A Machine Learning Approach to Distinguishing between Multiple Sclerosis and Cerebral Small Vessel Disease

by

Daniel Eftekhar

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science, Biomedical Engineering

Institute of Biomaterials & Biomedical Engineering
University of Toronto

© Copyright by Daniel Eftekhar 2018
A Machine Learning Approach to Distinguishing between Multiple Sclerosis and Cerebral Small Vessel Disease

Daniel Eftekhari
Master of Applied Science, Biomedical Engineering
Institute of Biomaterials & Biomedical Engineering
University of Toronto
2018

Abstract

We sought to develop an automatic diagnostic algorithm to distinguish between relapsing-remitting multiple sclerosis (RRMS) and cerebral small vessel disease (SVD) using neuroimaging and clinical features.

A mixture of t-distributions and a spatial heuristic algorithm were developed for the automatic segmentation of lesions on two MRI sequences. Combined lesion probability maps of RRMS and SVD were subsequently developed using a derivation set of patients. A novel cross entropy image distance metric was used to quantify the similarity of new cases to each disease. Bayesian learning algorithms using non-informative priors were trained using neuroimaging features and clinical features. Model hyperparameters were tuned using Monte Carlo cross validation.

The model consisting of both neuroimaging and clinical features misclassified one case out of 21 on the test set when distinguishing between RRMS and SVD.
As the societal and monetary cost of both diseases is high, this work has real potential for clinical impact.
Acknowledgements

I would like to thank my supervisors, Dr. Richard Aviv and Dr. Pascal Tyrrell, for their continued support and guidance throughout the course of this project. I would also like to thank Dr. Anne Martel and Dr. Alan Moody for their insightful input during the project. Most of all, I’d like to thank my mother for giving me every opportunity to succeed in life, and for fostering in me a love for learning and discovery.
# Table of Contents

Acknowledgements ........................................................................................................ iv
List of Tables ...................................................................................................................... vii
List of Figures ..................................................................................................................... ix
List of Abbreviations ......................................................................................................... xiv

1. Introduction ................................................................................................................... 1

2. Literature Review ......................................................................................................... 4
   2.1 Multiple Sclerosis ..................................................................................................... 4
   2.1.1 Multiple Sclerosis – Neuroimaging ...................................................................... 5
   2.2 Cerebral Small Vessel Disease ............................................................................... 7
       2.2.1 Cerebral Small Vessel Disease – Neuroimaging .............................................. 7
   2.3 Differences between Multiple Sclerosis and Cerebral Small Vessel Disease .......... 8
   2.4 Automatic Brain Tissue and Lesion Segmentation ................................................. 10
       2.4.1 Mixture Model & Expectation Maximization Algorithms .......................... 10
       2.4.2 Markov and Conditional Random Fields ..................................................... 12
       2.4.3 Deep Learning .............................................................................................. 13
       2.4.4 Other Approaches ......................................................................................... 14
   2.5 Co-registration ....................................................................................................... 14
   2.6 Automatic Algorithms for Diagnosis .................................................................... 15
       2.6.1 Logistic Regression ....................................................................................... 15
       2.6.2 Naïve Bayes .................................................................................................. 18
       2.6.3 Artificial Neural Networks ........................................................................... 21
   2.7 Related Work .......................................................................................................... 21
       2.7.1 Feature Extraction Approaches .................................................................... 22
       2.7.2 Image Retrieval & Similarity Approaches ................................................... 22
       2.7.3 Deep Learning Approaches ......................................................................... 22

3. Methods ......................................................................................................................... 23
   3.1 Neuroimaging ......................................................................................................... 24
       3.1.1 Patient Cohort ............................................................................................... 25
       3.1.2 Image Acquisition and Ground Truth Tracings .......................................... 25
       3.1.3 Brain Extraction & Bias Field Correction .................................................... 26
       3.1.4 Automatic Lesion Segmentation .................................................................. 27
           3.1.4.1 White Matter Lesion Segmentation ..................................................... 28
           3.1.4.2 T1 Black Hole & Lacune Segmentation ............................................. 31
           3.1.4.3 Minimizing False Positives .................................................................. 39
       3.1.5 Co-registration ............................................................................................... 41
           3.1.5.1 Lesion Filling ....................................................................................... 41
           3.1.5.2 Applying the Co-registration ............................................................... 42
       3.1.6 Segmentation Performance Metrics ................................................................ 43
       3.1.7 Combined Lesion Probability Maps .............................................................. 43
           3.1.7.1 Partitioning the Data and Cross Validation ........................................ 45
           3.1.7.2 Monte Carlo Cross Validation ......................................................... 46
           3.1.7.3 Neuroimaging Similarity Score ......................................................... 46
   3.2 Disease Classification .............................................................................................. 49
List of Tables

Table 1 Clinical features known to distinguish multiple sclerosis and cerebral small vessel disease. ................................................................................................................................. 8

Table 2 Neuroimaging features known to distinguish multiple sclerosis and cerebral small vessel disease. ................................................................................................................................. 9

Table 3 Intraclass correlation coefficient for total lesion volume between our segmentation algorithms and the ground truth tracings for two different datasets. Results are presented using two-way random effects with absolute agreement and single measures. ........................................... 59

Table 4 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes. Mean performance of the cross entropy image similarity metric for the best performing hyperparameters is provided. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. .................................................................................. 64

Table 5 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes. First and second rows: restricting $\epsilon$ to be the same for both diseases. Third row: Allowing $\epsilon$ to vary between the two diseases. Mean performance of the image similarity metrics for the optimal hyperparameter settings using the Gaussian Naïve Bayes classifier, without and with the non-informative prior. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. Using the non-informative prior decreases derivation set performance slightly, but importantly results in greater performance on the validation set (p < 0.05). ........................................................................................................... 68

Table 6 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. The average class means and standard deviations by lesion type for the Gaussian Naïve Bayes classifier are provided. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. The significant difference in class means for both WMLs and BHs suggests that discriminating between lesion subtypes is important in distinguishing between the two diseases. ............... 69
Table 7 Mean performance for distinguishing between RRMS vs. SVD and RRMS & SPMS vs. SVD using clinical features alone. The model used is the Naïve Bayes classifier with a non-informative prior. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. Higher performance is achieved when distinguishing between RRMS vs. SVD compared to RRMS & SPMS vs. SVD (p < 0.05), due to the smaller age discrepancy and smaller sex ratio difference between SPMS and SVD patients, compared to RRMS and SVD patients.

Table 8 RRMS vs. SVD. Mean performance using combined neuroimaging and clinical features, for the best hyperparameter settings. The model used is the Naïve Bayes classifier with a non-informative prior, and allows $\epsilon$ to vary between the two classes. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors.
List of Figures

Figure 1 Methodology. The left side represents the image processing & analysis component. The right side represents the aggregation of neuroimaging and clinical features, as well as the statistical analysis. 24

Figure 2 a) FLAIR image of an RRMS impaired patient. Arrows point to white matter lesions. b) T1 image of the same patient. Arrow points to a T1 black hole. c) Ground truth segmentation (green = grey matter, blue = white matter, purple = basal ganglia, orange = thalamus, beige = white matter lesion, white = T1 black hole). 25

Figure 3 a) T1 of an RRMS impaired patient b) Result of applying brain extraction. 27

Figure 4 (Modified color scheme) a) T1 of an RRMS impaired patient b) Result of applying bias field correction. 27

Figure 5 a) T1 of an RRMS impaired patient. Black holes (blue arrow) often appear adjacent to the ventricles (orange arrow) and other CSF (green arrow). b) SPM’s segmentation probability map for CSF. The associated black holes from a) have been assigned to CSF. 31

Figure 6 SPM’s white matter prior probability map. Prior probability of white matter near the ventricles (blue arrow) is not high enough to counteract the hypointensity of black holes on T1. Therefore, SPM assigns hypointense voxels in this area to CSF. 32

Figure 7 a) Original white matter prior probability map. b) After the transformation. 34

Figure 8 a) Transformed white matter prior probability map with no smoothing b) Gaussian smoothing with $\sigma = 1$ c) Gaussian smoothing with $\sigma = 2$ d) Gaussian smoothing with $\sigma = 3$. 34

Figure 9 a) T1 image of an RRMS impaired patient. b) Initial T1 black hole segmentation – clearly it is an over-segmentation, since it includes much of the CSF in its segmentation. 35

Figure 10 a) T1 image of an RRMS impaired patient. b) T1 black hole segmentation after the erosions and dilations – we are now left with an under-segmentation. 36
Figure 11 Region growing algorithm. a) 1 denotes a lesional voxel in the over-segmentation alone, and 2 denotes a lesional voxel in both the under and over-segmentations. b) Voxels labeled 3 are the result of the region growing algorithm.............................................................. 37

Figure 12 a) T1 image of an RRMS impaired patient. b) over-segmentation c) under-segmentation d) result of region growing algorithm from under-segmentation to over-segmentation .......................................................................................................................... 39

Figure 13 a) FLAIR of an RRMS impaired patient. The blue arrow points at the corpus callosum. b) Erroneous (blue arrow) white matter lesion detected. ....................................................... 39

Figure 14 a) The original CSF segmentation produced by SPM b) Result of thresholding, applying the Gaussian filter, and the Prewitt filter c) Result of applying the convex hull to the image in b). ...................................................................................................................................... 40

Figure 15 Left: FLAIR of an RRMS impaired patient. Center: corresponding white matter lesion segmentation. Right: filled FLAIR image............................................................. 41

Figure 16 Left: T1 of an RRMS impaired patient. Center: corresponding T1 black hole lesion segmentation. Right: filled T1 image.......................................................................................... 42

Figure 17 a) Reference image: MNI-152 T1. b) RRMS patient’s T1 scan in MNI space c) FLAIR in MNI space after being co-registered to the filled T1 scan. d) Lesion map in MNI space...................................................................................................................................... 42

Figure 18 (Left) T1 MNI template space image. Combined probability map from one Monte Carlo simulation for RRMS (center) and SVD (right)......................................................... 45

Figure 19 A schematic of the four approaches used for classification. 1) Using neuroimaging alone, but without distinguishing between white matter lesions and black holes. 2) Using neuroimaging alone, and distinguishing white matter lesions and black holes. 3) Using clinical features alone. 4) Using a combination of neuroimaging and clinical features........................................... 50

Figure 20 Bland-Altman plot for the total lesion volume between the lesion segmentation algorithm and the ground truth tracings on the RRMS dataset.................................................. 61
Figure 21 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon = 2^2$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. There is good separability between the two classes on the derivation set, but without significant overfitting, as exemplified by the ROC and PRC, which leads to good performance on the validation set. ................................................................. 64

Figure 22 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon = 2^8$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. Although the separability between classes is very good on the derivation set, the model is overfitting, as exemplified by the ROC and PRC, and thus the classifier performs poorly on the validation set. ........................................................................................................ 65

Figure 23 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon = 2^1$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. Observing the ROC and PRC, the model cannot distinguish between the two classes very well on the derivation set, and is therefore underfitting. This leads to poor performance in distinguishing the classes on the validation set. ........................................................................... 66

Figure 24 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and allowing for different values of $\epsilon$ for each disease. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon_{MS} = 2^4$, $\epsilon_{SVD} = 2^5$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. There is very good separability between the two classes on the derivation set, but without significant overfitting, as exemplified by the ROC and PRC, which leads to good performance on the validation set. ...................................................................................... 67
Figure 25 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. A Gaussian Naïve Bayes model constructed from a single Monte Carlo simulation using the optimal hyperparameter settings. Left: without the non-informative prior, right: with the non-informative prior. Blue corresponds to RRMS cases, and red to SVD cases. Circles represent derivation set cases used to construct the model, and crosses represents validation set cases. For both models, there is good separation between the two classes on the derivation set, which leads to good performance on the validation set. However, the model with the non-informative prior performs slightly better, in part because the non-informative prior causes the model to make less extreme predictions near the decision boundary.

Figure 26 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Left column: without the non-informative prior, right column: with the non-informative prior. Varying $\epsilon_{BH}$ from high (top) to low (bottom) leads to smaller separation in BH scores between the two classes. High values of $\epsilon_{BH}$ lead to overfitting to the derivation set, and small values lead to model underfitting. Intermediate values of $\epsilon_{BH}$ (center) lead to optimal performance.

Figure 27 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Left column: without the non-informative prior, right column: with the non-informative prior. Varying $\epsilon_{WML}$ from high (top) to low (bottom) leads to smaller separation in WML scores between the two classes. High values of $\epsilon_{WML}$ lead to overfitting to the derivation set, and small values lead to model underfitting. Intermediate values of $\epsilon_{WML}$ (center) lead to optimal performance.

Figure 28 RRMS vs. SVD. Monte Carlo simulation performances converged after approximately 100 simulations. Top-left: model developed using a single combined probability map for white matter lesions and T1 black holes. Top-right: model developed using separate combined probability maps for white matter lesions and T1 black holes. Bottom-left: model developed using clinical features alone. Bottom-right: model developed using combined neuroimaging and clinical features.

Figure 29 A visualization of the algorithm’s decision-making for an MS case. Positive values contribute to a diagnosis of MS, and negative values to a diagnosis of SVD. Left) white matter
lesions - the algorithm has discovered, from data alone, that periventricular lesions are associated with a diagnosis of MS. Right) black holes – most of the lesions contribute to a diagnosis of MS, except for one region which is the anterior limb of the internal capsule, a region known to affect SVD patients. The algorithm consolidates the findings from both regions, and correctly classifies the patient with MS. ................................................................. 77
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>BH</td>
<td>T1 Black Hole</td>
</tr>
<tr>
<td>BN</td>
<td>Bayesian Network</td>
</tr>
<tr>
<td>CDSS</td>
<td>Clinical Decision Support System</td>
</tr>
<tr>
<td>CIS</td>
<td>Clinically Isolated Syndrome</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>CNS</td>
<td>Central Nervous System</td>
</tr>
<tr>
<td>CPM</td>
<td>Combined Probability Map</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Random Field</td>
</tr>
<tr>
<td>CSF</td>
<td>Cerebrospinal Fluid</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization (Algorithm)</td>
</tr>
<tr>
<td>FLAIR</td>
<td>Fluid-Attenuated Inversion Recovery MRI</td>
</tr>
<tr>
<td>GM</td>
<td>Grey Matter</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass Correlation Coefficient</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo (Simulation)</td>
</tr>
<tr>
<td>MCC</td>
<td>Matthews Correlation Coefficient</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov Random Field</td>
</tr>
<tr>
<td>MNI</td>
<td>Montreal Neurological Institute</td>
</tr>
<tr>
<td>MS</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>NAGM</td>
<td>Normal Appearing Grey Matter</td>
</tr>
<tr>
<td>NAWM</td>
<td>Normal Appearing White Matter</td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>O()</td>
<td>Big O notation</td>
</tr>
<tr>
<td>PRC</td>
<td>Precision Recall Curve</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>RRMS</td>
<td>Relapsing-Remitting Multiple Sclerosis</td>
</tr>
<tr>
<td>SPM</td>
<td>Statistical Parametric Mapping</td>
</tr>
<tr>
<td>SPMS</td>
<td>Secondary Progressive Multiple Sclerosis</td>
</tr>
<tr>
<td>SVD</td>
<td>Cerebral Small Vessel Disease</td>
</tr>
<tr>
<td>T1</td>
<td>T1 Weighted MRI</td>
</tr>
<tr>
<td>T2</td>
<td>T2 Weighted MRI</td>
</tr>
<tr>
<td>TLV</td>
<td>Total Lesion Volume</td>
</tr>
<tr>
<td>TPM</td>
<td>Tissue Probability Map</td>
</tr>
<tr>
<td>WM</td>
<td>White Matter</td>
</tr>
<tr>
<td>WMH</td>
<td>White Matter Hyperintensity</td>
</tr>
<tr>
<td>WML</td>
<td>White Matter Lesion</td>
</tr>
</tbody>
</table>
1. Introduction

Diseases of the central nervous system (CNS) such as multiple sclerosis (MS) and cerebral small vessel disease (SVD) significantly impact affected patients. MS is a demyelinating disease of cerebral white matter (WM), and is the most common autoimmune disorder of the CNS. SVD is a disease of vascular origin (Shi and Wardlaw 2016) which leads to damage in the WM of the brain (Pantoni 2010). Despite differences in their underlying pathophysiology and epidemiology, they may be difficult to reliably distinguish using only neuroimaging (Aliaga and Barkhof 2014) (Geraldes et al. 2016) and clinical features, posing a challenge to the attending clinician. As early diagnosis and treatment may be associated with better long-term outcome in both diseases (Giovannoni et al. 2016) (Issac et al. 2015), and the societal and monetary cost for both diseases is potentially high, particularly for misdiagnosis in MS (Solomon and Corboy 2017), an accurate diagnostic algorithm that distinguishes between the two diseases may be of clinical and socioeconomic value.

From an automatic medical diagnosis viewpoint, distinguishing between MS and SVD is a binary classification problem, with neuroimaging and clinical features as the predictors. Traditional rules-based algorithms and decision processes, including those currently employed by clinicians (J Thompson et al. 2017), may not be capable of entirely capturing the complexity of the classification task (Geraldes et al. 2018), due in part to the heterogeneity of the diseases (Pantoni 2010) (J Thompson et al. 2017). Machine learning provides an alternative decision-making paradigm, in which complex patterns can be learned from data alone. The advantages of machine learning over other approaches, such as rules-based algorithms, becomes particularly apparent as the complexity of the medical diagnosis task increases (Wagholikar, Sundararajan, and Deshpande 2012). This motivates our data-driven approach to automatic medical diagnosis using machine learning.
However, the high representational power of some machine learning algorithms is not without their downside. Indeed, many machine learning algorithms, and more generally clinical decision support systems (CDSSs), have not gained widespread acceptance in the clinical setting because they do not provide explicit reasoning for the classification decisions they make, thus limiting their decision-making interpretability (Vellido, Martín-Guerrero, and Lisboa 2012). This has led to the moniker of the “black box” model, and as such their predictions must not be acted upon in blind faith (Tulio Ribeiro, Singh, and Guestrin 2016). Thus, even for algorithms that generally outperform clinicians in certain tasks, the possibility that the model could make errors that humans would not make means that human supervision cannot be relinquished (Lipton 2016). To gain widespread acceptance in the clinical setting, not only must a CDSS provide an accurate diagnosis, but its decision-making must be readily interpretable.

