THE REPUBLIC OF TURKEY

BAHÇEŞEHİR UNIVERSITY

AUTOMATIC SEGMENTATION, QUANTIFICATION AND RATING OF PERIVENTRICULAR WHITE MATTER HYPERINTENSITY IN DEMENTED PATIENTS

Doctoral Thesis

LEONARDO O. IHEME

İSTANBUL, 2018

THE REPUBLIC OF TURKEY
BAHÇEŞEHİR UNIVERSITY

GRADUATE SCHOOL OF NATURAL & APPLIED SCIENCES
ELECTRICAL & ELECTRONICS ENGINEERING

AUTOMATIC SEGMENTATION, QUANTIFICATION AND RATING OF PERIVENTRICULAR WHITE MATTER HYPERINTENSITY IN DEMENTED PATIENTS

Doctoral Thesis

LEONARDO O. IHEME

Supervisor: ASSOC. PROF. DR. DEVRİM ÜNAY

İSTANBUL, 2018

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
ELECTRICAL AND ELECTRONICS ENGINEERING

Name of the thesis: Automatic Segmentation, Quantification And Rating Of Periventricular White Matter Hyperintensity In Demented Patients
Name/Last Name of the Student: Leonardo O. Iheme
Date of the Defense of Thesis:04.06.2018

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assist. Prof. Yücel Batu SALMAN
Graduate School Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Arts.

Assist. Prof. Cavit Fatih KÜÇÜKTEZCAN
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Arts.

Examining Comittee Members	Signature
Thesis Supervisor Assoc. Prof. Dr. Devrim UNAY	
Member Assoc. Prof. Dr. Burcu TUNGA	
Member Assoc. Prof. Dr. Mehmet Alper TUNGA	
Member Asst. Prof. Dr. Mehmet TÜRKAN	
Member Asst. Prof. Dr. Tevfik AYTEKIN	

To Moji Iheme

AKNOWLEDGEMENTS

I would like to thank my supervisor, Associate Professor Devrim Ünay, for guiding me through my entire PhD; for his patience, encouragement and kindness. This work would not have been possible without him.

I am immensely grateful to our collaborators at Bayındır Hospital, especially Assoc. Prof Dr. Melek Kandemir and Assoc.Prof. Dr. Zehra Betül Yalçıner for their invaluable professional contribution during the course of the study. To my PhD thesis committee members, Assoc. Prof. Dr. Mehmet Alper Tunga, Asst. Prof. Dr. Tevfik Aytekin, Assoc. Prof. Dr. Burcu Tunga, and Asst. Prof. Dr. Mehmet Türkan for their academic guidance.

To my colleagues at the Medical Image Processing Lab at Bahçeşehir University, Osman Başkaya and Dr. İlkay Öksüz: I am thankful for the brainstorming sessions, coding tips and the positive work environment.

The completion of this work would not have been possible without the support from CETC Istanbul, particularly, Jasmin Sakka, Kursat Ergul & Kemal Koskas. Finally yet importantly, I acknowledge my: Dr. Moji, Ije, Reni & Dr. Andee for their unconditional support and love.

İstanbul, 2018 Leonardo O. IHEME

ABSTRACT

AUTOMATIC SEGMENTATION, QUANTIFICATION AND RATING OF PERIVENTRICULAR WHITE MATTER HYPERINTENSITY IN DEMENTED PATIENTS

Leonardo O. Iheme

Electrical and Electronics Engineering

Thesis Supervisor: Assoc. Prof. Dr. Devrim ÜNAY

June 2018 56 pages

This thesis details the automatic segmentation and novel rating of periventricular white matter hyperintensity as well as the two datasets on which the segmentation and rating algorithms were tested. White matter hyperintensity, as revealed by various studies, is a biomarker of Alzheimer's disease among others. The two types of white matter lesions (periventricular and deep), based on their spatial and anatomical location, point to distinctive neurodegenerative diseases. The perivetricular white matter lesions are associated with dementia, which is why our study focuses on the segmentation and rating of such lesions. The algorithms presented were tested on two datasets that have manual segmentation masks and possess different characteristics, but were rated by the same experts.

Three lesion segmentation algorithms are presented and the superiority of Conditional Random Fields over the thresholding and Support Vector Machines based algorithms is demonstrated. The relatively low performance of the thresholding based algorithm was expected due to its simple and fast but rigid nature. The results, presented in terms of the Dice score, improved with support vector machines. However, since neither of algorithms take pixel neighborhood dependencies into account, they did not perform well, particularly on subjects with small lesions. Conditional random fields on the other hand, produced highly accurate segmentation results. We concluded that the application of machine learning based algorithms are necessary for the automatic segmentation of white matter lesions. Furthermore, the result of segmentation is improved by the incorporation of neighborhood dependencies.

The white matter hyperintensity signature is a novel contribution of our work. It is developed as a new tool for visualization and automatic rating of periventricular white matter. Based on the rating scale developed by our clinical experts, features from the signature were extracted, a classifier was trained and periventricular grades of future unseen subjects were predicted. The results obtained from the MS Challenge dataset were excellent while those from the Bayindir dataset were acceptable. After analyses and discussions, we acknowledge that our prediction accuracy would improve with a more balanced dataset even though the problem of data imbalance is prevalent in the fields of machine learning and medicine.

