. S., Poppe, P. A., Geurts, J. J. G., Vrenken, H., Jonker, W. M. A., … Barkhof, F. (2009). Subtraction MR Images in a Multiple Sclerosis Multicenter Clinical Trial Setting. *Radiology*, *250*(2), 506–514. https://doi.org/10.1148/radiol.2501080480

Moraal, B., Pohl, C., Uitdehaag, B. M. J., Polman, C. H., Edan, G., Freedman, M. S., … Barkhof, F. (2009). Magnetic resonance imaging predictors of conversion to multiple

sclerosis in the BENEFIT study. *Archives of Neurology*, *66*(11), 1345–1352. https://doi.org/10.1001/archneurol.2009.243

Moraal, B., Wattjes, M. P., Geurts, J. J. G., Knol, D. L., van Schijndel, R. A., Pouwels, P. J. W., … Barkhof, F. (2010). Improved Detection of Active Multiple Sclerosis Lesions: 3D Subtraction Imaging. *Radiology*, *255*(1), 154–163. https://doi.org/10.1148/radiol.09090814

Mostert, J. P., Koch, M. W., Steen, C., Heersema, D. J., De Groot, J. C., & De Keyser, J. (2010). T2 lesions and rate of progression of disability in multiple sclerosis. *European Journal of Neurology*. https://doi.org/10.1111/j.1468-1331.2010.03093.x

Multiple Sclerosis International Federation. (2014). Atlas of MS 2013: Mapping Multiple Sclerosis Around the World. *Multiple Sclerosis International Federation*. https://doi.org/10.1093/brain/awm236

Newton, B. D., Wright, K., Winkler, M. D., Bovis, F., Takahashi, M., Dimitrov, I. E., … Okuda, D. T. (2017). Three-Dimensional Shape and Surface Features Distinguish Multiple Sclerosis Lesions from Nonspecific White Matter Disease. *Journal of Neuroimaging*, *27*(6), 613–619. https://doi.org/10.1111/jon.12449

Polman, C., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., … Wolinsky, J. S. (2011). Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals of Neurology*, *69*(2), 292–302. https://doi.org/10.1002/ana.22366

Rovira, Á., Wattjes, M. P., Tintoré, M., Tur, C., Yousry, T. A., Sormani, M. P., … Montalban, X. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis - Clinical implementation in the diagnostic process. *Nature Reviews Neurology*, *11*(8), 471–482. https://doi.org/10.1038/nrneurol.2015.106

Ruet, A., Arrambide, G., Brochet, B., Auger, C., Simon, E., Rovira, À., … Tintoré, M. (2014). Early predictors of multiple sclerosis after a typical clinically isolated syndrome. *Multiple Sclerosis Journal*, *20*(13), 1721–1726. https://doi.org/10.1177/

Schmidt, M. A., Linker, R. A., Lang, S., Lücking, H., Engelhorn, T., Kloska, S., … Dankerl, P. (2018). FLAIRfusion Processing with Contrast Inversion: Improving Detection and Reading Time of New Cerebral MS Lesions. *Clinical Neuroradiology*, *28*(3), 367–376. https://doi.org/10.1007/s00062-017-0567-y

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., … Mühlau, M. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage*, *59*(4), 3774–3783. https://doi.org/10.1016/j.neuroimage.2011.11.032

Sinnecker, T., Clarke, M. A., Meier, D., Enzinger, C., Calabrese, M., De Stefano, N., … Wuerfel, J. (2019). Evaluation of the Central Vein Sign as a Diagnostic Imaging Biomarker in Multiple Sclerosis. *JAMA Neurology*. https://doi.org/10.1001/jamaneurol.2019.2478

Stewart, W. A., Hall, L. D., Berry, K., & Paty, D. W. (1984). CORRELATION BETWEEN NMR SCAN AND BRAIN SLICE DATA IN MULTIPLE SCLEROSIS. *The Lancet*. https://doi.org/10.1016/S0140-6736(84)90584-1