In this work, we consolidate the requirements for an accurate and interpretable model. Our aims and objectives are:

**Aim:** To develop a machine learning classification algorithm that accurately distinguishes between MS and SVD using neuroimaging and clinical features.

**Hypothesis:** A machine learning classification algorithm that combines neuroimaging and clinical features will result in accurate disease discrimination.

**Objective 1:** To develop an algorithm that automatically segments white matter lesions and T1 black holes on MRI using image processing and statistical & machine learning approaches, and to evaluate model performance using an independent test dataset.

**Objective 2:** To develop a machine learning algorithm that determines the most probable diagnosis using quantitative neuroimaging and clinical features, and to evaluate model performance using an independent test dataset.
Outline: Section 2 is a review of the literature including computational approaches to neuroimaging segmentation and machine learning algorithms for medical diagnosis, and provides an overview of the clinical manifestations of each disease and their clinical discrimination. The methods are presented in section 3, and the results in section 4. We discuss the study’s results and their implications, and provide future directions in section 5, and make concluding remarks in section 6.
2. Literature Review

The neuroimaging markers of MS and SVD and their clinical manifestations will be reviewed. Discriminating features are highlighted. Lesion segmentation techniques and algorithms for automatic diagnosis are presented thereafter.

2.1 Multiple Sclerosis

MS affects over two million people worldwide, with disease prevalence varying considerably with geography (higher prevalence further from the equator) and ethnicity (higher prevalence among Caucasians) (Compston and Coles 2008). Women are affected twice as much as men, although some studies suggest as much as a seven-fold difference between the sexes (Sicotte 2011). Patients are typically diagnosed between the ages of 20 and 50, with up to 6% of patients having clinical onset after the age of 50 (de Seze et al. 2005). Initial diagnosis is rare in children and those over 60 years of age. Life expectancy is reduced in MS patients by as much as 7-14 years, although this figure may vary according to age of diagnosis and rate of disease progression (Scalfari et al. 2013). Early detection and treatment is associated with better long-term outcome (Giovannoni et al. 2016).

Patients can present with a wide variety of clinical symptoms, such as numbness, tingling, weakness, vision loss, gait impairment, incoordination, imbalance and bladder dysfunction (Lublin and Reingold 1996). These symptoms can manifest widely in different parts of the body. There are several types of MS, distinguished by pattern of disease progression. Patients with early clinical signs of MS, and with a single attack, which is signified by worsening of symptoms over a short period of time, have clinically isolated syndrome (CIS). Relapsing-remitting MS (RRMS) is characterized by attacks separated by months to years, with periods of remission in between. Approximately 80-90% of MS patients have RRMS (Gelfand 2014) (Sicotte 2011). 65% of RRMS patients develop secondary progressive MS (SPMS), which is
characterized by progressive worsening of symptoms over time with almost no remissions (Gelfand 2014) (Sicotte 2011). As a result of the more advanced stage of the disease, SPMS patients have higher total lesion loads than RRMS patients (Thompson et al. 1990).

Primary progressive MS (PPMS) is characterized by slowly worsening symptom conditions over time, with little to no remissions, and usually has a later age of onset than RRMS (Sicotte 2011). PPMS accounts for approximately 10-15% of all new MS diagnoses, and prevalence is approximately equal between men and women (Massimo Filippi and Rocca 2011) (Sicotte 2011). Finally, progressive-relapsing MS (PRMS) is characterized by an initially progressive form, but transitions into a relapsing form, and accounts for approximately 5% of cases (Lublin and Reingold 1996).

2.1.1 Multiple Sclerosis – Neuroimaging

In conjunction with clinical features, neuroimaging, most commonly MRI, is used for the diagnosis and disease activity monitoring of MS. For an overview of the physics and engineering underlying MRI acquisition, see (J. L. Prince and Links 2014). Among the most commonly used MRI sequences are T2, fluid-attenuated inversion recovery (FLAIR), and T1. T2 shows abnormalities in 95% of MS patients, emphasizing the importance of neuroimaging in disease detection. On T2 and FLAIR, lesions caused by WM inflammation, known as WM lesions (WMLs), appear hyperintense relative to adjacent normal appearing WM (NAWM). This is due to edema, decreased myelin content and glial scarring (P. M. Matthews et al. 2016). T1 black holes (BHs) are hypointensities on T1 with signal lower than that of grey matter (GM), and signify axonal loss and WM destruction (P. M. Matthews et al. 2016) (Tam et al. 2012). BHs are correlated with disease progression (Tam et al. 2012) and disability measures (Barkhof 1999) (Truyen et al. 1996).

Supratentorial lesions are ovoid in shape and located in the periventricular region (Dawson’s fingers), diffusely in the juxtacortical region (U-fibers), as well as throughout the cerebral WM
(P. M. Matthews et al. 2016). The corpus callosum is usually affected (Gean-Marton et al. 1991), as well as the WM near the trigone and temporal horns of the lateral ventricles (Aliaga and Barkhof 2014). In the infratentorial region, the floor of the fourth ventricle, cerebellar peduncles, medulla oblongata, and pons are affected (Aliaga and Barkhof 2014). Lesions are distributed bilaterally (Aliaga and Barkhof 2014).

A lack of clinical symptoms and a normal neurological examination, paired with MRI findings indicative of MS, has been termed preclinical MS (Sicotte 2011). Importantly, a third of patients presenting with T2 lesions but with normal neurological examination develop CIS over the course of two years (Lebrun et al. 2008) (Okuda et al. 2009). This emphasizes the importance of using MRI as an early diagnostic measure (Sicotte 2011), not least because early treatment is associated with slower disease progression in RRMS, and reduced risk of developing clinically-defined MS in those presenting with CIS (P. M. Matthews et al. 2016). Furthermore, T2 lesion load has been reported to increase by as much as 5-10% per year in patients with RRMS and SPMS (Paty et al. 1994), emphasizing the importance of early detection using MRI. Finally, MRI detects inflammatory activity at a rate seven to ten times faster than clinical events (Miller et al. 1988), which are often infrequent and ill-defined (Frank et al. 1994), making MRI an invaluable tool in the diagnosis of MS.

Reducing acquisition slice thickness, from 5 mm to 3 mm, has been shown to increase the detection of small lesions significantly (M Filippi et al. 1995), important because white matter lesion load is correlated with physical disability (Stankiewicz et al. 2011). Therefore, this study makes use of sequences with slick thicknesses no greater than 3 mm for MS patients. Lesion volume detection also depends on the magnetic field strength of the MRI scanner, with increases in detection between 1.5T and 3T scanners, particularly in the periventricular WM, as well as the cortical and juxtacortical regions (Stankiewicz et al. 2011). Consequently, we employ 3T magnetic field strength scanners in this study.
Although cortical lesions have been associated with disability progression and cognitive impairment (Calabrese, Filippi, and Gallo 2010) (Massimo Filippi and Rocca 2011), their detection is made difficult for a number of reasons. These include their small size, the need for very high magnetic field strength scanners which may not be available in most clinical centers, partial volume effects from CSF, and low contrast between cortical lesions and normal appearing GM (NAGM) (Calabrese, Filippi, and Gallo 2010) (Massimo Filippi and Rocca 2011) (P. M. Matthews et al. 2016). Therefore, the analysis of cortical lesions is excluded from this study.

2.2 Cerebral Small Vessel Disease

SVD is a contributing factor of stroke, both ischemic and haemorrhagic, as well as dementia (Pantoni 2010). It has an incidence rate of 5-10% in patients 20-40 years old, and is very common in adults over the age of 60, with prevalence reaching nearly 100% (Charil et al. 2006). In addition to cognitive impairment, increased risk of poor gait and balance, depression, and other physical disabilities is high. Early diagnosis and treatment may lead to better patient outcome, and the disease may be partially reversible if treated early (Issac et al. 2015).

2.2.1 Cerebral Small Vessel Disease – Neuroimaging

On neuroimaging, SVD presents with WM hyperintensities (WMHs), small subcortical infarcts, lacunes, visible perivascular spaces, cerebral microbleeds, and brain atrophy (Wardlaw, Smith, and Dichgans 2013). Due to the prevalence of WMHs in SVD, T1, T2 and FLAIR sequences are commonly used. As with MS, WMHs appear hyperintense on T2 and FLAIR, and WM destruction appears hypointense on T1. WMHs are distributed bilaterally, with varying shape depending on severity (Shi and Wardlaw 2016), although a rounded appearance is common (Wardlaw, Smith, and Dichgans 2013). Lesions are primarily supratentorial (Pantoni 2010), and can be found in the periventricular region and deep WM, the basal ganglia, and in cerebral WM (Wardlaw, Smith, and Dichgans 2013). Lacunes have reduced signal on T1 and as with BHs in MS, can in addition to WMHs, be indicative of disease
progression (Wardlaw, Smith, and Dichgans 2013) and neurological disability (Pantoni 2010). Lacunes can be round or ovoid, and are typically between 3-15 mm in diameter (Shi and Wardlaw 2016).

Lesions are more easily discerned on 3T scanners than 1.5 T scanners (Wardlaw et al. 2013). Due to the small size of certain lesions, a 5 mm slice thickness is recommended, with 1 mm in plane resolution (Wardlaw et al. 2013). However, if slice thickness is less than 2 mm, signal-to-noise ratio is compromised, which can complicate lesion identification (Salamon 2014). Therefore, this study employs 3T magnetic field strength scanners with slice thicknesses no greater than 5 mm for SVD patients.

2.3 Differences between Multiple Sclerosis and Cerebral Small Vessel Disease

Although MS and SVD have similar presentation, evaluating their neuroimaging and clinical features in context can help in their discrimination. Table 1 and Table 2 contrast the clinical and neuroimaging features of MS and SVD.

| Table 1 Clinical features known to distinguish multiple sclerosis and cerebral small vessel disease. |
|---|---|---|---|
| **Prevalence** | Multiple Sclerosis: ~ 0.1% | Small Vessel Disease: > 95% |
| Age at first diagnosis (Majority of Patients) | 20 - 40 | > 60 |
| Age at first diagnosis (Minority of Patients) | > 50 | 20 - 40 |
| Sex (M:F) | 1:2 | 1:1 |
| Numbness, Tingling, Weakness | ✓ | ✓ |
| Vision Loss | ✓ | ✓ |
| Gait Impairment | ✓ | ✓ |
| Incoordination, Imbalance | ✓ | ✓ |
| Bladder Dysfunction | ✓ | ✓ |
| Dementia | ✓ | ✓ |
| Stroke | ✓ | ✓ |
Table 2 Neuroimaging features known to distinguish multiple sclerosis and cerebral small vessel disease.

<table>
<thead>
<tr>
<th>Lesion Location</th>
<th>Multiple Sclerosis</th>
<th>Small Vessel Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juxtacortical</td>
<td>Common</td>
<td>Less Common</td>
</tr>
<tr>
<td>Temporal Lobe</td>
<td>Common</td>
<td>Less Common</td>
</tr>
<tr>
<td>Corpus Callosum</td>
<td>Common</td>
<td>Less Common</td>
</tr>
<tr>
<td>Basal Ganglia</td>
<td>Less Common</td>
<td>Common</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>Common</td>
<td>Less Common</td>
</tr>
<tr>
<td>Cerebellar Peduncles</td>
<td>Common</td>
<td>Less Common</td>
</tr>
<tr>
<td>Pons</td>
<td>Common</td>
<td>Common</td>
</tr>
</tbody>
</table>

Lesion Distribution

<table>
<thead>
<tr>
<th></th>
<th>Bilateral</th>
<th>Bilateral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supratentorial</td>
<td>Supratentorial &gt;</td>
<td>Supratentorial &gt;&gt;</td>
</tr>
<tr>
<td>Infratentorial</td>
<td>Infratentorial</td>
<td>Infratentorial</td>
</tr>
</tbody>
</table>

The criteria used for the diagnosis of MS are highly sensitive, but have low precision – they can reliably identify patients with MS, but should not be used to distinguish MS from disorders with similar presentation (Brownlee et al. 2017) (Geraldes et al. 2018), among which SVD is particularly common (Aliaga and Barkhof 2014). Although alternative tests can aid the attending clinician in such cases (Geraldes et al. 2018), these tests may not be widely available. Increased patient age further complicates diagnosis, as WMHs of vascular origin can contribute to total lesion load. Finally, studies which have examined the MRI criteria of MS have typically excluded patients over the age of 50 (Brownlee et al. 2017), leading to increased uncertainty in the discrimination of MS and SVD in older patients.

When making the diagnosis of MS, there exists a tension between the cost due to delayed diagnosis, as early treatment may be associated with better long-term outcome, and the cost associated with misdiagnosis and consequently unnecessary treatment, which can cause considerable harm and even death to the patient (Solomon and Corboy 2017). The nature of this tension may differ for SVD, as the need for early treatment is less emphasized. This presents a classic trade-off between the false positive and false negative rate of a diagnostic test, and calls for careful consideration of the associated costs in context of the diseases, potential costs to the healthcare system, and the combined expertise of physicians and CDSSs to make a diagnosis. For example, if the medical diagnosis test is used for screening, then low false negative rates are desirable.
2.4 Automatic Brain Tissue and Lesion Segmentation

Automatic segmentation encompasses the set of algorithmic approaches to distinguishing brain regions of interest on neuroimaging without manual intervention. We investigate techniques for segmenting brain tissue, such as GM, WM, and CSF, and for segmenting WMLs & WMHs, and BHs & lacunes. Due to their similar presentation, we will not draw any distinction between approaches for WML and WMH segmentation, or between BH and lacune segmentation.

The automatic segmentation of WMLs has been the focus of several research groups (García-Lorenzo et al. 2013) (Lladó et al. 2012) over the past decade. Commonly employed techniques include unsupervised learning algorithms such as the Gaussian mixture model (GMM), Markov random fields (MRFs) and outlier detection, and supervised learning algorithms such as the k-nearest neighbors (k-NN), random forests and adaboost, and neural networks, particularly those with deep architectures. More conventional image processing techniques, such as thresholding algorithms, edge detection and texture analysis, have also been employed. The following subsections will explore these techniques in some detail. We also investigate lesion segmentation algorithms for BHs and lacunes, although WML segmentation has received more attention in the research community.

2.4.1 Mixture Model & Expectation Maximization Algorithms

Mixture models are an intuitive choice for segmentation tasks, such as brain tissue segmentation and WML segmentation, since voxel intensity can distinguish tissue classes. The number of mixture components is usually designated by the experimenter, who knows a priori how many mixtures are appropriate. Spatial prior distributions for GM, WM, and CSF can guide the segmentation algorithm.
The prototypical mixture model is the GMM with \( K \) user-designated classes, which in the case of GM, WM, and CSF segmentation, corresponds to \( K = 3 \). The \( k^{th} \) mixture component has probability density function

\[
f(x^{(n)}; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp \left[ -\frac{1}{2} (x^{(n)} - \mu_k)^T \Sigma_k^{-1} (x^{(n)} - \mu_k) \right]
\]

where \( x^{(n)} \) is the feature vector corresponding to the \( n^{th} \) observation, \( \mu_k \) is the mean of the \( k^{th} \) component, \( \Sigma_k \) is the covariance matrix of the \( k^{th} \) component, and \( d \) is the dimensionality of the feature space. If voxel intensity is the only feature used, then \( d = 1 \), which reduces the model to the univariate case. Although using additional features, such as the 3D coordinates of the voxels, can increase the descriptiveness of the feature space, this can be problematic due to the curse of dimensionality.

The algorithm begins by initializing \( k \) mixture components, each parameterized by a mean and covariance. The initial parameterization can be random, or a clustering algorithm such as k-means can be used for initialization. This step can be important, as mixture models are known to be sensitive to their initial parametrization. Alternatively, several models with different random initializations can be evaluated, and thereafter the model with the highest log likelihood is selected. A final alternative is to use prior spatial probability maps for initialization, the advantage of this approach being that prior knowledge of GM, WM, and CSF spatial distributions can facilitate segmentation.

The expectation maximization (EM) algorithm adjusts the parameters of each mixture component so that the probability that the mixtures would have generated the data assigned to them is maximized. In the expectation (E) step, we calculate the probability that each mixture was responsible for generating the observations \( x^{(n)} \), given parameters \( \mu_k, \Sigma_k \) and \( \pi_k \) (the latter term being the weighting coefficient, which is the probability that the data was sampled from a particular mixture component). In the maximization (M) step, the parameters \( \mu_k, \Sigma_k \) and \( \pi_k \) are re-estimated to maximize the likelihood function. To summarize, we optimize the
parameters of each mixture component to maximize the probability that they would together generate the data. The E and M steps are repeated iteratively, until convergence or until a desired level of performance is reached. We leave the derivation of the EM algorithm for GMMs to any standard reference on machine learning, as previously published (Bishop 2006) (Murphy 2012).