Keywords: White matter hyperintensity segmentation, Periventricular lesions, White matter hyperintensity signature, Support vector machines, Conditional random fields

ÖZET

DEMANSLI HASTALARDA PERVENTRİKÜLER BEYAZ CEVHER HİPERİNTENSİTESİNİN OTOMATİK BÖLÜTLENMESİ, ÖLÇÜMÜ VE DERECELENDİRİLMESİ

Leonardo O. Iheme

Elektrik-Elektronik Mühendisliği

Tez Danışmanı: Doç.Dr. Devrim ÜNAY

Haziran 2018 56 Sayfa

Bu tez, periventriküler beyaz cevher hiperintensitesinin otomatik bölütlenmesi ve özgün biçimde derecelendirilmesinin yanısıra bölütleme ve derecelendirme yöntemlerinin sınandığı iki verisetini detaylandırmaktadır. Geçmiş çalışmalarda gösterildiği üzere beyaz cevher hiperintensitesi Alzheimer hastalığının biyo-işaretlerinden biridir. Beyaz cevher lezyonlarının iki türü (periventriküler ve derin) lezyonların uzamsal ve anatomik lokalizasyonlarına bağlı olarak tanımlanır ve farklı nörodejeneratif hastalıklara işaret eder. Periventriküler beyaz cevher lezyonları demans ile ilişkilendirilmiş olduğundan çalışmamız bu lezyonların bölütlenmesi ve derecelendirilmesine odaklanmıştır. Sunulan yöntemler elle bölütleme maskelerini de içeren, farklı özelliklere sahip, görsel derecelendirilmeleri aynı uzmanlar tarafından gerçekleştirilmiş iki verisetinde sınanmıştır.

Üç lezyon bölütleme yöntemi sunulmuş, şartlı rastgelen alanlar temelli yöntemin eşikleme ve destek vektör makinaları temelli olanlara üstünlüğü gösterilmiştir. Eşikleme temelli yöntemin nispeten düşük başarımı, yöntemin basit ve hızlı ancak esnek olmayan doğasına bağlıdır. Dice örtüşme skoru cinsinden sunulan sonuçlar destek vektör makinalarının başarımının daha yüksek olduğunu göstermiştir. Ancak her iki yöntemde de piksel komşuluk ilişkileri dikkate alınmadığından, özellikle küçük lezyonlu vakalarda başarımları iyi değildir. Öte yandan şartlı rastgele alanlar oldukça iyi bölütleme sonuçları üretmiştir. Ayrıca, komşuluk ilişkileri de kullanıldığında bölütleme yöntemlerinin başarımı iyileşmektedir.

Çalışmamızın özgün katkılarından biri beyaz cevher hiperintensite imzasıdır. Periventriküler beyaz cevher hiperintensitesinin görselleştirilmesinde ve otomatik derecelendirilmesinde yeni bir araç olarak geliştirilmiştir. Birlikte çalıştığımız tıbbi uzmanlar tarafından geliştirilen derecelendirme skalası temel alınarak imzanın öznitelikleri çıkartılmış, bir sınıflandırıcı eğitilmiş ve yeni vakaların periventriküler dereceleri kestirilmiştir. Yöntemin başarımı MS Challenge verisetinde çok yüksek iken Bayındır verisetinde makul seviyededir. Detaylı analizler ve tartışmalar sonucunda, dengesiz veri problemine makina öğrenmesi ve sağlık alanlarında sıkça rastlandığının altını çizerek, daha dengeli bir veriseti kullanıldığında yöntemimizin kestirim doğruluğunun (prediction accuracy) artacağını öngörüyoruz.

Anahtar Kelimeler: Beyaz cevher hiperintensitesi bölütleme, Periventriküler lezyonlar, Beyaz cevher hiperintensitesi imzası, destek vektör makinaları, Koşullu rastgele alanlar

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

2D	:	2 Dimensional
3D	:	3 Dimensional
AH	:	Anterior Horn
ALVIN	:	Automatic Lateral Ventricle delIneatioN
BET	:	Brain Extraction Tool
CDR	:	Clinical Dementia Rating
CNN	:	Convolutional Neural Network
CRF	:	Conditional Random Field
CSF	:	Cerebro–Spinal Fluid
CV	:	Cross Validation
DNN	:	Deep Neural Network
DWML	:	Deep White Matter Lesion
FLAIR	:	Fluid Attenuated Inversion Recovery
fMRI	:	Functional Magnetic Resonance Imaging
FP	:	False Positive
GDS	:	Generic Depression Scale
HMM	:	Hidden Markov Model
IHF	:	Inter-Hemispheric Fissure
L-BFGS	:	Limited memory Broyden–Fletcher–Goldfarb–Shanno
LoG	:	Laplacian of Gaussian
MMSE	:	Mini Mental State Exam
MRF	:	Markov Random Field
MRI	:	Magnetic Resonance Imaging
MS	:	Multiple Sclerosis
PD	:	Proton Density
PH	:	Posterior Horn
PWMH	:	Periventricular White Matter Hyperintensity
PWMHS	:	Periventricular White Matter Hyperintensity Signature
PWML	:	Periventricular White Matter Lesion
SVM	:	Support Vector Machines
THR	:	Thresholding
WB	:	Whole Brain
WMH	:	White Matter Hyperintensity
WML	:	White Matter Lesion