Swanton, J. K., Rovira, A., Tintore, M., Altmann, D. R., Barkhof, F., Filippi, M., … Miller, D. H. (2007). MRI criteria for multiple sclerosis in patients presenting with clinically isolated syndromes: a multicentre retrospective study. *Lancet Neurology*. https://doi.org/10.1016/S1474-4422(07)70176-X

Sweeney, E. M., Shinohara, R. T., Shea, C. D., Reich, D. S., & Crainiceanu, C. M. (2013). Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *American Journal of Neuroradiology*, *34*(1), 68–73. https://doi.org/10.3174/ajnr.A3172

Tallantyre, E. C., Brookes, M. J., Dixon, J. E., Morgan, P. S., Evangelou, N., & Morris, P. G. (2008). Demonstrating the perivascular distribution of ms lesions in vivo with 7-tesla MRI. *Neurology*, *70*(22), 2076–2078. https://doi.org/10.1212/01.wnl.0000313377.49555.2e

Tan, I. L., Van Schijndel, R. A., Fazekas, F., Filippi, M., Freitag, P., Miller, D. H., … Barkhof, F. (2002). Image registration and subtraction to detect active T2 lesions in MS: An interobserver study. *Journal of Neurology*, *249*(6), 767–773. https://doi.org/10.1007/s00415-002-0712-6

Tan, I. L., van Schijndel, R. A., Pouwels, P. J. W., Adèr, H. J., & Barkhof, F. (2002). Serial Isotropic Three-Dimensional Fast FLAIR Imaging: Using Image Registration and Subtraction to Reveal Active Multiple Sclerosis Lesions. *American Journal of Roentgenology*, *179*(3), 777–782. https://doi.org/10.2214/ajr.179.3.1790777

Tan, I. L., Van Schijndel, R. A., Pouwels, P. J. W., Van Walderveen, M. A. A., Reichenbach, J. R., Manoliu, R. A., & Barkhof, F. (2000). MR venography of multiple sclerosis. *American Journal of Neuroradiology*, *21*(6), 1039–1042. https://doi.org/10871010

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., … Cohen, J. A. (2017). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, *4422*(17). https://doi.org/10.1016/S1474-4422(17)30470-2

Thompson, A. J., Baranzini, S. E., Geurts, J., Hemmer, B., & Ciccarelli, O. (2018). Seminar Multiple sclerosis. *Www.Thelancet.Com*. https://doi.org/10.1016/S0140-6736(08)61620-7

Tintoré, M., Rovira, A., Río, J., Nos, C., Grivé, E., Sastre-Garriga, J., … Montalban, X. (2003). New diagnostic criteria for multiple sclerosis: Application in first demyelinating episode. *Neurology*. https://doi.org/10.1212/WNL.60.1.27

Tintoré, M., Rovira, A., Río, J., Nos, C., Grivé, E., Téllez, N., … Montalban, X. (2006). Baseline MRI predicts future attacks and disability in clinically isolated syndromes.

Neurology, *67*(6), 968–972. https://doi.org/10.1212/01.wnl.0000237354.10144.ec

Tintoré, Mar, Rovira, A., Río, J., Tur, C., Pelayo, R., Nos, C., … Montalban, X. (2008). Do oligoclonal bands add information to MRI in first attacks of multiple sclerosis? *Neurology*, *70*(13 PART 2), 1079–1083. https://doi.org/10.1212/01.wnl.0000280576.73609.c6

Tossberg, J. T., Crooke, P. S., Henderson, M. a, Sriram, S., Mrelashvili, D., Vosslamber, S., … Aune, T. M. (2013). Using biomarkers to predict progression from clinically isolated syndrome to multiple sclerosis. *Journal of Clinical Bioinformatics*, *3*(1), 18. https://doi.org/10.1186/2043-9113-3-18

Viglietta, V., Baecher-Allan, C., Weiner, H. L., & Hafler, D. A. (2004). Loss of Functional Suppression by CD4 $^{+}$ CD25 $^{+}$ Regulatory T Cells in Patients with Multiple Sclerosis. *The Journal of Experimental Medicine*. https://doi.org/10.1084/jem.20031579