Concretely, we initialize the model with a set of parameters \((\mu_k, \Sigma_k, \pi_k)\), subject to \(\sum_{k=1}^{K} \pi_k = 1\), then iterate until convergence using the following two steps:

1. **E-step**: Assign responsibilities \(r_k^{(n)}\) to each of the \(k\) mixture components for each of the \(n\) data points:

   \[
   r_k^{(n)} = \frac{\pi_k f(x^{(n)}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f(x^{(n)}; \mu_j, \Sigma_j)}
   \]

2. **M-step**: Re-estimate parameters given current responsibilities

   \[
   \mu_k = \frac{1}{\sum_{n=1}^{N} r_k^{(n)}} \sum_{n=1}^{N} r_k^{(n)} x^{(n)}
   \]

   \[
   \Sigma_k = \frac{1}{\sum_{n=1}^{N} r_k^{(n)}} \sum_{n=1}^{N} r_k^{(n)} (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T
   \]

   \[
   \pi_k = \frac{\sum_{n=1}^{N} r_k^{(n)}}{N}
   \]

3. After the E and M steps, evaluate the log likelihood to check for convergence

   \[
   \log L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k f(x^{(n)}; \mu_k, \Sigma_k)
   \]

### 2.4.2 Markov and Conditional Random Fields

The seminal paper of (Geman and Geman 1984) demonstrated how principles of statistical mechanics could translate into robust methods for image segmentation. These methods are now
collectively known as Markov random fields (MRFs), which as their core principle encourage smooth segmentations by minimizing an energy function – a measure of the dissimilarity between a pixel’s current classification and that of its neighbors. MRFs are typically used as a segmentation refinement tool; whereby the initial segmentation is often produced using conventional image processing techniques (Kato and Pong 2001), as well as other machine learning algorithms (Roy et al. 2015). After preliminary exploration of MRFs in this work, we concluded that their benefits of encouraging smooth segmentations were not appropriate, as small lesions with irregular shapes were sometimes erroneously discarded. However, a more thorough investigation may be warranted for future work. Furthermore, conditional random fields (CRFs), the discriminative counterparts of MRFs (S. J. D. Prince 2012), have shown promise in preserving edge detail in 3D medical image segmentation (Kamnitsas et al. 2017). They have also shown promise in detecting gadolinium-enhancing lesions (Karimaghaloo et al. 2012), which is of value due to the prevalence of contrast agents in MRI imaging.

2.4.3 Deep Learning

Convolutional neural networks (CNNs), a type of artificial neural network (ANN) loosely based on the mammalian visual cortex, have drawn much attention in the last half-decade due to their remarkable image classification & segmentation performances (Long, Shelhamer, and Darrell 2014). They have furthermore demonstrated significant success in WML segmentation for MS (Brosch et al. 2016) (Valverde et al. 2017), and SVD (Ghafoorian et al. 2016) (Li et al. 2018). However, CNNs typically require large training times, are computationally expensive, perform best when evaluated for a single disease, and without modification can only be used on images of the same size as those used in training. In this work, we sought to develop algorithms which are computationally inexpensive, require little to no hyperparameter tuning, can be used for the segmentation of WMLs on both MS and SVD, and can be applied to images of arbitrary size and resolution. Nevertheless, we note that deep learning approaches are the most promising technique for future direction.
2.4.4 Other Approaches

Although edge detection techniques have shown strong performance on tasks with synthetic lesions, we found their performance to be questionable on natural images, where WMLs are demarcated by fuzzy edges at best (Roy et al. 2015). We investigated texture analysis techniques, but found those too suffer from the heterogeneity of texture within WMLs. We found that thresholding techniques, such as Otsu’s method (Otsu 1979), were poor for both brain tissue segmentation and lesion segmentation, due to the method’s assumption that groups are normally distributed with equal within-group variances (Xue and Titterington 2011). This assumption is particularly detrimental when the groups have unequal sizes, as is very often the case for WMLs (relatively few) vs. NAWM (very many in comparison). Minimum-error thresholding techniques (Kittler and Illingworth 1986), and entropic thresholding techniques (Kapur, Sahoo, and Wong 1985), also performed poorly, for reasons similar to those listed above.

In comparison to brain tissue and WML segmentation, we found comparatively little literature for BH or lacune segmentation. The approaches we did find employed contour models (Tam et al. 2012), as well as morphological operations (Y. Wang et al. 2012) and fuzzy connectivity (Datta et al. 2006). More recently, CNNs were employed to detect lacunes in SVD (Ghafoorian et al. 2017). However, as these approaches made use of T2 or FLAIR sequences to assist in the segmentation, and some were not fully automatic, we sought to develop novel and automatic BH segmentation algorithms that could be performed using T1 sequences alone.

2.5 Co-registration

Co-registration is a technique used to create a common space for all images, so that inter and intra-patient comparisons can be made. Conventionally, co-registration software is used to apply an affine transformation to the patient’s native space image, with the same transformation subsequently applied to corresponding lesion masks. The template space is developed using a combination of healthy controls - this study made use of a T1 Montreal
Neurological Institute (MNI) template, derived from 152 healthy patients who were age-matched with our study’s patients.

Since template space images are formed using healthy controls, WMLs and BHs pose a challenge to co-registration (Ashburner and Friston 2000) (Tam et al. 2012), as the intensity distribution of lesions skews the co-registration. Lesion filling is an approach used to replace WMLs and BHs with normal appearing tissue prior to co-registration. Once lesion filling is complete, co-registration is performed using interpolation and re-slicing algorithms. A commonly used toolbox for image co-registration is the statistical parametric mapping (SPM12) package, which is available as an extension to MATLAB (“Statistical Parametric Mapping” 2014). It employs normalized mutual information for co-registration estimation (Studholme, Hill, and Hawkes 1999) and uses a 4th degree B-spline function by default for re-slicing.

2.6 Automatic Algorithms for Diagnosis

We review a small but important subset of algorithms used in the automatic medical diagnosis domain and explore how well they are suited for our prediction task.

2.6.1 Logistic Regression

Logistic regression has been a popular classification algorithm in various clinical domains. This is in part due to its higher interpretability in comparison to other inference algorithms - for example, the natural exponent of the model coefficients can readily be interpreted as the odds ratios (Szumilas 2010). As a discriminate model, it estimates the relation between predictors and class labels directly. Therefore, it enjoys the benefits associated with discriminative models, such as better asymptotic performance using fewer model parameters than generative models (Ng and Jordan 2002).
We provide a brief derivation of the logistic regression model for the binary classification case and explore some of the challenges associated with the algorithm in practice. We omit some details from the derivation, and direct the interested reader to any standard textbook on machine learning for further reference (Bishop 2006) (Murphy 2012).

Let $\sigma(z)$ denote the logistic sigmoid function

$$
\sigma(z) = \frac{1}{1 + \exp(-z)}
$$

Then the logistic regression model, parameterized by $\beta, \beta_0$ (the $\beta$ are known as the weights, and $\beta_0$ is the bias/intercept), gives the probability of $x^{(n)}$ belonging to class $C = 0$ as

$$
p(C = 0|x^{(n)}; \beta, \beta_0) = \sigma(\beta^T x^{(n)} + \beta_0)
$$

and therefore, the probability for class $C = 1$ is given by

$$
p(C = 1|x^{(n)}; \beta, \beta_0) = 1 - \sigma(\beta^T x^{(n)} + \beta_0)
$$

Our objective is to optimize the parameters such that the likelihood

$$
L(\beta, \beta_0) = \prod_{n=1}^{N} p(y^{(n)}|x^{(n)}; \beta, \beta_0)
$$

is maximized, where $y^{(n)}$ denotes the actual class label for $x^{(n)}$. Alternatively, we can minimize the negative log likelihood (the log is taken for convenience and numerical reasons), defined as

$$
\zeta(\beta, \beta_0) = -\log(L(\beta, \beta_0))
$$

where $\zeta$ is the cost. The cost can be re-written as

$$
\zeta(\beta, \beta_0) = -\sum_{n=1}^{N} y^{(n)} \log \left(1 - p(\hat{y} = 0|x^{(n)}; \beta, \beta_0)\right) - \sum_{n=1}^{N} (1 - y^{(n)}) \log \left(p(\hat{y} = 0|x^{(n)}; \beta, \beta_0)\right)
$$
which is also known as the cross entropy error function. What remains is estimating the parameters $\beta, \beta_0$, which can be done using first or second-order iterative optimization algorithms, such as gradient descent and Newton’s method, respectively. For gradient descent, the updates are

$$
\begin{align*}
\beta &\leftarrow \beta - \lambda \frac{\partial \text{loss}}{\partial \beta} \\
\beta_0 &\leftarrow \beta_0 - \lambda \frac{\partial \text{loss}}{\partial \beta_0}
\end{align*}
$$

where $\lambda$ is the learning rate, a model hyperparameter. Since this is a convex optimization problem in $\beta$, the loss has a single minimum - the global minimum.

To prevent overfitting to the training data, and to avoid unstable parameter estimates in the case of completely separable or quasi-separable classes, a regularization term $\alpha$, which is also a model hyperparameter, is often included in the logistic regression model. In this case, the likelihood function is given by

$$
L(\beta, \beta_0) = p(\beta) \prod_{n=1}^{N} p(y^{(n)}|x^{(n)}; \beta, \beta_0)
$$

where

$$
p(\beta) = N(\beta; 0, \alpha^{-1}I) = \frac{\alpha}{\sqrt{2\pi}} \exp\left(-\frac{\alpha}{2} \beta^T \beta\right)
$$

This can be interpreted as placing a Gaussian prior with zero mean and precision $\alpha$ on the model weights and is equivalent to penalizing the Euclidean norm of the model weights. The approach is also known as ridge regression or logistic regression with L2 regularization. Importantly, when such regularization is employed, feature scaling is needed. However, for some predictors there may be no well-established method for feature scaling – for example in the case of scaling patient age. Furthermore, feature scaling lowers model interpretability due to its effect on the model coefficients. This presents an important trade-off in model selection, that of interpretability vs. effectiveness.
2.6.2 Naïve Bayes

As generative models, the Naïve Bayes family of algorithms model the joint probability of the inputs and class labels and use Bayes’ theorem to compute the posterior probability of a class given input data and class priors. Like the logistic regression model, the Naïve Bayes family of classifiers can be easily interpreted. They make the simplifying assumption that the features are independent given the class. Although this assumption may weaken model power, under many circumstances it does not affect classifier performance (Zhang 2004), and furthermore may increase model generalizability when the dataset is small (Hand and Yu 2001). This is important in our work as we, like many studies concerned with automatic medical diagnosis, are limited to a small dataset. Furthermore, the independence of features helps limit the curse of dimensionality. Despite their relatively strong performance in classification tasks, Naïve Bayes algorithms are poor “soft” classifiers, as their class probability estimates are often biased towards 0 or 1 (Hand and Yu 2001) (Niculescu-Mizil and Caruana 2005). Therefore, the model should be investigated for classification purposes. If interpreting an accurate probability of diagnosis is needed, probability calibration approaches must be investigated, such as binning (Zadrozny and Elkan 2001) or Platt scaling and isotonic regression (Niculescu-Mizil and Caruana 2005).

We provide a brief derivation of the model, noting its important properties as we proceed. Recall Bayes’ rule, which states that

\[
p(C^{(n)} = k|x^{(n)}) = \frac{p(C^{(n)} = k)p(x^{(n)}|C^{(n)} = k)}{p(x^{(n)})}
\]

where there are \( K \) classes, \( y^{(n)} \) is the actual label for the \( n^{th} \) data point, and \( x^{(n)} = x_1^{(n)}, ..., x_d^{(n)} \) is the corresponding \( d \)-dimensional input vector. The independence of feature assumption implies that

\[
p(x^{(n)}|C^{(n)} = k) = \prod_{j=1}^{d} p(x_j^{(n)}|C^{(n)} = k)
\]
Observing that the evidence \( p(x^{(n)}) \) is constant given the input, and thus has no effect on the parameter estimates, we can simplify the formulation to

\[
p(C^{(n)} = k | x^{(n)}) \propto p(C^{(n)} = k) \prod_{j=1}^{d} p(x_j^{(n)} | C^{(n)} = k)
\]

which is composed of the class prior probabilities \( p(C^{(n)} = k) \) and the likelihood \( \prod_{j=1}^{d} p(x_j^{(n)} | C^{(n)} = k) \). The class prior probabilities can be estimated using the proportion of each class in the training set, or alternatively can be fixed in advance. We will come back to this point in the Methods section.

What remains is predicting the class of \( x^{(n)} \), denoted by \( \hat{y}^{(n)} \), according to the candidate class \( C^{(n)} = k \) with the highest posterior probability. This is known as the maximum a posteriori classification rule

\[
\hat{y}^{(n)} = \arg\max_{k \in \{1, \ldots, K\}} p(C^{(n)} = k) \prod_{j=1}^{d} p(x_j^{(n)} | C^{(n)} = k)
\]

For binary predictors, the Bernoulli Naïve Bayes classifier is used to model the likelihood of each feature given the class:

\[
p(x_j^{(n)} | C^{(n)} = k) = \theta_{j,k}^{x_j^{(n)}} (1 - \theta_{j,k})^{1-x_j^{(n)}}
\]

where \( N_k \) is the count of class \( k \) in the training set, and \( \theta_{j,k} \) is the probability of the \( k^{th} \) class for the \( j^{th} \) predictor. Its maximum likelihood estimate is given by

\[
\hat{\theta}_{j,k} = \frac{\sum_{n} x_j^{(n)}}{N_k}
\]

which has the intuitive interpretation of the proportion of times the \( j^{th} \) feature occurs for the \( k^{th} \) class in the training set.
For continuous predictors, a Gaussian distribution is used to model the likelihood of each feature given the class

\[ p(x_j^{(n)} | C^{(n)} = k) = \frac{1}{\sqrt{2\pi\sigma_{j,k}}} \exp \left[ -\frac{1}{2\sigma^2_{j,k}} (x_j^{(n)} - \mu_{j,k})^2 \right] \]

where \( \mu_{j,k} \) and \( \sigma_{j,k} \), the mean and standard deviation for the \( j^{th} \) feature and \( k^{th} \) class, are estimated using maximum likelihood

\[
\hat{\mu}_{j,k} = \frac{\sum_n^{N_k} x_j^{(n)}}{N_k}, \hat{\sigma}_{j,k} = \frac{\sum_n^{N_k} (x_j^{(n)} - \hat{\mu}_{j,k})^2}{N_k}.
\]

Intuitively, these estimates give the sample mean and variance of the training set for the \( j^{th} \) feature and \( k^{th} \) class. We denote this model as the Gaussian Naïve Bayes classifier.

Although the maximum likelihood estimates of model parameters are consistent (Newey and McFadden 1994), they suffer for small datasets because they are determined by a limited number of observations, which can lead to overfitting. We will address this problem by employing a Bayesian approach in the Methods section and demonstrate its impact on model performance in the Results section.

Finally, Bayesian networks, a family of models which encompasses the Naïve Bayes classifier as a special case, can be used to capture dependencies between input features. For example, dependencies between age and total lesion volume could be accounted for by the tree-augmented Naïve Bayes classifier (Friedman, Geiger, and Goldszmidt 1997). We leave their study to future work, as a simple extension of the models developed here.
2.6.3 Artificial Neural Networks

Artificial neural networks (ANNs) have enjoyed a large surge in popularity for supervised learning tasks in the past decade, as the computing power and dataset sizes necessary for their effective use have increased. Using this model, each patient’s neuroimaging and clinical features would correspond to input features, and the ANN output would be the disease classification probability. The backpropagation algorithm would be used to learn the weights that minimize the cross entropy loss between predicted class probabilities and labels.

ANNs can identify abstract and conceptually hierarchical relations between input features, explaining their success in very complex classification tasks involving multiple features. However, since we aim to significantly reduce the dimensionality of the neuroimaging and clinical features using statistical & machine learning approaches, both to improve model generalization and for model interpretability purposes (Vellido, Martín-Guerrero, and Lisboa 2012), the representative power of ANNs may be superfluous. There are other challenges that would complicate their use. First, ANNs are prone to overfitting to training data, which is particularly problematic given our small dataset size. Second, ANNs typically have several hyperparameters that must be tuned. These include the number of hidden layers, the number of units per layer, the activation function between hidden layers, the learning rate, the weight decay rate, batch size during training, and dropout rate in each of the layers. Nevertheless, we expect ANN performance would scale and outperform other models with increasing dataset size, and advances in automatic hyperparameter optimization (Snoek, Larochelle, and Adams 2012) (Snoek et al. 2015) may render the task less laborious.

2.7 Related Work

Here, we review three paradigms for automatic disease discrimination and diagnosis in the context of medical imaging. Our approach differs from these paradigms in that we use only lesion location and its associated probability in a given disease to distinguish between diseases.
2.7.1 Feature Extraction Approaches

Feature extraction approaches revolve around extracting features from regions of interest and using these to distinguish diseases. In (Vemuri et al. 2008), a support vector machine classifier was trained using features extracted from brain matter and patient demographics and clinical features to distinguish between Alzheimer’s patients and healthy controls. In (Toews et al. 2010), feature-based morphometry, a probabilistic approach to distinguishing between disease and non-disease using scale-invariant image features, is used to quantify feature variability between subject groups, which enables disease discrimination. In (Rana et al. 2015), features extracted from five regions affected in Parkinson’s disease are used to train a support vector machine classifier to distinguish healthy controls from patients with Parkinson’s disease.

2.7.2 Image Retrieval & Similarity Approaches

These approaches use a distance metric to calculate the similarity between images within and between disease groups. See (Akgül, Ünay, and Ekin 2009) as an example of using a nearest neighbor classifier to detect Alzheimer’s disease.

2.7.3 Deep Learning Approaches

In this paradigm, deep learning is used to compute a function between input image features and disease categories. In (Zou et al. 2017), 3-D convolutional neural networks are applied to functional and structural MRI to detect attention deficit hyperactivity disorder (ADHD). In (Yoo et al. 2018), unsupervised deep learning is used to learn latent spatial features from multimodal MRI to detect MS in its early stage.
3. Methods

*Objective 1:* To develop an algorithm that automatically segments white matter lesions and T1 black holes on MRI using image processing and statistical & machine learning approaches.