1. INTRODUCTION

1.1 MOTIVATION

The development and application of machine learning algorithms, with specific application to medical imaging, is yet to reach its apex. Though applications are abundant, and several sites have, to some extent, succeeded in the automation of diagnostic processes; there is yet to be a single claim to total automation. Human intervention still prevails to a large extent.

Delving into the sphere of brain imaging, specifically MRI, the most popular application of machine learning, which is segmentation, has received a huge amount of attention in recent years and this is for good reasons. Medical practitioners have relied on visual markers, and mental estimates for diagnosis. However, this method has been found to be subjective and sometimes inconsistent due to human factors. It has led to attempts, by researchers, to automate the segmentation process using computer algorithms. While this has been a talking point between engineers and medical experts, both communities agree that an accurate segmentation algorithm will advance diagnoses in the field.

Due to the complexity and multiple functions of the different parts of the human brain, it is expedient to focus on one aspect at a time in order to comprehensively understand it and solve associated problems thereof. White matter-associated irregularities have been found to be a common denominator in various neuro-degenerative diseases including Multiple Sclerosis (MS), Parkinson's, Hypertension, and Alzheimer's which is associated with dementia. Thus, the importance of the study of white matter and its automatic segmentation cannot be overemphasized.

Abnormalities in human brain white matter manifest in various ways depending on the disease and the imaging modality. Focusing on the study of white matter in demented patients via MRIs, the characteristics can be thoroughly studied and modeled in order to automate diagnoses. In MRIs, White matter possesses distinct characteristic appearance and properties that make it attractive to both medical practitioners and engineers for the purpose of visual examination and quantitative analyses respectively.

As attractive as this field of study may seem, the availability of test data has been a bane on the progress that has been made thus far. Building a database requires a close

collaboration between medical and technical personnel, which may be achievable but also laborious. Furthermore, the availability of public databases is limited in most cases to the proprietary body.

This thesis will take the aforementioned facts into account and offer an automatic segmentation, quantification and rating of white matter hyperintensities in demented patients along with a supporting reference database.

1.2 NOVEL CONTRIBUTIONS OF THE THESIS

The contributions of the thesis are as follows:

i Database: we present a database that consists of the MRIs of 103 subjects; out of which 52 are training data and the rest are used for testing. Our database is unique because our team of experts have manually annotated and graded the periventricular white matter lesions of the test subjects. Other sources of data do not have such annotations and ratings available

ii Automatic periventricular white matter hyperintensity segmentation using Conditional Random Fields and automatic lateral Ventricle delIneation. Among the Previous studies that have focused on automatic white matter segmentation, only a few have separated the periventricular white matter from the deep white matter and furthermore, the employment of CRF to achieve the task that has never been done and therefore is a novelty.

iii Automatic rating of periventricular white matter hyperintensity using the periventricular white matter hyperintensity signature. This novel idea is an extension of the segmentation and is yet to be explored by other researchers. It has been built from the ground up because of the gap that we noticed in the field: wherein automatic segmentations are manually rated. Rating has been performed manually by experts based on visual cues but with the process having been automated, it is consistent and non-subjective.

1.3 PLAN

The thesis is organized as follows: in Chapter 2, we present the current related state-of-the-art works contributed by other researchers in the field of white matter hyperintensity segmentation. In Chapter 3, we present details of the data used in this study as well as the

preprocessing regime. Chapters 4 and 5 detail the segmentation and automatic rating respectively. Finally, conclusions are drawn in Chapter 6 along with future directions of the work.

2. STATE-OF-THE-ART

2.1 INTRODUCTION

White matter hyperintensity (WMH) is the bright signal observed in FLAIR MRIs of subjects with vascular dementia, Alzheimer's disease (AD) as well as in some healthy elderly subjects ((Admiraal-Behloul, et al., 2005); (Besson, Ebmeier, Best, & Smith, 1988) (Jokinen, et al., 2005)). It is well known that the increased T2 relaxation time of WMH regions is as a result of a wide range of pathological processes including edema, inflammation, demyelination, axonal loss and gliosis (Gibson, Gao, Black, & Lobaugh, 2010).The segmentation of WMH regions has been motivated by the possibility that WMHs have consequences for cognition in the elderly.