Vural, G., Keklikoğlu, H. D., Temel, Ş., Deniz, O., & Ercan, K. (2013). Comparison of double inversion recovery and conventional magnetic resonance brain imaging in patients with multiple sclerosis and relations with disease disability. *Neuroradiology Journal*, *26*(2), 133–142. https://doi.org/10.1177/197140091302600201

Wadell, H. (1935). Volume, Shape, and Roundness of Quartz Particles. *The Journal of Geology*, *43*(3), 250–280. https://doi.org/10.1086/624298

Wattjes, M. P., Lutterbey, G. G., Gieseke, J., Träber, F., Klotz, L., Schmidt, S., & Schild, H. H. (2007). Double inversion recovery brain imaging at 3T: Diagnostic value in the detection of multiple sclerosis lesions. *American Journal of Neuroradiology*.

Wottschel, V., Alexander, D. C., Kwok, P. P., Chard, D. T., Stromillo, M. L., De Stefano, N., … Ciccarelli, O. (2015). Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clinical*, *7*, 281–287. https://doi.org/10.1016/j.nicl.2014.11.021

Zhang, H., Alberts, E., Pongratz, V., Mühlau, M., Zimmer, C., Wiestler, B., & Eichinger, P. (2018). Predicting conversion from clinically isolated syndrome to multiple sclerosis–An imaging-based machine learning approach. *NeuroImage: Clinical*. https://doi.org/10.1016/j.nicl.2018.11.003

Zhang, W. Y., & Hou, Y. L. (2013). Prognostic value of magnetic resonance imaging in patients with clinically isolated syndrome conversion to multiple sclerosis: A meta-analysis. *Neurology India*. https://doi.org/10.4103/0028-3886.115058

Zijdenbos, A. P., Forghani, R., & Evans, A. C. (2002). Automatic “pipeline” analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2002.806283

Zivadinov, R., Havrdová, E., Bergsland, N., Tyblova, M., Hagemeier, J., Seidl, Z., … Horáková, D. (2013). Thalamic atrophy is associated with development of clinically

definite multiple sclerosis. *Radiology*, *268*(3), 831–841. https://doi.org/10.1148/radiol.13122424

6 Appendix

6.1 Supplementary materials

6.1.1 MRI parameters

The imaging parameters were as follows:

3D DIR: Acquired voxel size, 1.2 × 1.2 × 1.3 mm_3; acquisition matrix, 208 × 208; field of view, 250; repetition time (TR) 5500 ms; echo time (TE) 328 ms; inversion time (TI) 2550 ms; TSE factor 173; number of slices 300; acquisition time 6 min; plane, sagittal.

3D FLAIR: Acquired voxel size, 1.03 x 1.03 x 1.5 mm^3; acquisition matrix, 224 x 154; field of view, 230; TR, 10000 ms; TE, 140 ms; TSE factor, 20; number of slices, 96; acquisition time, 5 min; plane, axial.

3D T1: Acquired voxel size, 1 x 1 x 1 mm^3; acquisition matrix, 240 x 240; field of view, 240; TR, 9 ms; TE, 4 ms; number of slices, 170; acquisition time, 6 min; plane, sagittal.

3D T2-TSE: Acquired voxel size, 1.03 × 1.03 × 1.5 mm_3; acquisition matrix 224 × 162; field of view 230; TR 4000–6000 ms (variable); TE 35 ms; TSE factor 7; number of slices 96; acquisition time 5 min; plane, axial.

6.1.2 Associated data

Exact numbers of the boxplots that are shown in Figure 6 on page 21:

	Standard	DIR map	FLAIR map
Mean [min]	9.6	3.1	3.1
Median [min]	8	2	2
25%-interquartile [min]	5	1	2
75%-interquartile [min]	12	3	3
Minimum [min]	0	1	1
Maximum [min]	42	22	25

Numbers to the boxplots in Figure 14 on page 39:

	Mean	Median	25%-IQ	75%-IQ	Mini-mum	Maxi-mum
Mean volume - CIS [mm^3]	70.99	49.55	33.90	84.85	22.00	341.00
Mean volume–MS [mm^3]	135.64	103.50	61.58	171.75	22,70	671.00
Minimum SVR–CIS [mm^{-1}]	1.2162	1.1150	0.9952	1.3750	0.6180	2.0300
Minimum SVR–MS [mm^{-1}]	0.5474	0.5530	0.4990	1.1175	0.2520	1.7500
Minimum sphericity–CIS	0.6292	0.9490	0.5590	0.6845	0.4240	0.8260
Minimum sphericity–MS	0.5474	0.5530	0.4990	0.6252	0.2520	0.6960

6.1.3 MATLAB script for calculating subtraction maps

This script calculates the subtraction maps for serial MR images in NIFTI file format without further post-processing steps. Line 14-20 show an automatically generated code by SPM 12. The follow-up image is named “post.nii” and the baseline image is named “pre.nii”.

```
12  %SPM-generated code: (used to reslice and coregister)
13
14  nrun = 1; % enter the number of runs here
15  jobfile = {'batch_reg_job.m'}; %working directory must be inserted manually
16  jobs = repmat(jobfile, 1, nrun);
17  inputs = cell(0, nrun);
18
19  spm('defaults', 'FMRI');
20  spm_jobman('run', jobs, inputs{:});
21  %SPM-generated code ends here, generates 'rpre.nii' which is 'pre.nii' registered on 'post.nii'
22
23  % Import post.nii and rpre.nii
24
25  rprevol = spm_vol('rpre.nii');
26  rpostvol = spm_vol('post.nii');
27  pre=spm_read_vols(rprevol);
28  post=spm_read_vols(rpostvol);
29
30  %calculate subtraction
31  sub  =  post-pre;
32
33  %write subtraction into file
34  subvol  = rpostvol;
35  subvol.fname = 'sub.nii';
36  spm_write_vol(subvol,sub);
```