*Objective 2:* To develop a machine learning algorithm that determines the most probable diagnosis using quantitative neuroimaging and clinical features.

We divide this section in two parts: the analysis of neuroimaging, and subsequently its use in combination with clinical features for developing the diagnostic algorithm. Figure 1 provides a schematic of the methodology.
Figure 1 Methodology. The left side represents the image processing & analysis component. The right side represents the aggregation of neuroimaging and clinical features, as well as the statistical analysis.

3.1 Neuroimaging

This subsection outlines the novel segmentation algorithms developed, the creation of the combined probability maps, and the derivation of the image similarity metrics.
3.1.1 Patient Cohort

The dataset consists of 39 RRMS (median age 48 (31 – 60), 12 males and 27 females), 44 SPMS (median age 55.5 (38 – 82), 17 males and 27 females) and 72 SVD patients (median age 73 (52 – 86), 44 males and 28 females), all of which were administered at the same center.

3.1.2 Image Acquisition and Ground Truth Tracings

For RRMS, the MRI modalities employed were FLAIR (3T, 2D, 3 mm slice thickness) and pre-contrast T1 (3T, 3D, 1.2 mm slice thickness). For SVD, the MRI modalities employed were FLAIR (3T, 2D, 5 mm slice thickness) and pre-contrast T1 (3T, 3D, 1.4 mm slice thickness).

Lesion tracings for the MS cases were done by a neuro-radiologist. BHs were traced in T1 space. The proton density T2 W and FLAIR sequences, along with their associated ground truth tracings, were co-registered to T1 space, as this target space resulted in the best co-registration. Figure 2a shows the FLAIR scan of a cognitively impaired MS patient in T1 space. Figure 2b shows the same patient’s T1 scan. Figure 2c shows the ground truth tracing.

![Figure 2](image)

*Figure 2 a) FLAIR image of an RRMS impaired patient. Arrows point to white matter lesions. b) T1 image of the same patient. Arrow points to a T1 black hole. c) Ground truth segmentation*
(green = grey matter, blue = white matter, purple = basal ganglia, orange = thalamus, beige = white matter lesion, white = T1 black hole).

3.1.3 Brain Extraction & Bias Field Correction

Using default settings, SPM’s brain tissue segmentation module was used for brain extraction on T1 (“Statistical Parametric Mapping” 2014). Voxels with \( p_{GM} + p_{WM} + p_{CSP} < 0.5 \), where \( p \) corresponds to tissue probability, were assigned to background intensity. Figure 3 shows the effect of brain extraction on a T1 image. Bias field correction was applied to T1 (“Statistical Parametric Mapping” 2014) and FLAIR (Schmidt 2017) images, to eliminate spatially varying intensity across the brain due to field inhomogeneities. The result is a more uniform distribution of intensity within a given tissue class across the brain. Figure 4 shows the effect of bias field correction on a T1 image, illustrated with a modified color scheme to make the effect clearer.

We did not employ intensity normalization, as the approach is sensitive to scanner type and differences in acquisition parameters; scans can present with orders of magnitude differences in intensity, as well as differing contrast between brain tissue. Furthermore, as our method, described below, is robust to the intensity scale of brain tissue (R. Wang et al. 2015), normalization was not required.
3.1.4 Automatic Lesion Segmentation

We outline our methodology for WML segmentation on FLAIR, which uses a mixture of $t$-distributions algorithm as an intermediate step. A novel spatial-heuristic segmentation algorithm was developed to identify BHs on T1. We outline our segmentation refinement techniques thereafter.
3.1.4.1 White Matter Lesion Segmentation

We begin by describing our algorithm for brain tissue segmentation on FLAIR. Despite their well-understood properties and relative simplicity, GMMs are not always an appropriate choice of model for segmentation tasks, as they are not robust to outliers and to non-normality in class distributions. To represent outliers, some practitioners have compensated by adding an additional mixture component to the GMM. However, there is little theoretical justification for this approach, and it is not expected to work in cases where the noise is not uniform (Peel and McLachlan 2000). Although the central limit theorem states that a linear combination of random variables tends toward a normal distribution in the limit of infinite samples, the samples must be independently and identically distributed. This is not the case for the distribution of voxel intensities in the brain, as brain regions are inter-related, motivating the need for an alternative model.

The automatic segmentation literature is rich in the variety of mixture models employed; for example see (R. Wang et al. 2015) for their use of the Gumbel and Fréchet distributions. However, we investigated the mixture of $t$-distributions, as this model remedies the problems associated with GMMs. In contrast to the exponentially decaying tails of the normal distribution, the tails of the $t$-distribution are heavy, making it less sensitive to outliers, and thus more robust for brain tissue segmentation problems with hyperintense lesions. This is particularly important for the segmentation of FLAIR sequences into GM, WM and CSF, as the distribution of these tissue classes can contain a high proportion of outliers, and violate the assumptions of normality. Furthermore, as the degrees of freedom $v$ increase, the $t$-distribution approaches the normal distribution, thus making the mixture of $t$-distributions a useful generalization of the GMM. As the mixture of $t$-distributions is closely related to the GMM, the model gives tractable maximum likelihood estimates using the EM algorithm (Peel and McLachlan 2000) (Sfikas, Nikou, and Galatsanos 2007). There has been interest in further extending the statistical robustness of mixture models, such as with skew-normal and skew-$t$ distributions, however parameter estimation in such techniques is typically computationally expensive, and not necessarily appropriate for our task.
Analogous to GMMs, we proceed by defining the probability density function of the \( t \)-distribution, which is given by

\[
f(x^{(n)}; \mu_k, \Sigma_k, v_k) = \frac{\Gamma\left(\frac{v_k + d}{2}\right)}{(\pi v_k)^{d/2}} \left| \Sigma_k \right|^{-\frac{1}{2}} \frac{1}{\left[1 + \frac{1}{v_k} \delta(x^{(n)}, \mu_k; \Sigma_k)^2\right]^{\frac{v_k + d}{2}}}
\]

where \( \delta(x^{(n)}, \mu_k; \Sigma_k) = (x^{(n)} - \mu_k)^T \Sigma_k^{-1} (x^{(n)} - \mu_k) \) is the Mahalanobis distance and \( \Gamma \) is the Gamma function.

Analogous to the E and M steps of the GMM, we compute the following updates:

1. E-step

\[
r_k^{(n)} = \frac{\pi_k f(x^{(n)}; \mu_k, \Sigma_k, v_k)}{\sum_{j=1}^{K} \pi_j f(x^{(n)}; \mu_j, \Sigma_j, v_j)}
\]

\[
u_k^{(n)} = \frac{v_k + d}{v_k + \delta(x^{(n)}, \mu_k; \Sigma_k)}
\]

where \( u_k^{(n)} \) acts as a weighting coefficient, such that outliers have reduced weight on parameter updates during the M-step

2. M-step

\[
\mu_k = \frac{\sum_{n=1}^{N} r_k^{(n)} u_k^{(n)} x^{(n)}}{\sum_{n=1}^{N} r_k^{(n)} u_k^{(n)}}
\]

\[
\Sigma_k = \frac{\sum_{n=1}^{N} r_k^{(n)} u_k^{(n)} (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T}{\sum_{n=1}^{N} r_k^{(n)}}
\]

\[
\pi_k = \frac{\sum_{n=1}^{N} r_k^{(n)}}{N}
\]

And the updated estimate of \( v_k \) is given by the solution to \( v_k^{+1} \) in the equation below, which is not available in analytical form, and thus must be solved numerically, using for example the Newton-Raphson method.
\[
\log \left( \frac{v_{k+1}}{2} \right) - \psi \left( \frac{v_{k+1}}{2} \right) + 1 - \log \left( \frac{v_k + d}{2} \right) + \frac{\sum_{n=1}^{N} r_k^{(n)} \left( \log u_k^{(n)} - u_k^{(n)} \right)}{\sum_{n=1}^{N} r_k^{(n)}} + \psi \left( \frac{v_k + d}{2} \right) = 0
\]

where \( \psi \) is the digamma function.

3. After the E and M steps, the log likelihood is evaluated to check for convergence.

\[
\log L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k f(x^{(n)}; \mu_k, \Sigma_k)
\]

We employ the mixture of \( t \)-distributions for the segmentation of FLAIR into GM, WM, and CSF, and thereafter assign lesion probability values to WM voxels. Before employing the mixture of \( t \)-distributions, a rough co-registration to an MNI space template is performed (Schmidt 2017). Because the FLAIR sequence may contain lesions, this co-registration may affect tissue class labels. However, since only a rough co-registration estimate is needed, and corrections will automatically be made using the EM algorithm, it does not impact performance significantly.

We use GM, WM, and CSF spatial probability maps as priors to initialize the EM algorithm, and thereafter apply the EM algorithm until either the log likelihood convergences or after 100 iterations. We then apply a logistic regression model developed by (Schmidt 2017) to estimate the probability of a given voxel being WML. Crucially, we replace the MS lesion spatial prior used by the logistic regression with a homogenous spatial prior. This is critical, as the segmentation algorithm must remain impartial to disease type, whether MS or SVD. The resulting probabilistic lesion map is then binarized at 0.5, as this threshold was found to result in strong segmentation performance. Finally, we apply a post-processing algorithm to segment juxtacortical lesions.
3.1.4.2 T1 Black Hole & Lacune Segmentation

Although T1 BHs coexist with hyperintensities on FLAIR, we sought to develop a novel algorithm using only T1 for two reasons. The first is that, as we are interested in co-registering the FLAIR sequence to the T1, it is important that BHs on T1 be lesion filled prior to this, as the presence of lesions can skew the co-registration process. Second, we develop this method from a convenience standpoint, allowing BHs to be segmented using just one image modality.

An obstacle to the automatic segmentation of BHs is the difficulty of distinguishing them from background CSF – particularly the ventricles. SPM’s segmentation typically assigns BHs to CSF with very high probability. To illustrate this, Figure 5b shows the CSF probability map produced by SPM’s segmentation for the associated T1 image (Figure 5a). Observe that SPM assigns BHs to CSF. We would instead like BHs to be assigned as WM, so that they can thereafter be differentiated from NAWM.

![Figure 5](image)

*Figure 5* a) T1 of an RRMS impaired patient. Black holes (blue arrow) often appear adjacent to the ventricles (orange arrow) and other CSF (green arrow). b) SPM’s segmentation probability map for CSF. The associated black holes from a) have been assigned to CSF.

The MNI-space WM prior probability map used by SPM, shown in Figure 6, shows that the prior tissue probability map for WM in regions corresponding to where BHs are most commonly found, such as adjacent to the ventricles, is relatively low. This results in areas of
hypointensity on T1 being labeled as CSF. To address this, we transform the WM prior probability, for the purposes of BH lesion segmentation alone, to bias SPM’s segmentation algorithm towards labeling BHs as WM.

![Figure 6 SPM's white matter prior probability map. Prior probability of white matter near the ventricles (blue arrow) is not high enough to counteract the hypointensity of black holes on T1. Therefore, SPM assigns hypointense voxels in this area to CSF.](image)

We developed a novel spatial heuristic algorithm to solve this segmentation challenge. We use the heuristic that most BHs are centered around the ventricles, and diminish in size and likelihood as we move further from the center of a slice, to guide us in the development of the algorithm. Without loss of generality, we assume isotropic images in the following derivation.

Denote the center of a given axial slice as \((m, n)\), and the dimensions of this slice as \((a, b)\). This gives the relation

\[
(m, n) = \left(\frac{a}{2}, \frac{b}{2}\right)
\]

The distance from the center of the image to any coordinate \((i, j)\) is given by

\[
d_{i,j} = \sqrt{(i - m)^2 + (j - n)^2}
\]

Further,
\[ D = \sqrt{m^2 + n^2} \]

is the distance from the center of the image to the furthest point in the slice.

We now present the transformation applied to the WM prior probability map.

\[ p_{\text{WTM}}(i, j) = \frac{1}{1 + \exp\left\{ -\left( \frac{D}{2d_{i,j}} \right)^2 [p_{\text{WM}}(i, j) - 0.5] \right\}} \]

where \( p_{\text{WM}}(i, j) \) is the probability of WM at \((i, j)\) in the original WM prior probability map. This process is repeated for every axial \((k^{th})\) slice in the image. Following this transformation, the six prior probability maps (each corresponding to different brain tissue) must be re-normalized. Here, the softmax transformation is used

\[ p(i, j, k, r) = \frac{\exp\left( \frac{p_{\text{WTM}}(i, j, k, r) - 0.5}{\tau} \right)}{\sum_{s=1}^{6} \exp\left( \frac{p_{\text{WTM}}(i, j, k, s) - 0.5}{\tau} \right)} \]

where \( p_{\text{WTM}}(i, j, k, r) \) represents the probability of the transformed prior probability map at position \((i, j, k)\), for the \(r^{th}\) prior. \(\tau\), a hyperparameter, is set to \(\frac{1}{12}\).

The original and resulting probability maps for WM are shown in Figure 7 for comparison. For moderately high probabilities of WM near the center, the transformation increases the prior probability of WM significantly. For low probabilities near the center, it reduces the prior probability of WM significantly. The effect of this pattern decays sigmoidally as we move farther from the center of the slice. This improves the detection of BHs, since as seen before they are often located near the ventricles (which in MNI space corresponds to the center of the image).
Figure 7 a) Original white matter prior probability map. b) After the transformation.

(Note: for the remainder of this section, we will refer to the original tissue prior probability maps as $TPM_1$ and the transformed ones as $TPM_2$. This will be important for distinguishing how we derive our tissue probability maps.)

Despite the improvements obtained after applying this transformation, the WM prior probability values are so high in regions near the ventricles that actual CSF is labeled as WM. Therefore, we applied Gaussian smoothing to smooth the probability density. Figure 8 shows $TPM_2$ after applying different-sized smoothing kernels - we selected $\sigma = 1$ in our final model, as this threshold was found to result in strong segmentation performance.

Figure 8 a) Transformed white matter prior probability map with no smoothing b) Gaussian smoothing with $\sigma = 1$ c) Gaussian smoothing with $\sigma = 2$ d) Gaussian smoothing with $\sigma = 3$. 
BHs were subsequently identified using $TPM_2$ as the tissue prior probability maps. SPM’s default parameters were used, with the exception that two Gaussian distributions were used to model the intensity distribution of WM, as opposed to one, since the distribution for BHs varies highly from that of NAWM.

SPM’s resultant WM probability map was thresholded at 0.5. We then applied a bounding threshold on the T1, taken relative to the mean GM intensity, on the corresponding regions in the T1 image, so that very low intensity voxels corresponding to CSF and high intensity voxels corresponding to NAWM would be excluded. A sample segmentation is provided in Figure 9. Although this segmentation captures the BHs, it is a gross over-segmentation, due in part by the increased prior probability values for WM near CSF.

![Figure 9 a) T1 image of an RRMS impaired patient. b) Initial T1 black hole segmentation – clearly it is an over-segmentation, since it includes much of the CSF in its segmentation.](image)

We used the following approach to correct the over-segmentation:

1. Dilate SPM’s GM mask (obtained using regular prior probability maps $TPM_1$) by 1 mm and remove voxels in the over-segmentation that intersect with the resultant GM mask.
2. Dilate SPM’s CSF mask (derived using the transformed prior probability maps $TPM_2$) by 1 mm and remove voxels in the over-segmentation that intersect with the resultant GM mask.

3. Erode the remaining BH lesion mask by 1 mm.

Dilation of GM was motivated by the large number of false positives in CSF near the GM. Dilation of CSF was similarly motivated. Finally, eroding the BH mask removed spurious lesions.

The result of these dilation and erosion operations yields the BH segmentation mask shown in Figure 10. Now, we have an under-segmentation of BHs.

![Figure 10 a) T1 image of an RRMS impaired patient. b) T1 black hole segmentation after the erosions and dilations – we are now left with an under-segmentation.](image)

To consolidate the under and over-segmentations, we used a region growing algorithm:

1. Dilate the under-segmentation.
2. Multiply the result of step 1. by the over-segmentation mask.

which was repeated until convergence.
This algorithm effectively uses the lesional voxels in the under-segmentation as seeds, which thereafter grow towards the over-segmentation. Figure 11 illustrates the algorithm using a cartoon depiction of a hypothetical slice. Voxels labeled 1 signify those found in the over-segmentation alone, and voxels labeled 2 signify those found in both the under and over-segmentations. The algorithm labels any $2 - 1$ or $2 - 2$ connected regions as BHs, and any $1 - 1$ connected segments are discarded.

![Region growing algorithm](image)

*Figure 11 Region growing algorithm. a) 1 denotes a lesional voxel in the over-segmentation alone, and 2 denotes a lesional voxel in both the under and over-segmentations. b) Voxels labeled 3 are the result of the region growing algorithm.*

This region-growing procedure is equivalent to performing a simple depth-first search on an undirected graph using voxels with value 2 as sources. Specifically, we perform the depth-first search procedure presented in Algorithm 1, treating each voxel with value 2 as a source, and recursively searching for any 8-connected neighbors in the axial plane with value 1 or 2 to grow into.

We use pseudocode to convey the essential function of the algorithm; the actual implementation depends on the programming language of choice. We assume the undirected graph $G$ represents the image, with each voxel $q$ in $G$ represented as a vertex with property *value*. The subroutine $\text{neighborhood}(G, q)$ returns the set of vertices that are neighbors of vertex $q$ in graph $G$.