Pathologies reveal that the presence of white matter lesions (WML), appearing in T2-FLAIR MR images, in the brain is a symptom of many neurological disorders including multiple sclerosis (MS), Parkinson's disease, hypertension, stroke, and dementia that will eventually lead to death ((Capizzano, et al., 2004); (Brickman, Muraskin, & Zimmerman, 2009); (DeCarli, et al., 1995)). Depending on the distribution and spatial localization of WMLs, the associated diseases can be narrowed down. Anterior subcortical WMHs have been associated with normal aging while lesions located at the posterior subcortical region have been strongly tied to cognitive disorders ((Sheline, et al., 2008); (Besson, Ebmeier, Best, & Smith, 1988); (Jokinen, et al., 2005)). Furthermore, WMLs located in the periventricular watershed area (periventricular border) have been associated with a decline in processing speed and executive function ((Prins, et al., 2005); (Van den Heuvel, et al., 2006)). In (Van Straaten, et al., 2006) the authors found that the increase in progression of WMLs in this area was associated with a greater deterioration in global cognitive score. Therefore, the accurate quantification of WMLs is worth an underscore given the neurodegenerative associations and correlations pointed out by various studies.

WMLs have been broadly classified into two, namely deep white matter lesions (DWML) and periventricular white matter lesions (PWML). (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987) defined PWML as WMLs contiguous with the margins of each lateral ventricle and DWMLs as those separate to it. Even though a dichotomy exists between the classification; functional as well as histopathological correlates, in addition to

etiological factors, have proven that a separation between the two categories is relevant, (Kim, MacFall, & Payne, 2008).

2.2 WHITE MATTER HYPERINTENSITY DATABASE

The collection of data in a study of this nature is as important as the development of an algorithm because the system cannot be tested without available data. It is safe to say that data collection and algorithm development go hand in hand. However, just as algorithm development poses challenges, so does data collection.

There exists a fair amount of publicly available databases however, most of these were not specifically collected for WML studies, (Maniega, et al., 2015). A lot of effort has been made with regard to data collection but the limited amount of training data has been a bane on the progress of automation. The data provided by segmentation algorithm challenges such as (Carass A. , et al., 2017) and (Bento, Souza, Lotufo, Frayne, & Rittner, 2018) have been a step in the right direction even though the datasets are only partially released for various reasons.

It is worth noting that despite the fact that WML, its severity and rating thereof, have been extensively studied; a publicly available database that includes the expert rating is yet to be released. This lack of training data is prerequisite to automation of the rating process. Though mentions of WMH data and rating scales have been made ((Mäntylä, et al., 1997); (Wardlaw, Ferguson, & Graham, 2004); (Beyer, Aarsland, Greve, & Larsen, 2006)), the rating-image pairs are not explicitly available.

2.3 AUTOMATIC SEGMENTATION OF WHITE MATTER HYPERINTENSITY

In order to achieve accuracy and consistency in quantifying WMHs, it is imperative to have a dependable segmentation regime and it is little wonder that studies have invested so much on the automation of the segmentation process ((Van Leemput, et al., 1999); (Van Leemput, Maes, Vandermeulen, Colchester, & Suetens, 2001); (Maillard, et al., 2008); (Wu, et al., 2006); (Admiraal-Behloul, et al., 2005); (Jeon, et al., 2011)). In the recent past, some automated WMH segmentation methods were proposed in the literature, wherein the combined usage of multiple contrasts (T2-, T1-, and sometimes proton density-weighted (PD) images) for the purpose of obtaining an accurate segmentation result was employed ((Anbeek, Vincken, Osch, Bisschops, & Grond, 2004); (Schwartz,

et al., 2009)). While these studies excelled in their rights, they were not shy of shortcomings. The high computational cost and the availability of data were two factors that did not favor such studies.

Since WMLs appear as high intensity voxels in FLAIR MRIs, one would intuitively assume that a simple thresholding algorithm would achieve the required segmentation. The results obtained in (Iheme, et al., 2013) show that this is not the case. While thresholding is efficient in capturing the high intensity voxels, it fails to take context and the sequential nature of the data into account. Thus, the results of thresholding are riddled with false positives and would require extensive post processing regimes, (Gibson, Gao, Black, & Lobaugh, 2010).

The works of (Maillard, et al., 2008) and (Admiraal-Behloul, et al., 2005) present very good results of fully automated WMH detection techniques and are among a few that demonstrate the potential of the automatized detection methods. The current state-of-the-art WMH segmentation methods are based on artificial intelligence algorithms. (Admiraal-Behloul, et al., 2005) developed a two-level segmentation technique where the essential level involved the well-known fuzzy inference system. Using Support Vector Machines (SVM) combined with AdaBoost, (Lao, et al., Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine., 2008) achieved WMH segmentation. More recently, (Cerasa, et al., 2012) presented an application of Cellular Neural Networks (CNN) capable of automatically determining the lesion load in MS patients from MRI. The growing interest in machine learning based WML segmentation algorithms has led to review articles such as the one that was recently published by (Göçeri, Durá, & Gunay, 2016) as well as challenges such as (Carass A. , et al., 2017) where the leading results came from groups whose algorithms were machine learning based, (Catanese, Commowick, & Barillot, 2015).