6.1.4 Python script for the prediction project

This python script calculates the parameters of interest in the second project and also the descriptive statistics of these parameters:

```
from os import listdir,linesep
import nibabel as nib
import numpy as np
from skimage.measure import(mesh_surface_area,marching_cubes_lewiner)
from skimage.morphology import label
from scipy.stats import(skew,kurtosis)

path_to_images = "c:/users/bene/desktop/MS_CIS_Cohort/" #Should end with a "/"
all_files = listdir(path_to_images)
num_of_samples = int(len(all_files) / 3)

"""
Features:
---------
SampleID
Läsionszahl
Volumen (total, min, max, mean, sd)
Surface (min, max, mean, sd)
Sphericity (min, max, mean, sd)
svr (min, max, mean, sd)
Intensity_kurtosis_t1(total, min, max, mean, sd)
Intensity_skewness_t1(total, min, max, mean, sd)
Intensity_kurtosis_flair(total, min, max, mean, sd)
Intensity_skewness_flair(total, min, max, mean, sd)
"""

result = np.zeros(((num_of_samples),39))

for ii in (np.arange(num_of_samples)):
    sample = int(all_files[ii][7:13])
    result[(ii),0] = sample

    mask_file = nib.load(path_to_images + all_files[ii])
    mask = mask_file.get_data()

    flair_file = nib.load(path_to_images + all_files[ii+num_of_samples])
    flair = flair_file.get_data()

    t1_file = nib.load(path_to_images + all_files[ii+(num_of_samples*2)])
    t1 = t1_file.get_data()

    num_of_lesions = np.unique(label(mask)).max()
    mask_labels = label(mask)

    result[(ii),1] = num_of_lesions

    volumen = []
    surface = []
    sphericity = []
    svr = []
    t1_total = []
    t1_kurtosis = []
    t1_skewness = []
    flair_total = []
    flair_kurtosis = []
    flair_skewness = []

    for jj in (np.arange(1,num_of_lesions+1)):
        temp_mask = np.zeros(mask_labels.shape)
        temp_mask[mask_labels == jj] = 1

        cur_volume = len(temp_mask[temp_mask==1]) * 1 * 1 * 1
        volumen.append(cur_volume)

        verts, faces, normals, values = marching_cubes_lewiner(temp_mask,level=0.1,spacing=(1,1,1))
        cur_surface = mesh_surface_area(verts,faces)
        surface.append(cur_surface)

        cur_sphericity = (36 * np.pi * cur_volume ** 2) ** (1.0 / 3.0) / cur_surface
        sphericity.append(cur_sphericity)

        svr.append(cur_surface / cur_volume)

        t1_array = np.array(t1[temp_mask==1])
        t1_total.append(t1_array)
        t1_kurtosis.append(kurtosis(t1_array))
        t1_skewness.append(skew(t1_array))

        flair_array = np.array(flair[temp_mask==1])
        flair_total.append(flair_array)
        flair_kurtosis.append(kurtosis(flair_array))
        flair_skewness.append(skew(flair_array))
```

```
result[(ii),2] = np.sum(volumen)
result[(ii),3] = np.min(volumen)
result[(ii),4] = np.max(volumen)
result[(ii),5] = np.mean(volumen)
result[(ii),6] = np.std(volumen)

result[(ii),7] = np.min(surface)
result[(ii),8] = np.max(surface)
result[(ii),9] = np.mean(surface)
result[(ii),10] = np.std(surface)

result[(ii),11] = np.min(sphericity)
result[(ii),12] = np.max(sphericity)
result[(ii),13] = np.mean(sphericity)
result[(ii),14] = np.std(sphericity)

result[(ii),15] = np.min(svr)
result[(ii),16] = np.max(svr)
result[(ii),17] = np.mean(svr)
result[(ii),18] = np.std(svr)

t1_total = np.concatenate(t1_total)
result[(ii),19] = kurtosis(t1_total)
result[(ii),20] = np.min(t1_kurtosis)
result[(ii),21] = np.max(t1_kurtosis)
result[(ii),22] = np.mean(t1_kurtosis)
result[(ii),23] = np.std(t1_kurtosis)

result[(ii),24] = skew(t1_total)
result[(ii),25] = np.min(t1_skewness)
result[(ii),26] = np.max(t1_skewness)
result[(ii),27] = np.mean(t1_skewness)
result[(ii),28] = np.std(t1_skewness)

flair_total = np.concatenate(flair_total)
result[(ii),29] = kurtosis(flair_total)
result[(ii),30] = np.min(flair_kurtosis)
result[(ii),31] = np.max(flair_kurtosis)
result[(ii),32] = np.mean(flair_kurtosis)
result[(ii),33] = np.std(flair_kurtosis)

result[(ii),34] = skew(flair_total)
result[(ii),35] = np.min(flair_skewness)
result[(ii),36] = np.max(flair_skewness)
result[(ii),37] = np.mean(flair_skewness)
result[(ii),38] = np.std(flair_skewness)
```