*Algorithm 1. Depth-first graph search implementation of the region growing algorithm.*
1. \texttt{GRAPH\_SEARCH}(G)

2. \texttt{for } q \in G \\
3. \texttt{if } q.\texttt{value} == 2 \\
4. \texttt{q.\texttt{value} = 3} \\
5. \texttt{GRAPH\_GROW}(G, q)

1. \texttt{GRAPH\_GROW}(G, q)

2. \texttt{Q = neighborhood}(G, q) \\
3. \texttt{for } q' \in Q \\
4. \texttt{if } q'.\texttt{value} == 1 \texttt{or } q'.\texttt{value} == 2 \\
5. \texttt{q'.\texttt{value} = 3} \\
6. \texttt{GRAPH\_GROW}(G, q')

The resulting image \( G \) after \texttt{GRAPH\_SEARCH}(G) finishes executing is equivalent to the resulting image after the erosion-dilation procedure described above. The algorithm runs in time \( O(|V| + |E|) \), \( V \) is the set of vertices; the voxels that are \( 2 - 1 \) or \( 2 - 2 \) connected, \( E \) is the set of edges in the \( 2 - 1 \) and \( 2 - 2 \) connected segments, and \( | \cdot | \) denotes the cardinality of the set.

Finally, we applied a 2D morphological filling operation on a slice-by-slice basis to obtain smooth BH segmentations. The result of the application of this algorithm on our previous under and over-segmentations is shown in Figure 12. Clearly, it is a much better approximation of the actual BH lesion distribution.
Figure 12 a) T1 image of an RRMS impaired patient. b) over-segmentation c) under-segmentation d) result of region growing algorithm from under-segmentation to over-segmentation.

3.1.4.3 Minimizing False Positives

The corpus callosum, the largest WM structure in the brain, was often labeled as lesional on FLAIR by our automatic segmentation algorithm. Figure 13 shows a sample case. We sought to isolate for these erroneous lesions, and re-assign them to healthy tissue.

Figure 13 a) FLAIR of an RRMS impaired patient. The blue arrow points at the corpus callosum. b) Erroneous (blue arrow) white matter lesion detected.

The following algorithm was used to remove these lesions, on an axial slice-by-slice basis:
1. Threshold the CSF probability map produced by SPM (using \(TPM_1\)) at 0.5.
2. Apply a Gaussian blur with \(\sigma = 1\).
3. Apply the horizontal and vertical Prewitt edge detection filters, and calculate their Euclidean sum.
4. Find the convex hull of the resulting area, and re-assign tissue labeled as WML to NAWM in this area.

The first three steps can be taken together to isolate for the ventricles, which surround the corpus callosum. By thresholding and applying a Gaussian blur to Figure 5b, the background CSF in the brain is made spurious, while the integrity of the ventricles are maintained. The Prewitt filter identifies the edges of the ventricles. The result of these three steps is shown in Figure 14b. Thereafter, the convex hull of the image is taken, giving Figure 14c. This process isolates for the ventricles and as a result the corpus callosum, and enables the removal of erroneous lesions without encroaching on correctly labeled WMLs. Importantly, we do not remove any BHs using this method, as they can easily be confounded with the ventricles.

![Image of brain MRI with annotations](image)

*Figure 14 a) The original CSF segmentation produced by SPM b) Result of thresholding, applying the Gaussian filter, and the Prewitt filter c) Result of applying the convex hull to the image in b).*
3.1.5 Co-registration

3.1.5.1 Lesion Filling

We used the following procedure for lesion filling on axial FLAIR and T1. For each axial slice, each voxel corresponding to a lesion had its intensity value re-assigned to a random sample from a Gaussian distribution with mean equal to that of the healthy WM, and standard deviation one half of the healthy WM, as in (Valverde, Oliver, and Lladó 2014). Subsequently, we applied a Gaussian filter with $\sigma = 1$. This resulted in a normal appearance in the lesion-filled WM, and thus minimized error during the ensuing co-registration.

Figure 15 and Figure 16 show the result of lesion filling on axial FLAIR and T1, respectively. Our filling algorithm performs just as well as published methods, though this conclusion is admittedly qualitative in nature.

*Figure 15 Left: FLAIR of an RRMS impaired patient. Center: corresponding white matter lesion segmentation. Right: filled FLAIR image.*
3.1.5.2 Applying the Co-registration

SPM was used for co-registration. Measures of performance include how closely anatomical regions overlap between the template image and the co-registered one, whether voxel intensity values carry the same distribution across the brain, and the overall appearance of the transformed image. Figure 17 shows a representative case for the co-registration procedure.

Figure 16 Left: T1 of an RRMS impaired patient. Center: corresponding T1 black hole lesion segmentation. Right: filled T1 image.

Figure 17 a) Reference image: MNI-152 T1. b) RRMS patient’s T1 scan in MNI space c) FLAIR in MNI space after being co-registered to the filled T1 scan. d) Lesion map in MNI space.
3.1.6 Segmentation Performance Metrics

Total lesion volume (TLV) was calculated for the ground truth tracings and the automatic segmentations. The intraclass correlation coefficient (ICC) between the two raters was calculated using SPSS software. A two-way random effects with absolute agreement and single measures design was used, which is the appropriate choice as advocated by (E. Shrout and L. Fleiss 1979) and (Koo and Y Li 2016). Bland-Altman plots were used to assess whether there was a systematic bias for TLV across the two raters.

We used the publicly available ISBI dataset for MS lesion segmentation (Carass et al. 2017) as an independent dataset to evaluate our segmentation performance. We use the raw images, and not the pre-processed ones. A two-way random effects with absolute agreement and single measures design was used, however here we have three raters, as the ISBI data set has two ground truth tracings for each case, which were done by two independent raters. We do not include other measures for assessing segmentation performance, as we are interested in lesion detection, rather than accurate segmentation per se (Kamnitsas et al. 2017).

3.1.7 Combined Lesion Probability Maps

A neuroimaging combined probability map (CPM), as we define it here, represents the probability distribution of lesions in the brain for a disease. Such a CPM is derived from a training dataset. Thereafter, given a new image, its image similarity (which we define shortly) to the CPM determines how likely it is for that image to belong to a particular disease.
For both MS and SVD, the dataset was randomly split into training and test sets using an 80/20% split by disease. The test data was not used to assess model performance until the very end of the study, and was used exactly once.

We thereafter split the training data into derivation and validation sets using an 80/20% split – the derivation set is used to develop the CPMs, and the validation set to test performance and for hyperparameter tuning. Several approaches were considered for data partitioning and cross validation between the derivation and validation sets, two of which we mention here, and thereafter outline further in section 3.1.7.1. The first approach uses random partitioning to produce derivation and validation sets. This can be done via Monte Carlo cross validation or k-fold cross validation. Although this ensures randomization, a considerable number of permutations are necessary to find a stable estimate of model performance. However, each simulation is very computationally expensive, so that only a relatively small number of simulations can be reasonably performed. To resolve this challenge - that of obtaining stable performance estimates with only a small number of simulations, we only performed simulations for random splits which produced a similar fractional TLV distribution across the derivation and validation sets, where fractional TLV is defined as $\frac{TLV}{TLV+NAWM}$.

After partitioning the data, we use two different approaches for constructing CPMs from the derivation set, which we later compare in terms of classification performance. The first approach calculates a CPM for the combined WML and BH mask, so that no distinction is made between these lesion types. The second approach calculates two CPMs for each disease – one corresponding to WMLs, the other to BHs. This allows us to determine whether differentiating the lesion types leads to improved classifier performance. Regardless of the approach, the CPMs are constructed by summing the lesion masks, then normalizing by the number of patients, giving a spatial probability distribution. Figure 18 shows the same slice for CPMs derived for RRMS and SVD.
3.1.7.1 Partitioning the Data and Cross Validation

K-fold cross validation and Monte Carlo cross validation are two approaches to cross validation. In k-fold cross validation, we partition the data into $k$ segments, and develop the CPM using $k - 1$ groups, leaving one set out. This is repeated $k$ times, once for each unique set left out. K-fold cross validation is known to be a nearly unbiased estimator, however it has high variance. Further, although k-fold cross validation ensures randomization, it may produce unrepresentative populations with respect to fractional TLV distribution for the derivation and validation sets. In Monte Carlo cross validation, derivation and validation sets are assigned by randomly partitioning the data. This process is repeated until a sufficient number of simulations have been performed to satisfy statistical conditions, which may be very high to obtain stable estimates of performance. However, we can choose to only investigate the Monte Carlo simulations which produce splits with representative fractional TLVs between the derivation and validation sets. We describe this procedure next.

One way to ensure representative groups with respect to fractional TLV distribution is to use the DUPLEX algorithm (Snee 1977), which would compute the similarity of lesion distribution between images in feature space. However, due to the immense feature space of 3D medical

---

Figure 18 (Left) T1 MNI template space image. Combined probability map from one Monte Carlo simulation for RRMS (center) and SVD (right).
images (corresponding to the number of voxels), and hence the large computational load (Reitermanov 2010), we excluded this algorithm from further consideration. Another option is constrained optimization - ensuring that the mean fractional TLV across derivation and validation sets is approximately equal. However, this is insufficient, as we would like to ensure that the fractional TLV distribution between derivation and validation sets is similar. Consequently, we employ the non-parametric Kolmogorov-Smirnov test to ensure similar fractional TLV distributions between the derivation and validation sets for a particular disease. We require that \( p > 0.975 \) (two-sided test with \( \alpha = 0.05 \)), and ensure that the mean fractional TLV difference between the derivation and validation sets is less than 0.01.

### 3.1.7.2 Monte Carlo Cross Validation

We use a Monte Carlo simulation approach that guarantees each sample is included in the validation set on a rotating basis – that is, we constrain the algorithm so that on a rotating basis, one sample is preserved for the validation set. We first partition the training data into derivation and validation sets (as described in section 3.1.7.1). We then construct the CPMs from the derivation set, and subsequently use image similarity measures (discussed in section 3.1.7.3) to construct models, which we evaluate on the validation set. We use this Monte Carlo cross validation approach to obtain estimates of performance based on the hyperparameters of interest. We then construct our final model with the optimal hyperparameter settings on the entire training set, and evaluate this model on the test set.

### 3.1.7.3 Neuroimaging Similarity Score

We investigate two image similarity measures: the \( L^p \)-norm distance metric and our novel cross entropy image distance.

#### 3.1.7.3.1 Euclidean Distance

The Euclidean distance is defined as:
\[
\begin{align*}
    d &= \sqrt{\sum_{i,j,k} (\hat{p}_{i,j,k} - \hat{q}_{i,j,k})^2} \\
\end{align*}
\]

where \( \hat{p}_{i,j,k} \) denotes whether the voxel in the segmentation at coordinates \((i, j, k)\) is lesional (1) or not (0), and \( \hat{q} \) is the probability of class lesional in the CPM. Alternatively, as a generalization of the Euclidean distance, the \( L^p \)-norm can be used

\[
\begin{align*}
    d &= \sqrt[p]{\sum_{i,j,k} |\hat{p}_{i,j,k} - \hat{q}_{i,j,k}|^p} \\
\end{align*}
\]

where \( p = 2 \) reduces the \( L^p \)-norm to the Euclidean distance. Here, \( p \) is a hyperparameter, and in principle can be set by identifying, via grid search, the value which maximizes validation set performance.

Regardless of the value of \( p \), the \( L^p \)-norm is tabulated for each of the two CPMs. Subsequently, the score \( s \), which is bounded between \([0, 1]\), for each disease amongst the \( K \) possible diseases is

\[
\begin{align*}
    s(C = i) &= \frac{\sum_{j=1, j\neq i}^K d_j}{\sum_{k=1}^K d_k} \\
\end{align*}
\]

Importantly, by the properties of \( L^p \) spaces, the \( L^p \)-norm decreases monotonically as \( p \) increases. This implies that if the \( L^p \)-norm, with \( p \geq 1 \), cannot effectively be used to distinguish the two diseases, then any value greater than \( p \) will not be able to distinguish them either. Therefore, we restrict our analysis of the \( L^p \)-norm to \( p = 1 \), but also analyze \( p = 2 \) as a proof of concept of this principle.

3.1.7.3.2 Cross Entropy Image Similarity

The scoring metric here is the cross entropy loss function:
\[
\text{loss} = \frac{1}{\sum_{i,j,k} I(\hat{p}_{i,j,k} = 1)} \sum_{i,j,k} \min -\hat{p}_{i,j,k} \left( \log \hat{q}_{i,j,k} \right), \epsilon
\]

where, as before \(\hat{p}_{i,j,k}\) denotes whether the voxel in the segmentation at coordinates \((i,j,k)\) is lesional (1) or not (0), \(\hat{q}\) is the probability of class lesional in the CPM, \(I[\cdot]\) is the indicator function, and \(\log[\cdot]\) denotes the binary logarithm. \(\epsilon\) is a hyperparameter that bounds the log loss for a given voxel. This approach to bounding the log loss is equivalent in practice to the minmax rule for avoiding undefined values when computing the log loss. The optimal \(\epsilon\) for each disease is obtained by identifying, via grid search, the value which maximizes validation set performance. The hyperparameters used in the grid search are \(\{2^1, 2^2, \ldots, 2^{15}\}\). We use grid search, as opposed to randomly sampling from the set of possible hyperparameters, because the number of possible hyperparameters is not very large.

The score for each disease amongst the \(K\) possible diagnoses is

\[
s(C = i) = \frac{\sum_{j=1,j\neq i}^{K} \text{loss}_j}{\sum_{k=1}^{K} \text{loss}_k}
\]

### 3.1.7.3.3 Accounting for Spatial Differences

Co-registration inaccuracies, as well as anatomical differences between patients, can cause lesions located in similar brain regions for different patients not to map exactly onto each other, which can result in low image similarity even when lesion distribution is near-identical.

We considered two approaches to remedy this problem. The first was to apply smoothing operations to the CPMs (such as mean, median or Gaussian smoothing), and the second was to explicitly account for spatial differences in the distance metrics. We decided to pursue the latter approach, since there were fewer hyperparameters to consider, and filtering the CPM seemed intuitively a faulty route, since we would be discarding information from the CPM without significant gains to compensate for the information loss.
Concretely, we re-define our distance metrics as

\[ d = \sqrt[p]{\sum_{i,j,k} \min_{\gamma_1, \gamma_2, \gamma_3} |\hat{p}_{i,j,k} - \hat{q}_{i+\gamma_1,j+\gamma_2,k+\gamma_3}|^p}, \gamma_1, \gamma_2, \gamma_3 \in [-\gamma, \gamma] \]

and

\[ loss = \frac{1}{\sum_{i,j,k} I(\hat{p}_{i,j,k} = 1)} \sum_{i,j,k} \min_{\gamma_1, \gamma_2, \gamma_3} [\min \hat{p}_{i,j,k}(\log \hat{q}_{i+\gamma_1,j+\gamma_2,k+\gamma_3}, \epsilon)], \gamma_1, \gamma_2, \gamma_3 \in [-\gamma, \gamma] \]

That is, for any given voxel in the new segmentation, we select the minimum distance/loss calculated within a \( \gamma \) mm neighborhood around that voxel with respect to the CPM, so that spatial differences of a few millimeters between segmentations and the CPMs is not problematic. We set \( \gamma \) to 2 mm, since this accounted for differences due to co-registration inaccuracies and anatomical differences between patients.

### 3.2 Disease Classification

The first and second approaches to classification use neuroimaging alone, the third approach uses clinical features alone, and the fourth combines neuroimaging and clinical features. We investigate two approaches using neuroimaging to determine whether distinguishing between WMLs and BHs adds predictive value in disease discrimination. We will contrast the performance of the four approaches and determine whether combining neuroimaging and clinical features adds predictive value. In all four models, MS corresponds to a positive case, and SVD to a negative case, and we consider the classification of RRMS vs. SVD. For comparison, a model is developed to distinguish RRMS & SPMS vs. SVD using clinical features alone. Figure 19 provides a schematic of the four classification approaches.
We use the Naïve Bayes family of classifiers, as opposed to the logistic regression model, when multiple covariates are used for prediction. In addition to the advantages associated with the Naïve Bayes family of classifiers, which are listed in section 2.6, we found that the logistic regression model was prone to giving unstable parameter estimates depending on the distribution of features in the derivation set. Specifically, some of the Monte Carlo simulations produced splits with near-perfect separation or quasi-separation between classes in the derivation set. Logistic regression is notorious for its poor performance in such problems. We attempted to alleviate the separability problem, at the expense of model interpretability, by
using L1 (lasso) and L2 (ridge) regularization, as well as more robust variants of logistic regression such as Firth logistic regression (Firth 1993) (Heinze and Schemper 2002). However, these did not fully resolve the instability in model parameter estimates, which had either astronomical odds ratios or standard errors.

There are theoretical advantages to using Naïve Bayes classifiers as well. For example, the logistic regression model requires \( O(d) \) training cases, where \( d \) is the dimensionality of the feature space, to achieve its asymptotic classification performance (Ng and Jordan 2002). As the number of features in our model grew, for example by combining neuroimaging and clinical features, this asymptotic bound could have become more problematic. In contrast, (Ng and Jordan 2002) show that the Naïve Bayes family of classifiers require only \( O(\log d) \) samples to achieve their asymptotic performance. Therefore, although the logistic regression model generally obtains better asymptotic performance than the Naïve Bayes family of classifiers, Naïve Bayes classifiers reach their asymptotic performance more quickly, and thus can outperform the logistic regression model for small datasets.

To address class imbalance, all Naïve Bayes implementations have their class prior probabilities fixed at 0.5 for both MS and SVD. Since setting the class prior probabilities to 0.5 results in obtaining the maximum likelihood estimate of class predictions, we set the prediction probability threshold for MS as the proportion of MS cases in the derivation set. Furthermore, as we use the \( \phi \) coefficient, which is defined in the next subsection, to evaluate model performance, the final model performance is unaffected by class imbalances. In the Discussion section, we will investigate how incorporating the actual prior probability of each disease, as well as the relative cost of misdiagnosis, would change our model formulation in practice.