Graph based algorithms such as Markov Random Field (MRF), Hidden Markov Model (HMM) and Conditional Random Field (CRF) have gained popularity in the field of medical image segmentation since they are suitable for labeling sequential data. CRFs have been found to outperform the former. Precisely, (Karimaghaloo, et al., 2012) developed a framework of CRF for automatically segmenting gadolinium-enhancing MS lesions. Their study, briefly, showed that compared to other segmentation algorithms

including MRF and SVM, CRF produces less false positives and has a better overall segmentation performance.

Recently, researchers have turned their attention to deep learning techniques, to segment white matter hyperintensity. The work of (Mohsen, et al., 2017) focused on the application of deep convolution networks to solve white matter hyperintensity segmentation problems. Their main contribution was the inclusion of location sensitive features, which they showed, improved the results of conventional convolutional neural networks and a random forest classifier. (Ricardo, et al., 2018) went as far as applying CNNs to differentiate stroke lesions from WMLs. The trend in the research community seems to have shifted from graph based segmentation algorithms to Deep Neural Networks (DNNs). State of the art results have been obtained by combining both methods. Since traditional classification models rely on hand-crafted features, this intuitively presents a limitation. However combining the power of the deep features extracted by DNNs with the ability of CRFs to model the relationship between samples has shown promise in some computer vision applications (Fayao, Guosheng, & Chunhua, 2015), (Anurag, et al., 2018). As effective as these techniques have been, there is yet to be an explicit comparison between CRFs and DNNs.

In summary, it is evident that the automatic segmentation of WMLs, is of utmost importance to the quest for accurate and consistent diagnosis of the various associated pathologies. The review article, (Caligiuri, et al., 2015), provides a summary of 34 segmentation methods published between 1980 and 2014.

2.4 AUTOMATIC RATING OF WHITE MATTER HYPERINTENSITY

Various WML rating scales have been developed ((Gerard & Weisberg, 1986); (Shimada, Kawamoto, Matsubayashi, & Ozawa, 1990); (Erkinjuntti, et al., 1994); (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987); (Breteler, et al., 1994); (Wahlund, et al., 1990)) and are being used in clinical practice ((Yue, et al., 1997); (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987)) however, the abundance of rating scales and the somewhat dissimilarity that exists between the various sites' rating procedures and criteria can prove to be a curse rather than a solution to the problem since consistency is among other factors that could hamper the standardization of WML rating. In addition, as pointed out in (Van Straaten, et al., 2006), even though visual assessments

can detect differences in WML in some groups; they display ceiling effects and poor discrimination of absolute lesion volumes. To these ends, the need for developing an automated system that is able to learn and rate WML data homogeneously cannot be overemphasized. To the best of our knowledge, until the time of the completion of this work, an automated rating system has not yet been developed in order to fill the void, which is why we are introducing the Periventricular White Matter Hyperintensity Signature.

The need for rating patients WMLs has increased over the years and this is evident in the number of rating scales that have been developed and analyzed by many researchers from various groups around the world: (Mäntylä, et al., 1997) compared 13 different rating scales; in (Pantoni, et al., 2002) the degree of heterogeneity between six different scales was investigated and in (Kapeller, et al., 2003) three rating scales were compared. Though studies have reported relatively high inter-scale agreement, they have also expressed concerns about the assumptions that were made while the conversion from one scale to another was carried out: for example, (Pantoni, et al., 2002) applied a linear regression to partly nonlinear scales.

The abundance of these scales could lead to inconsistent conclusions, (Mäntylä, et al., 1997); even more worrying is the fact that though raters are required to be well trained and experienced in order to be accurate and consistent while rating patients, it is not always the case. Furthermore, the issue of inter-rater and even intra-rater agreement arises and needs to be tackled. (Kim, MacFall, & Payne, 2008) suggested that an automatic quantitative scheme could be more accommodating to differences in classification criteria. Hence, this thesis introduces and automated rating system.

3. DATASET & PRE-PROCESSING

3.1 INTRODUCTION

The data used in carrying out experiments is as important as the other details involved in it. It is little wonder that numerous data collection related topics, in different fields of study, are being widely studied. In the fields of medical imaging research and machine learning, the quality and the quantity of data collected greatly affect the results and the findings.

In general, the items in a medical imaging dataset include but are not limited to details about the following:

i the subjects,

ii their demographics,

iii the imaging modalities, and

iv image preprocessing.

In this work, we have collected data from two sources in order to validate the performance of our algorithms objectively. Our sources of data are:

i Bayındır Hospital, İçerenköy, Istanbul

ii The Multiple Sclerosis Lesion Segmentation Challenge, 2015.

For future reference, they will be referred to as *Bayindir data* and *MS Challenge data* respectively.

3.2 DATA CHARACTERISTICS

3.2.1 Bayindir Data

The Bayindir dataset consists of cross-sectional MR images of 102 subjects in the age range of 50 to 87 years that suffered from memory problems and forgetfulness. Each subject underwent a psychiatric history and mental status examination (MMSE and CDR); a social and medical history as well as a physical examination; structural magnetic resonance imaging scans (for some depressed subjects) and laboratory tests.