6.1.5 R Script for the prediction project

This script performs the oblique random forest classification and the three-fold cross-validation (line 1, line 6-25); and performs a bootstrapping approach to calculate the most important shape features (line 51-72).

```
1  library(obliqueRF)
2  library(pROC)
3  library(ggplot2)
4  library(caret)
5  #cis_lesion_features.txt
6  dat <- read.table(file.choose(),header=T,sep="\t",stringsAsFactors = F)
7
8  dat.calc <- dat[,2:ncol(dat)]
9
10 #k folds
11 set.seed(123456789)
12 k.folds <- createFolds(as.factor(dat.calc$ConvMS),k=3)
13
14 prob <- matrix(nrow=nrow(dat.calc),ncol=1)
15 img.matrix <- as.matrix(dat.calc[,2:19])
16 #img.matrix <- as.matrix(dat.calc[,c(2,20:39)]) #For intensity
17 conv.vector <- dat.calc$ConvMS
18 for (i in 1:length(k.folds)){
19   pred.matrix <- img.matrix[-unlist(k.folds[i]),]
20   cv.matrix <- img.matrix[unlist(k.folds[i]),]
21   conv.vector.cv <- conv.vector[-unlist(k.folds[i])]
22   orf.cv <- obliqueRF(pred.matrix,conv.vector.cv,ntree=300,mtry=3,training_method = "log")
23   prob.tmp <- predict(orf.cv,type=c("prob"),newdata=cv.matrix)
24   prob[unlist(k.folds[i]),] <- prob.tmp[,1]
25 }
```

```
51 #Bootstrap ORF importance
52 #Custom-built to avoid errors due to under-sampling one group (at least 3 samples per group)
53 #
54 img.matrix <- as.matrix(dat.calc[,2:19])
55 conv.vector <- dat.calc$ConvMS
56 set.seed(123456789)
57 for (bsn in 1:100){
58   bs <- sample(84,84,replace=T)
59   while (length(subset(conv.vector[bs],conv.vector[bs]==1)) *
60          length(subset(conv.vector[bs],conv.vector[bs]==0)) < 243) bs <- sample(84,84,replace=T)
61   imp.orf <- obliqueRF(img.matrix[bs,],conv.vector[bs],ntree=300,mtry=3,bImportance=T,training_method="log")$importance
62   if (bsn == 1) {res <- imp.orf}
63   else {res <- rbind(res,imp.orf)}
64   print(paste(bsn," of 100 iterations",sep=""));flush.console()
65 }
66 imp.mean <- apply(res,2,mean)
67 imp.mean <- scale(imp.mean)
68
69 vars <- as.factor(names(dat.calc)[2:19])
70 middle <- c(1,4,8.5,12.5,16.5)
71 end <- c(1.5,6.5,10.5,14.5,18.5)
72 x.labels <- c("LesionCount","Volume","Surface Area","Sphericity","SVR")
```

6.1.6 Software Reference and URLs

BrainSeg3D	Laboratory of Imaging Technologies, Faculty of Electrical Engineering, University of Ljubljana, Slovenia URL: https://www.quantim.eu/knowledge-base
ITK-Snap	Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* 2006 Jul 1;31(3):1116-28. URL: http://www.itksnap.org/
LST	Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of FLAIR-hyper-intense white-matter lesions in multiple sclerosis. Neuroimage, 59(4):3774–3783, 2012. URL: www.statisticalmodelling.de/lst.html
MATLAB	MATLAB and Statistics Toolbox, Natick, Massachusetts: The Math-Works Inc. URL: https://de.mathworks.com/
Python	Python Software Foundation. Python Language Reference. URL: http://www.python.org
R	R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/
Seg3D	Center for Integrative and Biomedical Computing (2016). Seg3D: Volumetric Image Segmentation and Visualization. Scientific Computing and Imaging Institute (SCI) URL: https://www.sci.utah.edu/cibc-software/seg3d.html
SPM12	The Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London, UK. URL: https://www.fil.ion.ucl.ac.uk/spm/software/download/
SPSS	IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp. URL: https://www.ibm.com/products/spss-statistics

6.1.7 MATLAB algorithms

ROBEX — Iglesias, J. E., Liu, C. Y., Thompson, P. M., & Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2011.2138152

N4 bias correction algorithm — Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2010.2046908

6.1.8 Python libraries

nibabel — Brett, M., Markiewicz, C. J., Hanke, M., Côté, M.-A., Cipollini, B., McCarthy, P., … freec84. (2020). nipy/nibabel: 3.2.1. https://doi.org/10.5281/ZENODO.4295521

NumPy — Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature 2020 585:7825*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Scikit-learn — Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

SciPy — Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Scikit-image — Van Der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., … Yu, T. (2014). Scikit-image: Image processing in python. *PeerJ*, *2014*(1). https://doi.org/10.7717/PEERJ.453

6.1.9 R packages

obliqueRF Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., & Hamprecht, F. A. (2011). On oblique random forests. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6912 LNAI, pp. 453–469). https://doi.org/10.1007/978-3-642-23783-6_29

ggplot Wickham, H. (2009). ggplot2. *Ggplot2*. *Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org. https://doi.org/10.1007/978-0-387-98141-3

6.2 List of publications

Zhang, H., Alberts, E., Pongratz, V., Mühlau, M., Zimmer, C., Wiestler, B., Eichinger, P. (2019). Predicting conversion from clinically isolated syndrome to multiple sclerosis–An imaging-based machine learning approach. *NeuroImage Clin.* https://doi.org/10.1016/j.nicl.2018.11.003