Due to the small dataset size, the maximum likelihood estimates for the Naïve Bayes model parameters were sometimes overly-confident, leading to sub-optimal performance on the validation set. As a result, we considered using maximum a posteriori estimation with
conjugate priors on the model parameters. Conjugate priors are particularly useful, as the posterior distribution is available in closed form. However, this presents the need to optimize the hyperparameters associated with the conjugate priors. Alternatively, these hyperparameters can be set in advance, but this could result in significant model bias for small datasets. To resolve this, we used Jeffreys prior (Jeffreys 1946), a non-informative prior, to obtain the parameter estimates. Further details on Jeffreys prior and a derivation of the resulting model are provided in Appendix A.

Finally, we justified our use of 100 Monte Carlo simulations by empirically assessing the convergence of model performance summary statistics as a function of the number of simulations.

### 3.2.1 Classification using Neuroimaging Alone

Recall the two approaches for constructing CPMs for each disease. The first creates a single CPM for the combined WML and BH lesion masks, and the second creates separate CPMs for the WML and BH lesion masks. We outline classification approaches for each. For each method, a model is developed to differentiate RRMS from SVD.

Our evaluation metric is Matthews correlation coefficient (MCC) (B. W. Matthews 1975), also known as the phi ($\phi$) coefficient. It is defined as

$$\phi = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where $TP$ is the number of true positives, $TN$ is the number of true negatives, $FP$ is the number of false positives, and $FN$ is the number of false negatives. We use this measure under the premise that the cost of a false positive is equivalent to that of a false negative – in this case, $\phi$ is regarded as one of the best summarizing statistics (Powers 2011), and it is robust to imbalanced class sizes in the derivation set (Boughorbel, Jarray, and El-Anbari 2017). Its value
lies within the range [-1, +1], with -1 representing complete disagreement in classification, 0 representing no better than random guessing, and +1 perfect agreement. If any of the four terms in the denominator equals zero, then $\phi$ is set to zero. In the Discussion section, we will investigate how incorporating the relative cost of false positives and false negatives would change our evaluation metric in practice.

Although we use $\phi$ as the evaluation metric for our models, we are interested in evaluating model performance in further detail. Specifically, we use precision/positive predictive value (PPV), sensitivity/recall, specificity, and negative predictive value (NPV) as other indicators of model performance. They are defined as follows:

\[
\text{precision/PPV} = \frac{TP}{TP + FP}
\]
\[
\text{sensitivity/recall} = \frac{TP}{TP + FN}
\]
\[
\text{specificity} = \frac{TN}{TN + FP}
\]
\[
NPV = \frac{TN}{TN + FN}
\]

3.2.1.1 Classification using a Unified Combined Probability Map

We determine the optimal threshold for the neuroimaging score by maximizing $\phi$. For a given Monte Carlo simulation and for each disease, we construct a CPM from the derivation set, and find the optimal threshold for a variety of hyperparameter settings. These optimal thresholds are then used for prediction on the validation set. This procedure is repeated for all Monte Carlo simulations, and the resulting performances are averaged to obtain a final performance measure. The hyperparameters giving the best performance with respect to $\phi$ on the validation set, across all Monte Carlo simulations, are the optimal hyperparameters. We
also measure the cross entropy loss between actual class labels and neuroimaging scores on the derivation set

\[
\zeta = - \sum_{n=1}^{N} p^{(n)} \log(q^{(n)}) - \sum_{n=1}^{N} (1 - p^{(n)}) \log(1 - q^{(n)})
\]

where \(p^{(n)}\) is the actual class label for the \(n^{th}\) case, and \(q^{(n)}\) is the neuroimaging score.

(Note: for clarity, the cross entropy loss measured here serves a different purpose than the cross entropy image similarity metric.)

3.2.1.2 Classification using Dual Combined Probability Maps

We make use of supervised learning algorithms which map the features corresponding to the WML and BH scores to class labels. Specifically, we use the Gaussian Naïve Bayes classifier, without and with the non-informative prior for comparison. The general approach is not unlike the one described in section 3.2.1.1. For a given Monte Carlo simulation and for each disease, we construct two CPMs from the derivation set. For each possible combination of hyperparameters, we develop Gaussian Naïve Bayes classifiers which are used for prediction on the validation set. The hyperparameters giving the best performance with respect to \(\phi\) on the validation set, across all Monte Carlo simulations, are the optimal hyperparameters.
3.2.2 Classification using Clinical Features Alone

We use two clinical features - patient age at the time of presentation, and sex – both of which are factors known to distinguish MS and SVD (see section 2). For a given Monte Carlo simulation, a Gaussian Naïve Bayes classifier is developed using patient age, and a Bernoulli Naïve Bayes classifier using patient sex. We then combine the probability estimates of the two models as described below to make a classification. Importantly, we use the same derivation and validation splits used when developing the models with neuroimaging features alone, in order to control for the possibility of varying performance depending on the derivation and validation splits. However, when we develop the model to compare RRMS & SPMS vs. SVD, we randomly split the training set into derivation and validation sets, as no counterpart model was developed to distinguish these diseases using neuroimaging alone.

We combine the probability estimates of the Gaussian Naïve Bayes and Bernoulli Naïve Bayes models as follows. For a given Monte Carlo simulation, denote the $n^{th}$ derivation case consisting of age alone as $t^{(n)}$, and the $n^{th}$ derivation case consisting of sex alone as $u^{(n)}$. The posterior probability for prediction of the $k^{th}$ class is given by

$$ p(C^{(n)} = k | t^{(n)}, u^{(n)}) = \frac{p(C^{(n)} = k)p(t^{(n)}|C^{(n)} = k)p(u^{(n)}|C^{(n)} = k)}{\sum_{l=1}^{K} p(C^{(n)} = l)p(t^{(n)}|C^{(n)} = l)p(u^{(n)}|C^{(n)} = l)} $$

Since the prior probability for each disease is set to 0.5, this reduces to

$$ p(C^{(n)} = k | t^{(n)}, u^{(n)}) = \frac{p(t^{(n)}|C^{(n)} = k)p(u^{(n)}|C^{(n)} = k)}{\sum_{l=1}^{K} p(t^{(n)}|C^{(n)} = l)p(u^{(n)}|C^{(n)} = l)} $$
3.2.3 Classification using Neuroimaging and Clinical Features

We combine the clinical features with the neuroimaging scores, which are derived using the dual CPMs, to obtain a combined neuroimaging and clinical feature set. This gives each patient four features, and thus for each Monte Carlo simulation four Naïve Bayes models are developed. After developing these four models, we combine their probability estimates to obtain the final posterior probability estimate. For a given Monte Carlo simulation, denote the age, sex, WML score, and BH score for the $n^{th}$ derivation case as $t^{(n)}$, $u^{(n)}$, $s_1^{(n)}$, $s_2^{(n)}$, respectively. The posterior probability for prediction of the $k^{th}$ class is given by

$$p(C^{(n)} = k | t^{(n)}, u^{(n)}, s_1^{(n)}, s_2^{(n)})$$

$$= \frac{p(C^{(n)} = k) p(t^{(n)} | C^{(n)} = k) p(u^{(n)} | C^{(n)} = k) p(s_1^{(n)} | C^{(n)} = k) p(s_2^{(n)} | C^{(n)} = k)}{\sum_{l=1}^{K} p(C^{(n)} = l) p(t^{(n)} | C^{(n)} = l) p(u^{(n)} | C^{(n)} = l) p(s_1^{(n)} | C^{(n)} = l) p(s_2^{(n)} | C^{(n)} = l)}$$

Since the prior probability for each disease is set to 0.5, this reduces to

$$p(C^{(n)} = k | t^{(n)}, u^{(n)}, s_1^{(n)}, s_2^{(n)})$$

$$= \frac{p(t^{(n)} | C^{(n)} = k) p(u^{(n)} | C^{(n)} = k) p(s_1^{(n)} | C^{(n)} = k) p(s_2^{(n)} | C^{(n)} = k)}{\sum_{l=1}^{K} p(t^{(n)} | C^{(n)} = l) p(u^{(n)} | C^{(n)} = l) p(s_1^{(n)} | C^{(n)} = l) p(s_2^{(n)} | C^{(n)} = l)}$$

3.2.4 Interpreting the Classification Decision

We explore two approaches to improve model interpretability. The first provides the relative importance of each feature in making the classification decision, and the second visualizes the brain regions which contributed most significantly to the classification decision.
3.2.4.1 Relative Importance of Features

As seen in section 3.2, each Naïve Bayes classifier outputs a probability of the data belonging to class $k$ as $p(C^{(n)} = k|f_i^{(n)})$, where $f_i$ corresponds to the $i^{th}$ feature, of which there are $J$ in total. For the model combining neuroimaging and clinical features, $J = 4$. We next evaluate the relative contribution of each feature, as a percentage, in making the classification decision.

Recall that the log-odds ratio, or logit of class probability predictions for a binary classification problem, is given by

$$\log \left( \frac{p(C^{(n)} = 1|f_i^{(n)})}{p(C^{(n)} = 0|f_i^{(n)})} \right)$$

which has an analogous role to the log-odds ratio in logistic regression. The relative weight of each feature is evaluated by taking the softmax of the logits

$$W_i = \frac{\exp \left( \log \left( \frac{p(C^{(n)} = 1|f_i^{(n)})}{p(C^{(n)} = 0|f_i^{(n)})} \right) \right)}{\sum_{j=1}^{J} \exp \left( \log \left( \frac{p(C^{(n)} = 1|f_j^{(n)})}{p(C^{(n)} = 0|f_j^{(n)})} \right) \right)}$$

which reduces to

$$W_i = \frac{p(C^{(n)} = 1|f_i^{(n)})}{\sum_{j=1}^{J} p(C^{(n)} = 1|f_j^{(n)})} \frac{p(C^{(n)} = 0|f_i^{(n)})}{p(C^{(n)} = 0|f_j^{(n)})}$$

For each prediction, we evaluate the relative importance of each feature using this equation.
3.2.4.2 Visualizing Significant Brain Regions

Recall each patient has their neuroimaging score for each disease and CPM computed as

$$loss = \frac{1}{\sum_{l,j,k} l(\hat{p}_{l,j,k} = 1)} \sum_{l,j,k} \min_{y_1,y_2,y_3} \left[ \min_{\gamma_1,\gamma_2,\gamma_3} \left\{ \hat{p}_{l,j,k} \left( \log \hat{q}_{l+y_1,j+y_2,k+y_3}, \epsilon \right) \right\}, \gamma_1,\gamma_2,\gamma_3 \in [-\gamma,\gamma] \right]$$

If we observe the voxel-wise loss instead, we obtain a new image $I$ with the following values for each voxel

$$I = \min_{y_1,y_2,y_3} \left[ \min_{\gamma_1,\gamma_2,\gamma_3} \left\{ \hat{p}_{l,j,k} \left( \log \hat{q}_{l+y_1,j+y_2,k+y_3}, \epsilon \right) \right\}, \gamma_1,\gamma_2,\gamma_3 \in [-\gamma,\gamma] \right]$$

Assuming separate CPMs are used for WMLs and BHs, and we allow for $\epsilon$ to take different values for each disease, the four resulting images for a given patient are $I_{\epsilon_{MS,WML}}, I_{\epsilon_{SVD,WML}}, I_{\epsilon_{MS,BH}}, I_{\epsilon_{SVD,BH}}$. For each lesion type, the log ratio $r$ of the images can be used to visualize which brain regions contributed most significantly to the classification decision

$$r_{WML} = \log \left( \frac{I_{\epsilon_{SVD,WML}}}{I_{\epsilon_{MS,WML}}} \right) = \log(I_{\epsilon_{SVD,WML}}) - \log(I_{\epsilon_{MS,WML}})$$

$$r_{BH} = \log \left( \frac{I_{\epsilon_{SVD,BH}}}{I_{\epsilon_{MS,BH}}} \right) = \log(I_{\epsilon_{SVD,BH}}) - \log(I_{\epsilon_{MS,BH}}).$$
4 Results

We use $\alpha = 0.05$ to assess statistical significance. Statistics which are averaged across Monte Carlo simulations are presented with standard errors. These are computed as $\frac{\sigma}{\sqrt{N}}$, where $\sigma$ is the standard deviation of the statistic across the $N$ Monte Carlo simulations. All other statistics are presented with standard deviations. We use a two-tailed $t$-test when comparing the classification performance of two models across the 100 Monte Carlo simulations.

4.1 Automatic Lesion Segmentation Performance

Table 3 presents the segmentation performance on our RRMS dataset, as well as the raw images from the publicly available ISBI MS dataset, which was used to validate model performance independently. All results are presented using two-way random effects with absolute agreement and single measures. Note that SPM failed to segment brain matter in four cases from the publicly available ISBI MS dataset. As this preliminary segmentation is necessary in our methodological pipeline, these cases were excluded from further evaluation.

<table>
<thead>
<tr>
<th></th>
<th>RRMS Dataset</th>
<th>Intraclass Correlation Coefficient</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Ground Truth Raters</td>
<td>1</td>
<td>0.966</td>
<td>[0.935, 0.982]</td>
</tr>
<tr>
<td></td>
<td>ISBI Dataset</td>
<td>2</td>
<td>0.950</td>
</tr>
</tbody>
</table>

The one sample $t$-test for the difference of means between the two raters on the RRMS dataset resulted in $p > 0.05$, which signifies no systematic bias in TLV estimation. Figure 20 shows the Bland-Altman plot for TLV on our RRMS dataset. The middle horizontal line shows the average lesion volume difference between the lesion masks from the automatic segmentation
algorithms and the ground truth tracings. This line corresponds nearly to zero, indicating little difference between the two raters. The upper and lower horizontal lines represent two standard deviations in lesion volume difference above and below the mean difference. There are two outliers, both corresponding to an under-estimation of TLV in the automatic segmentation. However, it should be noted that the ground truth tracings had imperfections themselves, in particular they sometimes over-zealously traced lesions, and thus large differences in TLV between the two raters should not necessarily be taken as a sign of poor algorithm performance.

There is a slight negative trend in the difference in TLV between the automatic segmentations and ground truth tracings. This is in fact typical of computer systems, and will be investigated further in the discussion section.
4.2 Classification Performance

We contrast the performance of four approaches – two of which use neuroimaging alone, one which uses clinical features alone, and the final approach which uses a combination of neuroimaging and clinical features. We analyze the performance of the classifiers in
distinguishing RRMS vs. SVD, as well as RRMS & SPMS vs. SVD for clinical features alone. Finally, we evaluate classifier performance using $\phi$, and use several other metrics as alternative indicators of performance. Throughout, a positive case corresponds to MS, and a negative case to SVD.

### 4.2.1 Methods 1 & 2 - Using Neuroimaging Alone

We exclude the results for the $L^p$-norm distance metric, as it led to poor classification performance. As such, results are only provided for our novel cross entropy image similarity metric. All results are for $\gamma = 2$ mm (i.e. a spatial tolerance of 2 mm in the axial, sagittal and coronal planes), as we found that spatial tolerance significantly improved performance.

#### 4.2.1.1 Method 1 - Unified Combined Probability Map

Table 4 shows the mean performance of the cross entropy image similarity metric for the best performing hyperparameter(s), when restricting $\epsilon$ to be the same for both diseases, and when allowing it to vary for the two diseases. Also listed are the average cross entropy losses between neuroimaging scores and labels for the derivation set. These results were obtained by averaging 100 Monte Carlo simulations and are presented with standard errors. Model performance improves when the optimal value of $\epsilon$ can vary between the two diseases ($p < 0.05$).

Figure 21 shows a receiver operating characteristic (ROC) curve and a precision recall curve (PRC), both constructed from the derivation set, for the cross entropy image similarity metric, with $\epsilon = 2^2$ when restricting the value of $\epsilon$ to be the same for both diseases. The associated histograms of disease scores, for the derivation and validation sets, and cross entropy loss for the derivation set are also provided in Figure 21. Figure 22 is for $\epsilon = 2^8$, and Figure 23 is for $\epsilon = 2^1$. 

62
These figures demonstrate the role of the hyperparameter $\epsilon$ in governing model generalization and in constraining model power. High values of $\epsilon$ lead to overfitting to the derivation set, since small differences in lesion distribution in the derivation set can have a very high effect on the image similarity score. Low values of $\epsilon$ lead to model underfitting on the derivation set, since differences in lesion load distributions are restricted. This explains why the model with a moderate value of $\epsilon$ performs better – it suffers less from the extremes of model overfitting and underfitting.

Furthermore, this notion of generalization has a relation with the cross entropy loss between neuroimaging scores and labels on the derivation set. A classifier which obtains a low cross entropy loss on the derivation set is akin to overfitting, since as in Figure 22, there is a range of scores where the classifier cannot effectively make predictions. A high cross entropy loss on the derivation set is akin to underfitting, since the model fails to separate the two diseases even on cases it has been trained on, as in Figure 23. The values of the cross entropy losses vary significantly with $\epsilon$, and moderate values of $\epsilon$ lead to better performance.

Figure 24 shows the model when allowing the optimal value of $\epsilon$ to vary between the two diseases. It achieves the best performance among the models analyzed thus far ($p < 0.05$). As there is a large difference in TLV between the two diseases, a different value of $\epsilon$ may allow for the intricacies of lesion distribution in each disease to be better captured when calculating the image similarity score.
Table 4 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes. Mean performance of the cross entropy image similarity metric for the best performing hyperparameters is provided. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors.