Table 3.1: Demographic features of Bayindir data

		AD[1] (n=50)		Depression (n=47)	
		Range	Mean ± SD[2]	Range	Mean ± SD
Age (years)		63 – 99	78.92 ± 6.37	60 – 86	69.85 ± 6.61
Gender	Female, n (%)	20 (40%)		30 (63.8%)	
	Male, n (%)	30 (60%)		17 (36.2%)	
Education (years)		0 – 18	9.40 ± 4.50	0 – 15	10.28 ± 3.82
Duration of complaints (years)		0 – 10	1.78 ± 1.98	0.1 – 17	2.68 ± 3.59
MMSE[3]		7 – 29	22.78 ± 6.03	20 – 30	27.83 ± 2.26
GDS[4]		0 – 22	9.38 ± 5.30	1 – 23	11.59 ± 5.74
CDR[5]		0.5 – 3	1.15 ± 0.66		

[1] Alzheimer's disease
[2] Standard deviation
[3] Mini Mental State Examination
[4] Geriatric depression scale
[5] Clinical dementia rating scale

The image acquisition protocol for each subject included T1, T2, and T2-FLAIR axial sequences acquired with a 1.5T scanner. Each image was thoroughly inspected for errors in the acquisition process and patients that had MRI contraindications were excluded from the study. The mean resolution of the 52 training images included in this dataset is 0.44×0.44×5.36 (mm) and standard deviation of 0.04×0.04×1.12.

All subjects included in the study gave a written consent for their data to be used anonymously for research purposes and their demographic information is summarized in Table 3.1.

3.2.2 MS Challenge Data

The MS Challenge dataset consists of cross-sectional MR images of patients mostly diagnosed with MS. The images are available in their native space as well as in preprocessed form from which the manual segmentations were obtained. The rating was performed by our domain experts but unlike the Bayindir dataset, the MS Challenge dataset went through a more rigorous preprocessing pipeline which can be found in (Dzung, et al., 2015).

Table 3.2 summarizes the characteristics of the two datasets used for training.

3.2.3 The Rating Scale

The WMH rating system developed by Dr. Melek Kandemir and Dr. Zehra Betül Yalçıner of Bayındır Hastanesi, İçerenköy, is similar to that described by (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987) and more recently (Valdés Hernández, et al., 2014). It is a sub-set of an entire rating scale which includes other scales for whole brain atrophy, sulcal atrophy, lateral ventricle atrophy, and infarcts. Further details of the rating scales and the work that has been done in their regard were published in (Iheme, et al., 2013) and (Cash, et al., 2015). The table of grades has also been presented in Appendix A.1 TABLE 1, VISUAL MRI RATING SCALE.

Table 3.2: Characteristics of the two datasets used for experiments in this study.

	Bayindir	MS Challenge
Quantity	52	21
Mean Dimension (in mm)	$0.44 \times 0.44 \times 5.36$	$1 \times 1 \times 1$
Standard deviation	$0.04 \times 0.04 \times 1.12$	$0 \times 0 \times 0$
Pre-processing	Mild	Rigorous

Though various visual rating scales exist, that of (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987) has been extensively studied and evaluated ((Mäntylä, et al., 1997); (Pantoni, et al., 2002)); and it has been found to, in most cases, significantly correlate with other existing scales. Accordingly in this study we use an adopted version of the Fazekas scale. Our rating scale is based on the anatomical location and the volumetric size of lesions. Depending on the aforementioned factors, patients are assigned to one of the PWMH grades shown in Table 3.3.

The grades, which range from 0 to 4 and imply increasing lesion load. Patients with little or no lesions are assigned to the 0 grade while those having lesions in small quantities and located at the tips of the anterior and posterior horns of the lateral ventricle, are labeled as grade 1 or *caps* patients. Grade 2 patients typically have lesions appearing (in the axial view) as a *thin line* around the lateral ventricle. Lesions found in grade 3 patients resemble a *halo* around the lateral ventricle appearing more severe in the temporal lobe and less at the tips of the anterior horn. Patients with the most severe cases of PWMH are assigned grade 4. The lesions of such patients appear to *diffuse* from the posterior and anterior horns of the ventricle to other parts of the occipital and frontal lobes respectively. Figure 3.1 depicts an example of each of these grades.

The major difference between our rating scale and the well-established Fazekas scale is that we have included an extra grade for patients with little or no lesions whereas the Fazekas scale does not make provision for this. The distribution of subjects in our datasets with respect to grades assigned by the expert raters is shown in **Table 3.4**

The relatively low number of patients in grades 3 and 4 is due to the difficulty in finding such patients with severe cases of PWMH, as mostly such patients are in critical states and cannot make it to the MRI machine for a scan.

Table 3.3: The Bayindir periventricular white matter hyperintensity rating scale

Grade	Name	Description
0	No lesion	Little or no periventricular lesions
1	Caps	Lesions located at the tips of the lateral ventricle posterior and anterior horns
2	Thin line	Resembles a rim, mostly located around the temporal horn of the lateral ventricle
3	Halo	Forming a halo around the lateral ventricle, more than a thin line
4	Irregular	Extending to the deep white matter

Figure 3.1: Exemplary FLAIR MR images of the five PWMH grades ordered from left-to-right in increasing lesion load as Grade 0, Grade 1, Grade 2, Grade 3, and Grade 4.