Eichinger, P., Schön, S., Pongratz, V., Wiestler, H., Zhang, H., Bussas, M., Hoshi, M.M., Kirschke, J., Berthele, A., Zimmer, C., Hemmer, B., Mühlau, M., Wiestler, B. (2019). Accuracy of unenhanced MRI in the detection of new brain lesions in multiple sclerosis. *Radiology 291*, 429–435. https://doi.org/10.1148/radiol.2019181568

Eichinger, P., Wiestler, H., Zhang, H., Biberacher, V., Kirschke, J.S., Zimmer, C., Mühlau, M., Wiestler, B. (2017). A novel imaging technique for better detecting new lesions in multiple sclerosis. *J. Neurol*. 264, 1909–1918. https://doi.org/10.1007/s00415-017-8576-y

6.3 Acknowledgments

I would like to express my deepest gratitude to my doctorate supervisor, PD Dr. med. Benedikt Wiestler of the department diagnostic and interventional neuroradiology at Klinikum rechts der Isar at Technical University Munich, who has guided and encouraged me through this project. Thank you for your kind welcome and your permanent support since our first encounter in May 2016. Thank you for your motivation to grow beyond my limits and for including me in your excellent research team. I greatly benefitted from your expertise in imaging and machine learning. It is an honor to be your first doctorate candidate and to research in such an exciting and highly promising topic of MS and AI.

I would also like to acknowledge Prof. Mark Mühlau of the neurology department at the Klinikum rechts der Isar as my doctorate mentor. Thank you for your continuous guidance and your valuable input throughout the years.

I would like to thank Dr. med. Paul Eichinger who introduced me to neuroradiology and the opportunity to do research in this field. He taught me a lot of the knowledge and introduced me to the computer skills I needed to complete this work. He showed me how to use MATLAB and SPM. His door was always open for my questions und trouble spots about my research and writing. I am grateful for his valuable comments as the second reader of this thesis and his general expertise in the field of neuroradiology.

I wish to thank my former doctorate supervisor, Prof. Dr. med. Claus Zimmer, Chief of the department diagnostic and interventional neuroradiology at Klinikum rechts der Isar atTechnical University Munich. I want to thank him for his regular evaluations of the process of this dissertation.

I would also like to thank Esther Alberts, a fellow doctoral candidate and expert who were involved in the coding of MATLAB and PYTHON models.

I like to pronounce my sincere gratitude to the Department of Neurology at the Klinikum Rechts der Isar. I wish to thank Prof. Dr. med. Bernhard Hemmer and Dr. med. Viola Pongratz, who kindly provided us with the data on MS patients utilized in this thesis. I also want to thank Dr. med. Hanni Wiestler, who was part of the subtraction map project and who supported me at that time with any question I had. I am very grateful for their supportive cooperation at all times.

Großen Dank an meine Familie, die mich den ganzen Weg von Beginn meines Medizinstudiums bis zur Abgabe dieser Dissertation immer unterstützt hat und diesen Weg überhaupt erst ermöglicht hat. In schwierigen Zeiten konnte ich immer auf aufmunternde Worte zählen und ich wurde darin bestärkt, dass kein Weg zu schwer ist und alles möglich ist.

Vielen lieben Dank an Sebastian für die unendliche emotionale und fachliche Unterstützung über die Jahre meines Studiums, während des Praktischen Jahrs und während der Doktorarbeit. Unabhängig von der physischen Entfernung oder Tageszeit konnte ich mich auf seine bestmögliche Unterstützung als Datenexperte und Ranger verlassen.

Ebenso möchte ich meinen lieben Freundinnen Stefanie und Angelika danken, mit denen ich über die Jahre viel und konstruktiven Austausch über Fallstricke und Probleme einer Promotion hatte. Ich durfte von ihren Erfahrungswerten und Ratschlägen sehr profitieren und freue mich über den zukünftigen intensiven Austausch in unserem weiteren Werdegang.