<table>
<thead>
<tr>
<th>Cross Entropy Image Similarity Metric</th>
<th>$\phi$</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>NPV</th>
<th>Cross Entropy Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common $\epsilon$ for MS and SVD $\epsilon = 2^2$</td>
<td>Derivation</td>
<td>0.865 ± 0.004</td>
<td>0.884 ± 0.005</td>
<td>0.946 ± 0.003</td>
<td>0.930 ± 0.004</td>
<td>0.970 ± 0.001</td>
</tr>
<tr>
<td>Validation</td>
<td>0.509 ± 0.019</td>
<td>0.740 ± 0.017</td>
<td>0.631 ± 0.019</td>
<td>0.849 ± 0.012</td>
<td>0.806 ± 0.008</td>
<td></td>
</tr>
<tr>
<td>Differing $\epsilon$ for MS and SVD $\epsilon_{MS} = 2^4$, $\epsilon_{SVD} = 2^5$</td>
<td>Derivation</td>
<td>0.973 ± 0.001</td>
<td>0.992 ± 0.002</td>
<td>0.973 ± 0.002</td>
<td>0.996 ± 0.001</td>
<td>0.986 ± 0.001</td>
</tr>
<tr>
<td>Validation</td>
<td>0.618 ± 0.020</td>
<td>0.774 ± 0.017</td>
<td>0.739 ± 0.021</td>
<td>0.863 ± 0.011</td>
<td>0.862 ± 0.009</td>
<td></td>
</tr>
</tbody>
</table>

Figure 21 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Receiver operating
characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon = 2^8$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. There is good separability between the two classes on the derivation set, but without significant overfitting, as exemplified by the ROC and PRC, which leads to good performance on the validation set.

Figure 22 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon = 2^8$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. Although the separability between classes is very good on the derivation set, the model is overfitting, as exemplified by the ROC and PRC, and thus the classifier performs poorly on the validation set.
Figure 23 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and restricting $\varepsilon$ to be the same for both diseases. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\varepsilon = 2^1$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. Observing the ROC and PRC, the model cannot distinguish between the two classes very well on the derivation set, and is therefore underfitting. This leads to poor performance in distinguishing the classes on the validation set.
Figure 24 RRMS vs. SVD using a single combined probability map for white matter lesions and T1 black holes, and allowing for different values of $\epsilon$ for each disease. Receiver operating characteristic curve and precision recall curve for a Monte Carlo simulation. $\epsilon_{MS} = 2^4$, $\epsilon_{SVD} = 2^5$. The histogram of scores for the derivation and validation sets is provided, with the cross entropy loss embedded. There is very good separability between the two classes on the derivation set, but without significant overfitting, as exemplified by the ROC and PRC, which leads to good performance on the validation set.

4.2.1.2 Method 2 - Dual Combined Probability Maps

We demonstrate the effect of using separate CPMs for WMLs and BHs in this section, as well as the effect of using the non-informative prior on model generalization and performance. Table 5 shows the mean performance of the cross entropy image similarity metric with separate CPMs for WMLs and BHs for the best performing hyperparameters, without and with the non-
informative prior. These results are obtained for the case where $\epsilon_{WML}$ and $\epsilon_{BH}$ are restricted to be the same for MS and SVD. Model performance improves with the non-informative prior ($p < 0.05$). A third model which makes use of the non-informative prior and allows the values of $\epsilon_{WML}$ and $\epsilon_{BH}$ to vary with disease is constructed. The results are obtained by averaging 100 Monte Carlo simulations, and are presented with standard errors.

Figure 25 shows a representative decision boundary in feature space for the Gaussian Naïve Bayes classifier for one of the 100 Monte Carlo simulations, without and with the non-informative prior. The figures are for the case where $\epsilon_{WML}$ and $\epsilon_{BH}$ are restricted to be the same for MS and SVD, as this helps make clear the effect of the hyperparameters $\epsilon_{WML}$ and $\epsilon_{BH}$ while controlling for their differences between the two diseases. Table 6 shows the average class mean ($\mu$) and standard deviation ($\sigma$) across 100 Monte Carlo simulations (see section 2.6.2 for an interpretation of these parameters), also for the case where $\epsilon_{WML}$ and $\epsilon_{BH}$ are restricted to be the same for MS and SVD. There is a significant difference in class means for both WMLs and BHs, suggesting that distinguishing the two lesion types adds predictive value.

**Table 5** RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes. First and second rows: restricting $\epsilon$ to be the same for both diseases. Third row: Allowing $\epsilon$ to vary between the two diseases. Mean performance of the image similarity metrics for the optimal hyperparameter settings using the Gaussian Naïve Bayes classifier, without and with the non-informative prior. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. Using the non-informative prior decreases derivation set performance slightly, but importantly results in greater performance on the validation set ($p < 0.05$).

<table>
<thead>
<tr>
<th>Gaussian Naïve Bayes Classifier</th>
<th>$\phi$</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o non-informative prior</td>
<td>Derivation</td>
<td>0.908 ± 0.003</td>
<td>0.916 ± 0.004</td>
<td>0.968 ± 0.002</td>
<td>0.951 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>0.635 ± 0.019</td>
<td>0.780 ± 0.015</td>
<td>0.761 ± 0.016</td>
<td>0.962 ± 0.010</td>
</tr>
<tr>
<td>w/ non-informative prior</td>
<td>Derivation</td>
<td>0.902 ± 0.003</td>
<td>0.908 ± 0.004</td>
<td>0.968 ± 0.002</td>
<td>0.946 ± 0.002</td>
</tr>
</tbody>
</table>
Table 6 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. The average class means and standard deviations by lesion type for the Gaussian Naive Bayes classifier are provided. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. The significant difference in class means for both WMLs and BHs suggests that discriminating between lesion subtypes is important in distinguishing between the two diseases.

<table>
<thead>
<tr>
<th>Model Parameter $\mu$</th>
<th>Model Parameter $\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Matter Lesion Scores</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td></td>
<td>Small Vessel Disease</td>
</tr>
<tr>
<td>Black Hole Scores</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td></td>
<td>Small Vessel Disease</td>
</tr>
</tbody>
</table>

Figure 25 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. A Gaussian Naive Bayes model constructed from a single Monte Carlo simulation using the optimal hyperparameter settings. Left: without the non-informative prior, right: with the non-informative prior. Blue corresponds to RRMS cases, and red to SVD cases. Circles represent derivation set cases used.
to construct the model, and crosses represents validation set cases. For both models, there is good separation between the two classes on the derivation set, which leads to good performance on the validation set. However, the model with the non-informative prior performs slightly better, in part because the non-informative prior causes the model to make less extreme predictions near the decision boundary.

Figure 26 demonstrates how the classification decision rule changes when varying $\epsilon_{BH}$ while keeping $\epsilon_{WML}$ fixed. For relatively large values of $\epsilon_{BH}$, there is a very large spread in BH scores between the two classes, whereas for relatively small values of $\epsilon_{BH}$, there is very little spread in BH scores between the two classes. Similarly, Figure 27 demonstrates how the classifier decision rule changes when varying $\epsilon_{WML}$ while keeping $\epsilon_{BH}$ fixed. For relatively large values of $\epsilon_{WML}$, there is a very large spread in WML scores between the two classes, whereas for relatively small values of $\epsilon_{WML}$, there is very little spread in WML scores between the two classes.

In both cases, there is an intuitive explanation for the pattern: large values of $\epsilon$ allow for large variations in cross entropy scores, thus relaxing the effect of the hyperparameter. Small values of $\epsilon$ restrict the variation in scores, thus constraining the model’s ability to distinguish the two diseases. Much like for the case of the unified CPMs, there is an analogous role of overfitting and underfitting for each of these hyperparameter settings. Large values of $\epsilon$ lead to near-perfect separability between the two classes on the derivation set, but poor performance on the validation set. This is because large values of $\epsilon$ magnify the impact of lesion differences on the derivation set, which may not generalize to the validation set. Small values of $\epsilon$ lead to underfitting on the derivation data, because the effect of the true underlying difference between the two classes is restricted. This in turn results in poor performance on the validation set. As with the unified CPMs, a balance between the two extremes leads to optimal performance.

Observe that the decision boundary between MS and SVD (the white line between the two classes in each figure) is very fine – most of the feature space corresponds to a probability
prediction of nearly 0% and 100% for each disease. This matches the conventional wisdom that the Naïve Bayes classifier biases its prediction probabilities towards 0 or 1, and may motivate the need for probability calibration to better represent the actual probability of disease, as investigated in section 2.6.2. However, the effect is more noticeable for the Naïve Bayes model without the non-informative prior. This is because the non-informative prior causes the model to make less extreme predictions near the decision boundary, and thus improves model generalization. Finally, observe that without the non-informative prior, most of the feature space corresponds to an SVD classification (in red). This is due to the class imbalance in the problem, which causes SVD to have a bigger effect on the model parameter estimates. The class predictions are more balanced when using the non-informative prior, as the extreme estimates of maximum likelihood are avoided. However, it should be noted that since we use $\phi$, and not accuracy to measure model performance, this does not affect the model performance for the best hyperparameter settings.
Figure 26 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Left column: without the non-informative prior, right column: with the non-informative prior. Varying $\epsilon_{BH}$ from high (top) to low (bottom) leads to smaller separation in BH scores between the two classes. High values of $\epsilon_{BH}$ lead to overfitting to the derivation set, and small values lead to model underfitting. Intermediate values of $\epsilon_{BH}$ (center) lead to optimal performance.
Figure 27 RRMS vs. SVD using separate combined probability maps for white matter lesions and T1 black holes, and restricting $\epsilon$ to be the same for both diseases. Left column: without the non-informative prior, right column: with the non-informative prior. Varying $\epsilon_{\text{WML}}$ from high (top) to low (bottom) leads to smaller separation in WML scores between the two classes.
High values of $\epsilon_{WML}$ lead to overfitting to the derivation set, and small values lead to model underfitting. Intermediate values of $\epsilon_{WML}$ (center) lead to optimal performance.

4.2.2 Method 3 - Using Clinical Features Alone

Table 7 presents the performance metrics for distinguishing between RRMS vs. SVD using clinical features alone. We also provide results for distinguishing between RRMS & SPMS vs. SVD to better appreciate how differences in patient age and sex for RRMS and SPMS affect classifier performance. The results were obtained by averaging 100 Monte Carlo simulations, and are presented with standard errors.

The classifier used to distinguish RRMS vs. SVD performs relatively well. This is because there is a higher proportion of females vs. males in RRMS compared to SVD, and the difference in age between the two diseases is larger. The model used to differentiate RRMS & SPMS vs. SVD does not perform as well ($p < 0.05$), however. This can be explained by the higher age of SPMS patients, and the more balanced female to male ratio, which reduces the discriminative power of these features in distinguishing between MS and SVD. Note that the model used to distinguish between RRMS vs. SVD achieves very high sensitivity (correctly identifies MS patients), but at the expense of lower precision (misdiagnoses patients with MS). This presents an important trade-off in disease discrimination.

Table 7 Mean performance for distinguishing between RRMS vs. SVD and RRMS & SPMS vs. SVD using clinical features alone. The model used is the Naïve Bayes classifier with a non-informative prior. Results are averaged over 100 Monte Carlo simulations, and are presented with standard errors. Higher performance is achieved when distinguishing between RRMS vs. SVD compared to RRMS & SPMS vs. SVD ($p < 0.05$), due to the smaller age discrepancy and
smaller sex ratio difference between SPMS and SVD patients, compared to RRMS and SVD patients.

<table>
<thead>
<tr>
<th>Naïve Bayes Classifier</th>
<th>Derivation Set</th>
<th>Validation Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRMS vs. SVD</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Sensitivity</td>
</tr>
<tr>
<td></td>
<td>0.865 ± 0.002</td>
<td>0.96 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>0.857 ± 0.008</td>
<td>0.95 ± 0.008</td>
</tr>
<tr>
<td>RRMS &amp; SPMS vs. SVD</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.715 ± 0.003</td>
<td>0.841 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>0.739 ± 0.011</td>
<td>0.906 ± 0.006</td>
</tr>
</tbody>
</table>

4.2.3 Method 4 - Using Neuroimaging and Clinical Features

A Naïve Bayes model using neuroimaging and clinical features as predictors, with a non-informative prior and allowing the value of $\epsilon$ to vary between the diseases, was developed. The performance for the model with the optimal hyperparameter settings is shown in Table 8 for RRMS vs. SVD. From the results, we can conclude that combining neuroimaging and clinical features leads to the best classifier performance ($p < 0.05$).

<table>
<thead>
<tr>
<th>Naïve Bayes Classifier</th>
<th>Derivation</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\phi$</td>
<td>Precision</td>
</tr>
<tr>
<td>$\epsilon_{MS, WML} = 2^4, \epsilon_{SVD, WML} = 2^2, \epsilon_{MS, BH} = 2^4, \epsilon_{SVD, BH} = 2^4$</td>
<td>1.000 ± 0.000</td>
<td>1.000 ± 0.000</td>
</tr>
<tr>
<td></td>
<td>0.971 ± 0.006</td>
<td>0.993 ± 0.004</td>
</tr>
</tbody>
</table>

4.2.4 Monte Carlo Simulation Performance Convergence

Figure 28 demonstrates empirically that 100 Monte Carlo cross validation simulations was sufficient for each of the four models proposed, as performance on the validation set convergences.
Figure 28 RRMS vs. SVD. Monte Carlo simulation performances converged after approximately 100 simulations. Top-left: model developed using a single combined probability map for white matter lesions and $T1$ black holes. Top-right: model developed using separate combined probability maps for white matter lesions and $T1$ black holes. Bottom-left: model developed using clinical features alone. Bottom-right: model developed using combined neuroimaging and clinical features.

4.2.5 Model Interpretability

Figure 29 shows $r_{WML}$ and $r_{BH}$, which are defined mathematically in section 3.2.4.2, for an MS case. For $r_{WML}$, most of the image has values above zero, which contributes to a neuroimaging score indicative of MS. Interestingly, the lesions occur in the periventricular region, which is known to occur more frequently in MS than in SVD. For $r_{BH}$, most of the
lesions are also indicative of MS, except for one lesion, which corresponds to the anterior limb of the internal capsule; a brain region regularly affected in SVD patients. The algorithm consolidates the scores from every region, separately for the two lesion types, to correctly diagnose the patient with MS.

These remarkable visual representations of the neuroimaging scores suggest that the algorithm’s findings can be consolidated with what is already clinically known, as shown in the literature view (see Table 2). Perhaps more importantly, the algorithm’s findings may serve to guide future clinical work in discovering brain regions of interest for disease discrimination. Finally, this visual representation serves as a decision support tool for the clinician using the algorithm.

Figure 29 A visualization of the algorithm’s decision-making for an MS case. Positive values contribute to a diagnosis of MS, and negative values to a diagnosis of SVD. Left) white matter lesions - the algorithm has discovered, from data alone, that periventricular lesions are associated with a diagnosis of MS. Right) black holes – most of the lesions contribute to a diagnosis of MS, except for one region which is the anterior limb of the internal capsule, a region known to affect SVD patients. The algorithm consolidates the findings from both regions, and correctly classifies the patient with MS.
4.2.6 Final Model Selection and Performance on Test Set

The model with the highest performance on the validation sets is the Naïve Bayes classifier with neuroimaging and clinical features as predictors (p < 0.05), as described in section 4.2.3. This model was trained on the entire training set, using the same hyperparameters that were found optimal in section 4.2.3. The performance of the model is excellent, it achieves perfect performance on the training set, and near-perfect performance on the test set, with only a single error on 21 cases. The only incorrectly classified patient was an MS patient - a male in their low-to-mid fifties with very low TLV – it would seem that, given the information available to make the classification, the algorithm’s decision-making was sensible. Of course, this is further evidenced by its correct classification of the remaining 20 cases.
5. Discussion

In this work, we have successfully developed automatic and accurate algorithms for distinguishing between MS and SVD. As the societal and monetary cost of both diseases is high, our work has real potential for clinical impact. Furthermore, our models have, from data alone, shown that combining neuroimaging and clinical features leads to better performance than using either in isolation, and that discriminating between WMLs and BHs leads to better performance. Although our primary aim has been distinguishing between MS and SVD, we have in the process developed novel and reliable WML and BH segmentation algorithms. Furthermore, many published lesion segmentation algorithms have been designed to perform well on a particular disease and require several image modalities in combination to perform well. In contrast, our algorithms are general-purpose WML and BH segmentation algorithms, and use only two image modalities which can be used in isolation. Our approach is not sensitive to the intensity scale or image contrast due to different scanners, as evidenced by our excellent performance on the publicly available ISBI dataset. Further, our model performed well on the ISBI dataset despite not taking advantage of the longitudinal nature of the dataset, and we used only the raw images for segmentation, as opposed to the pre-processed ones. Finally, our algorithms can be deployed on images of any size and resolution.

In comparison to the current diagnostic criteria used for the differential diagnosis of MS and SVD, our method is entirely data-driven, uses quantitative measures alone, and employs only neuroimaging and clinical features. Furthermore, our approach makes a diagnosis from one presentation alone, which is in contrast to the current diagnostic criteria of "dissemination of lesions in space and time" for MS (Polman et al. 2011) (J Thompson et al. 2017), which requires multiple presentations. However, we have considered the discrimination of only two diseases, which has inherently reduced the complexity of the task.
The interpretation of the results for each classifier is as follows. For neuroimaging alone, we often obtained nearly 100% accuracy on the derivation sets, but lower performance on the validation sets. This highlights a common phenomenon known as overfitting - images used to build the model are likely to be classified correctly, since the model is trained using these cases. The effect is particularly acute due to the curse of dimensionality - we have a large feature space (i.e. the number of voxels in the MRI scans) relative to the number of samples, which makes the classifier sensitive to small and possibly random fluctuations in lesion distribution. We chose to employ the $\phi$ coefficient for identifying the optimal threshold due to its insensitivity to class imbalances.