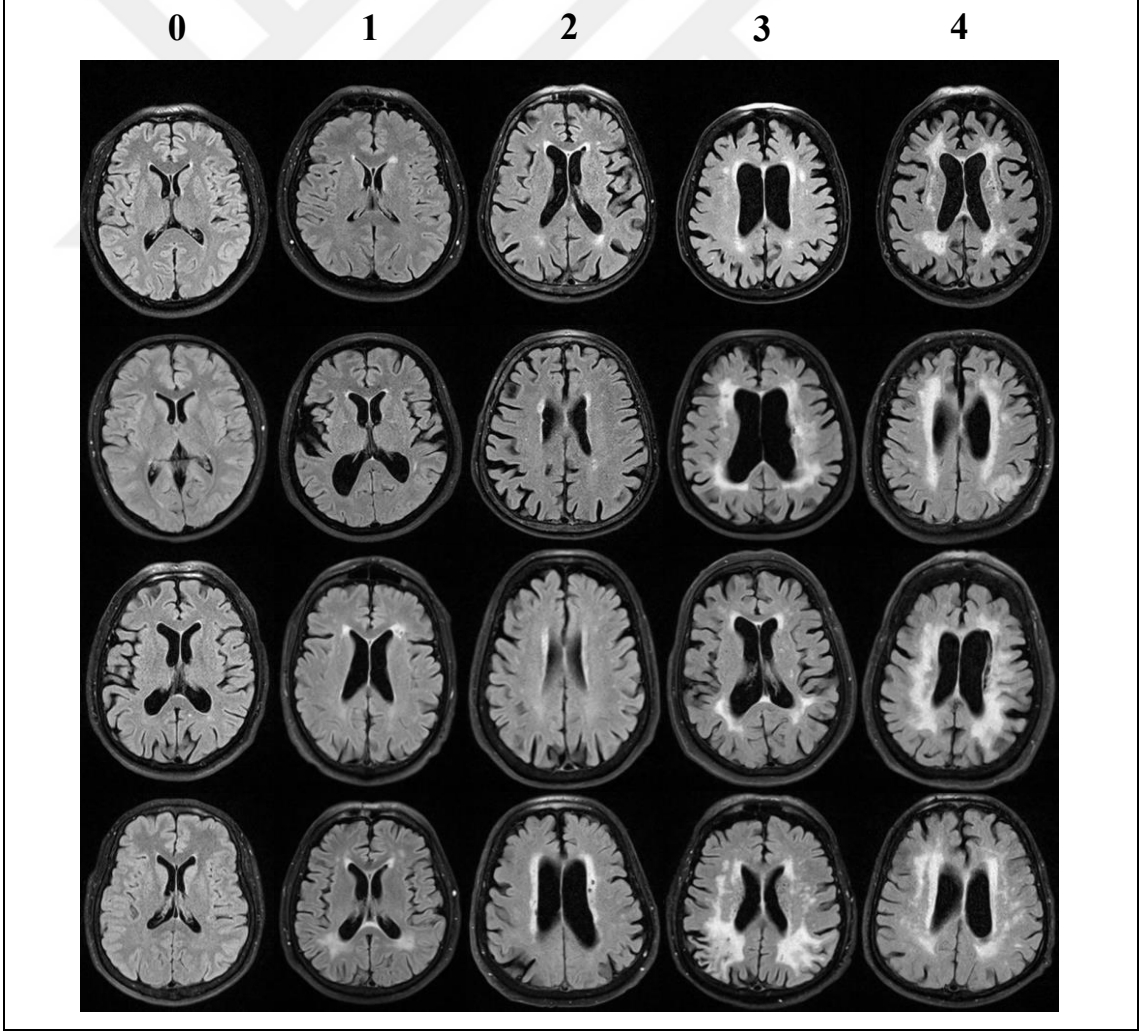
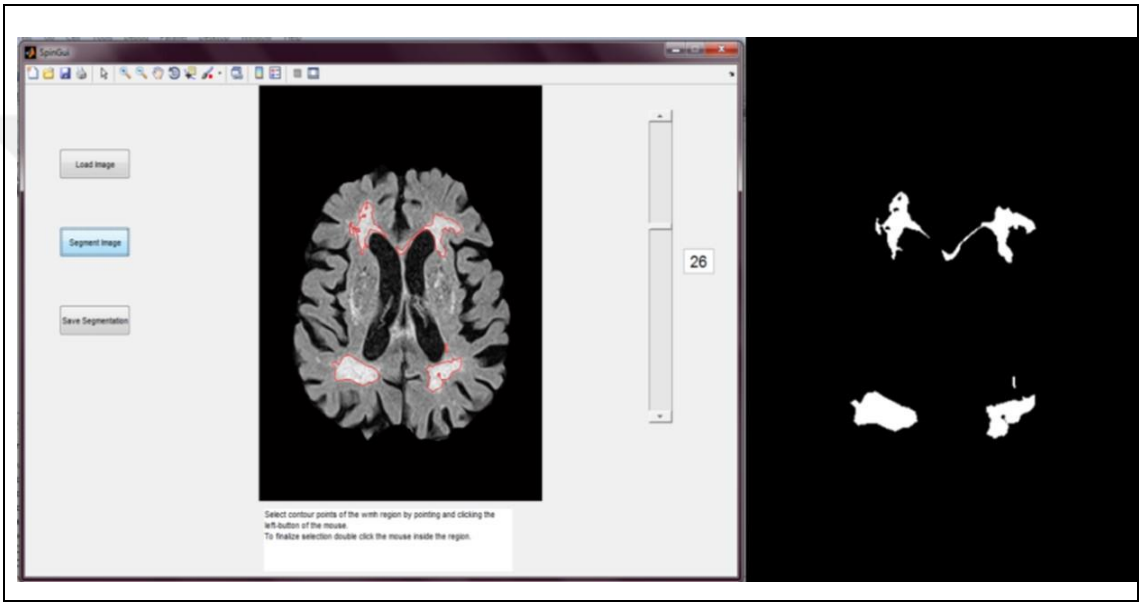