Although combining neuroimaging and clinical features led to the best classifier, using clinical features alone had a high classification performance. Viewed naively, this may suggest that neuroimaging is superfluous, given the added cost to the healthcare system associated with it. However, in terms of a diagnostic odds ratio, a performance increase from 90% to 95%, or 98% to 99%, presents a doubling in performance, as the error rate in the test decreases by a factor of two in each case. That is, improvements in performance should be viewed relative to the benchmark level of performance. Viewed in this sense, the increase in performance when combining neuroimaging and clinical features is significant.

Constructing separate CPMs for WMLs and BHs improved classifier performance. This may be because WMLs and BHs are manifestations of different underlying pathology, and thus a separate treatment of the two is advantageous. By extension, it is conceivable that distinguishing other disease subtypes, such as cortical lesions and brain atrophy (for longitudinal studies), as well as analyzing lesions by brain region, can further improve the discrimination of MS vs. SVD. However, we caution against using a priori knowledge indiscriminately, as this may limit models in discovering yet-unknown relations useful for disease discrimination. Finally, combining neuroimaging with clinical features led to better performance than using either in isolation.
Our results have provided empirical evidence for the need for Monte Carlo cross validation, as the uncertainty in validation set performance was high when employing a small number of simulations, but demonstrated convergence across 100 simulations. The need for Monte Carlo simulation was particularly important in this study, as we were restricted to a small sample size for model validation. This approach to cross validation maximizes dataset utility and can be used in other rare diseases.

There are several limitations to the current work. First, because our RRMS cases had higher total lesion volumes than the SVD cases, it is possible that our classifiers became sensitive to these differences, instead of identifying the true underlying differences in lesion distribution. Further investigation of this may prove valuable, as lesion loads in the population at large may differ in distribution from our dataset. One approach to isolate for the effect of lesion load would be to develop CPMs based on stratified lesion load groups – for example, within the derivation set, separate CPMs could be developed for lesion volumes between 0 – 5.0 cm³, 5.0 – 10.0 cm³, and 10.0 + cm³. Thereafter, given a new validation or test case, the image would first have its total lesion volume calculated using the lesion segmentation algorithms, and thereafter the appropriate CPM group would be used for disease discrimination. Nevertheless, the classifiers we have developed perform significantly better than using lesion load alone to distinguish the two diseases, providing strong evidence that the cross entropy image similarity metric is of significant value. Furthermore, such strategies can succumb to indiscriminate stratification, as it could be argued that perpetually stratifying by lesion load would recurrently increase performance on a given dataset. There is likely a balance in stratification by lesion load that would lead to the best classifier performance, and this balance would need to be validated with a sizeable dataset. Finally, we considered identifying the brain regions which contributed to disease discrimination best, as a verification of whether our data-driven approach reflected the findings from the summary tables in section 2.3. Our approach would have involved subtracting the CPMs for RRMS and SVD, and thereafter assessing whether the mean difference by brain region of interest was significantly different than zero via paired t-tests, while accounting for multiple comparisons. However, we suspected this would lead to the finding that lesions in most brain regions would be indicative of MS, due to the larger
average lesion load in MS compared to SVD in our dataset. Future steps include conducting the above analysis while controlling for the effects of total lesion volume, an analogous approach to developing CPMs by stratified lesion load groups as indicated above.

We have thus far evaluated our models by maximizing the $\phi$ coefficient. This inherently values the cost of misdiagnosis between the diseases equally - in our evaluations, we did not weigh the relative cost of misdiagnosis of MS and SVD. Nevertheless, quantifying this difference in cost is not trivial, as subjective measures such as impact on quality of life for misdiagnosis must be consolidated with quantitative measures such as the cost to the healthcare system for. This presents a trade-off in the false positive and false negative rate of diagnosis.

An important limitation of this study is the lack of ground truth tracings for SVD. We have not investigated segmentation algorithms for sagittal or coronal sequences, although in principle such images can be converted into axial ones. Another limitation is the small dataset size used. Finally, our segmentation algorithms have only been validated on pre-contrast MRI; performance on post-contrast MRI may differ significantly.

One important aspect of the model is its interpretability. We developed an approach to detail the relative importance of each feature in the classification decision, and more interestingly, which lesions in the brain were most influential in making the classification. Identifying the most influential brain regions would help consolidate the medical community’s current understanding of lesion load distribution for each disease, as discussed in the Literature Review, with the findings the algorithm makes. The increased interpretability of the algorithm would also facilitate its use as a CDSS.

Future steps include applying our algorithms to the discrimination of pediatric MS and vasculitis. Success would demonstrate the generalizability of our approach, both for evaluating and distinguishing other disease types, and to variability in patient demographics. Our model
can easily be extended to the discrimination of multiple diseases that have overlapping presentation and a variety of underlying disease processes. This would maximize its utility, as it is often the case that multiple diseases with similar presentation compete in the differential diagnosis. Disorders of interest include migraines, psychiatric disorders, acute disseminated encephalomyelitis (ADEM), and infectious diseases. As eluded to earlier, a future goal is to investigate the development of CPMs based on stratified lesion load groups – the purpose being to control for lesion load in the differential diagnosis. We envisage a larger dataset will be required, as partitioning our already limited dataset would further introduce variance into CPM lesion distribution.

Finally, we consider what adjustments would need to be made to deploy the model in practice. One aspect is a Bayesian framework for the prior probability of classes. That is, in the models developed thus far, we have considered the neuroimaging and clinical features that help make the classification, but we have done this without context of the prior probability of each disease. As shown in section 2, disease prevalence for MS is lower than for SVD. A Bayesian framework would integrate this prior probability in the decision-making process, although additional complexities could arise. For example, not only would the prior probability of each disease have to be known, but the likelihood that a patient would present themselves for an assessment – we suspect MS patients would present themselves more often, as their symptoms are uncharacteristic of the general population for their generally lower age group.
6. Conclusions

The algorithms developed and employed in this work have successfully distinguished, with high performance, two diseases of the CNS that have traditionally been considered difficult to distinguish using neuroimaging and clinical features alone. In addition, we found that combining neuroimaging and clinical features led to the best classifier performance in distinguishing MS from SVD.
7. References


Geraldes, Ruth, Margaret M Esiri, Gabriele De Luca, and Jacqueline Palace. 2016. “Age-Related Small Vessel Disease: A
Potential Contributor to Neurodegeneration in Multiple Sclerosis.” *Brain Pathology* (Zurich, Switzerland) 27.


88


Appendix A

We begin by deriving the maximum a posteriori parameter estimates for the Bernoulli and Gaussian distributions, and thereafter motivate the need for the Bayesian approach using non-informative priors. The derivation is similar in style to (S. J. D. Prince 2012).

Maximum A Posteriori

The conjugate prior of the Bernoulli distribution is the Beta distribution, whose probability density function is given by

\[ p(\theta_{j,k}) = \frac{\Gamma(\alpha_{j,k} + \beta_{j,k})}{\Gamma(\alpha_{j,k})\Gamma(\beta_{j,k})} \theta_{j,k}^{\alpha_{j,k}-1}(1 - \theta_{j,k})^{\beta_{j,k}-1}, \]

where \( \Gamma \) is the gamma function. This leads to the following maximum a posteriori estimate for \( \theta_{j,k} \)

\[ \hat{\theta}_{j,k} = \frac{\sum N_k x_j^{(n)} + \alpha_{j,k} - 1}{N_k + \beta_{j,k} + \alpha_{j,k} - 2}. \]

Thereafter, the probability of a new sample is estimated using the probability mass function of the Bernoulli distribution as usual. We can see that by setting \( \alpha \) and \( \beta \) to high values, we effectively introduce artificial “counts” for each feature, which leads to less extreme values for \( \theta \). The higher the values of \( \alpha \) and \( \beta \), the greater the effect of smoothing, the effect of which is more noticeable for small datasets. Note that setting \( \alpha = 1 \) and \( \beta = 1 \), which corresponds to uniform prior distribution, leads to the intuitive result of no smoothing, i.e. the maximum likelihood estimate.

The conjugate prior of the Gaussian distribution is the normal-inverse-gamma distribution, whose probability density function is given by
\[ p(\mu_{j,k}, \sigma_{j,k}) = \frac{\sqrt{\gamma_{j,k}} \beta_{j,k}^{\alpha_{j,k}}}{\sigma_{j,k} \sqrt{2\pi} \Gamma(\alpha_{j,k})} \left( 1 \sigma_{j,k}^{2} \right)^{\alpha_{j,k}+1} \exp\left[-\frac{2\beta_{j,k} + \gamma_{j,k} (\delta_{j,k} - \mu_{j,k})^{2}}{2\sigma_{j,k}^{2}}\right] \]

This leads to the following maximum a posteriori estimates for \( \mu_{j,k} \) and \( \sigma_{j,k} \)

\[ \hat{\mu}_{j,k} = \frac{\sum_{n}^{N_{k}} x_{j}^{(n)} + \gamma_{j,k} \delta_{j,k}}{N_{k} + \gamma_{j,k}}, \hat{\sigma}_{j,k}^{2} = \frac{\sum_{n}^{N_{k}} (x_{j}^{(n)} - \mu_{j,k})^{2} + 2\beta_{j,k} + \gamma_{j,k} (\delta_{j,k} - \mu_{j,k})^{2}}{N_{k} + 3 + 2\alpha_{j,k}} \]

Thereafter, the probability of a new sample is estimated using the probability density function of the Gaussian distribution as usual. Examining the formula for \( \hat{\mu} \), we can see there are two contributing components, the first corresponding to the maximum likelihood estimate, and the second to the prior. For small values of \( \gamma \), the prior has little effect and the parameter estimate approaches the maximum likelihood estimate, and for high values of \( \gamma \), the prior dominates the estimate. The formula for \( \hat{\sigma} \) has a similar interpretation when considering the model hyperparameters in combination.

Although the maximum a posteriori estimates minimize over-confidence in the model parameters, the hyperparameters \( (\alpha, \beta) \) and \( (\alpha, \beta, \gamma, \delta) \) for the Beta and normal-inverse-gamma distributions must be tuned. Alternatively, these parameters could be set \textit{a priori}, but this introduces bias into the estimates, which is particularly problematic for small datasets. An additional complication is that the effect of these hyperparameters depends on the scale of the feature, which is problematic for features of differing scale such as the neuroimaging score and patient age. This motivates the need for a Bayesian approach to parameter estimation using a non-informative prior.

**Bayesian Approach**

In the Bayesian approach, we calculate a posterior over the model parameters. For the Bernoulli distribution, this is given by
\[ p \left( \theta_{j,k} \mid x_j^{(n)} \ldots (N_k) \right) = \frac{\prod_{n=1}^{N_k} p\left(x_j^{(n)} \mid \theta_{j,k} \right) p\left(\theta_{j,k} \right)}{p\left(x_j^{(n)} \ldots (N_k) \right)} \]

\[ p \left( \theta_{j,k} \mid x_j^{(n)} \ldots (N_k) \right) = \frac{\prod_{n=1}^{N_k} \text{Bern}_j^{(n)}(\theta_{j,k}) \text{Beta}_{\theta_{j,k}}(\alpha_{j,k}, \beta_{j,k})}{p\left(x_j^{(n)} \ldots (N_k) \right)} \]

\[ p \left( \theta_{j,k} \mid x_j^{(n)} \ldots (N_k) \right) = \text{Beta}_{\theta_{j,k}}(\tilde{\alpha}_{j,k}, \tilde{\beta}_{j,k}) \]

where \( \tilde{\alpha}_{j,k} = \sum_{n=1}^{N_k} x_j^{(n)} + \alpha_{j,k} \) and \( \tilde{\beta}_{j,k} = \sum_{n=1}^{N_k} x_j^{(n)} + \beta_{j,k} \)

The advantage of using conjugate priors is now clear – the corresponding posterior distribution is available in closed form, with the same distribution as the prior but with different parameters.

To make class predictions for new data \( x^* \), we compute the weighted average of predictions given by the posterior distribution over parameters

\[ p(C^* = k \mid x_j^*) = \int p(x_j^* \mid \theta_{j,k}) p(\theta_{j,k} \mid x_j^{(n)} \ldots (N_k)) d\theta_{j,k} \]

\[ p(C^* = k \mid x_j^*) = \int \text{Bern}_j^{(n)}(\theta_{j,k}) \text{Beta}_{\theta_{j,k}}(\tilde{\alpha}_{j,k}, \tilde{\beta}_{j,k}) d\theta_{j,k} \]

\[ p(C^* = k \mid x_j^*) = \frac{\sum_{n=1}^{N_k} x_j^{(n)} + \alpha_{j,k}}{N_k + \beta_{j,k} + \alpha_{j,k}} \]

For the Gaussian distribution, the posterior over the model parameters is given by

\[ p \left( \mu_{j,k}, \sigma_{j,k}^2 \mid x_j^{(n)} \ldots (N_k) \right) = \frac{\prod_{n=1}^{N_k} p\left(x_j^{(n)} \mid \mu_{j,k}, \sigma_{j,k}^2 \right) p\left(\mu_{j,k}, \sigma_{j,k}^2 \right)}{p\left(x_j^{(n)} \ldots (N_k) \right)} \]

\[ p \left( \mu_{j,k}, \sigma_{j,k}^2 \mid x_j^{(n)} \ldots (N_k) \right) = \frac{\prod_{n=1}^{N_k} \text{Norm}_j^{(n)}(\mu_{j,k}, \sigma_{j,k}^2) \text{NormInvGamma}_{\mu_{j,k}, \sigma_{j,k}^2}(\alpha_{j,k}, \beta_{j,k}, \gamma_{j,k}, \delta_{j,k})}{p\left(x_j^{(n)} \ldots (N_k) \right)} \]

\[ \text{Norm}_j^{(n)}(\mu_{j,k}, \sigma_{j,k}^2) \text{NormInvGamma}_{\mu_{j,k}, \sigma_{j,k}^2}(\alpha_{j,k}, \beta_{j,k}, \gamma_{j,k}, \delta_{j,k}) \]
\[ p\left(\mu_{j,k}, \sigma_{j,k}^2 \mid x_j^{(n)}(N_k)\right) = \text{NormInvGamma}_{\mu_{j,k}, \sigma_{j,k}^2} (\bar{\alpha}_{j,k}, \bar{\beta}_{j,k}, \bar{\gamma}_{j,k}, \bar{\delta}_{j,k}) \]

where \( \bar{\alpha}_{j,k} = \alpha_{j,k} + \frac{N_k}{2}, \)

\[ \bar{\beta}_{j,k} = \frac{\sum_{j=1}^{N_k} x_j^{(n)} + \beta_{j,k}}{2} - \frac{(\beta_{j,k} + \sum_{j=1}^{N_k} x_j^{(n)})^2}{2(\beta_{j,k} + N_k)} , \]

\[ \bar{\gamma}_{j,k} = \gamma_{j,k} + N_k, \]

\[ \bar{\delta}_{j,k} = \frac{(\beta_{j,k} + \sum_{j=1}^{N_k} x_j^{(n)})}{\gamma_{j,k} + N_k} . \]

Again we see the advantage of using conjugate priors, as the corresponding posterior distribution is available in closed form, with the same distribution as the prior but with different parameters.

To make predictions for new data \( x^* \), we compute the weighted average of predictions given by the posterior distribution over parameters

\[ p(C^* = k \mid x_j^*) = \int \int p(x_j^* \mid \mu_{j,k}, \sigma_{j,k}^2) p\left(\mu_{j,k}, \sigma_{j,k}^2 \mid x_j^{(n)}(N_k)\right) d\mu_{j,k} d\sigma_{j,k} \]

\[ p(C^* = k \mid x_j^*) = \int \int \text{Norm}_{x_j^*} (\mu_{j,k}, \sigma_{j,k}^2) \text{NormInvGamma}_{\mu_{j,k}, \sigma_{j,k}^2} (\bar{\alpha}_{j,k}, \bar{\beta}_{j,k}, \bar{\gamma}_{j,k}, \bar{\delta}_{j,k}) d\mu_{j,k} d\sigma_{j,k} \]

\[ p(C^* = k \mid x_j^*) = \frac{1}{\sqrt{2\pi} \sqrt{\bar{\gamma}_{j,k}} \bar{\alpha}_{j,k} \Gamma(\bar{\gamma}_{j,k})} \frac{\bar{\alpha}_{j,k}}{\sqrt{\bar{\gamma}_{j,k} \bar{\beta}_{j,k}}} \Gamma(\bar{\alpha}_{j,k}) \]

where \( \bar{\alpha}_{j,k} = \bar{\alpha}_{j,k} + \frac{1}{2} \),

\[ \bar{\beta}_{j,k} = \frac{x_j^2}{2} + \bar{\beta}_{j,k} + \frac{\gamma_{j,k} \delta_{j,k}^2}{2} - \frac{(\bar{\gamma}_{j,k} \delta_{j,k} + x_j)^2}{2(\bar{\gamma}_{j,k} + 1)}, \]

\[ \bar{\gamma}_{j,k} = \bar{\gamma}_{j,k} + 1. \]

96
Now, we must optimize the hyperparameters of the Beta and normal-inverse-gamma conjugate prior distributions. Alternatively, we can use a non-informative prior, as this would limit model bias. Denote by $I(\theta)$ the Fisher information for an unknown parameter $\theta$ that models $X$. Under certain conditions of differentiability for $\theta$, we can write the Fisher information as

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log(f(X; \theta)) \mid \theta \right]$$

Jeffreys prior, a non-informative prior (Jeffreys 1946), is given by $p(\theta) \propto \sqrt{\text{det} I(\theta)}$. For a Bernoulli distribution, the non-informative prior is the Beta distribution with $\alpha = \frac{1}{2}, \beta = \frac{1}{2}$. Thus, we set these hyperparameters in the above Bayesian formulation for the Bernoulli distribution. For the Gaussian distribution, the non-informative prior is the normal-inverse-gamma distribution with $\alpha = 0, \beta = 0, \gamma = 0, \delta = 0$, and we set these accordingly as well.