Table 3.4: Grade distribution of the subjects in the two datasets.

Grade	Bayindir	MS Challenge
0	14	4
1	14	12
2	13	1
3	7	4
4	4	0

Figure 3.2: Screen shot of segmentation software and the resulting binary mask.



3.2.4 The Rating Process

Our expert team consists of two neurologists and two radiologists. The four experts are responsible for rating patients' PWMH among other clinical related tasks. For every grade, the experts have agreed on a reference subject image that can be compared to future images. After the radiologists deliver the MRIs of patients, all four experts meet for the purpose of rating. For every subject, they individually visually compare the image to the reference images then collectively come to a consensus grade. The individual visual assessment involves a mental optimization process, namely, trying to maximize the similarity between the subject's PWMH and the reference PWMHs. On the other hand, the collective consensus is a combination of a majority vote as well as a strong presentation of a case should there be a decision that largely differs from the majority.

3.3 DATA PRE-PROCESSING

3.3.1 Data Collection

For the Bayindir dataset, the first pre-processing step was skull stripping which was achieved using BET v2.1 (Brain Extraction Tool), (Smith, 2002) provided by the Oxford Centre for Functional MRI of the Brain (FMRIB) followed by bias field correction using the N4ITK MRI Bias Correction algorithm in the 3DSlicer tool, (Jennings, Fennessy, Sonka, & Buatti, 2013).

Manual segmentations were meticulously executed using our in-house matlab-based software (Figure 3.2) and validated by our medical experts respectively. The manual delineation was performed on the FLAIR images of subjects and the results produced were binary masks of the WMLs.

Figure 3.3: ALVIN Lateral Ventricle Segmentation flowchart.

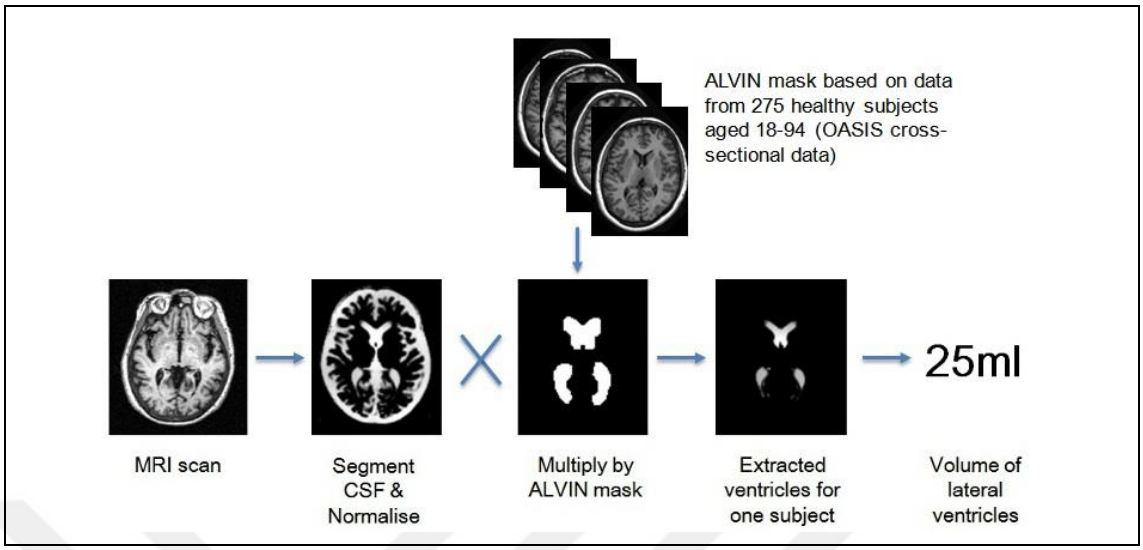
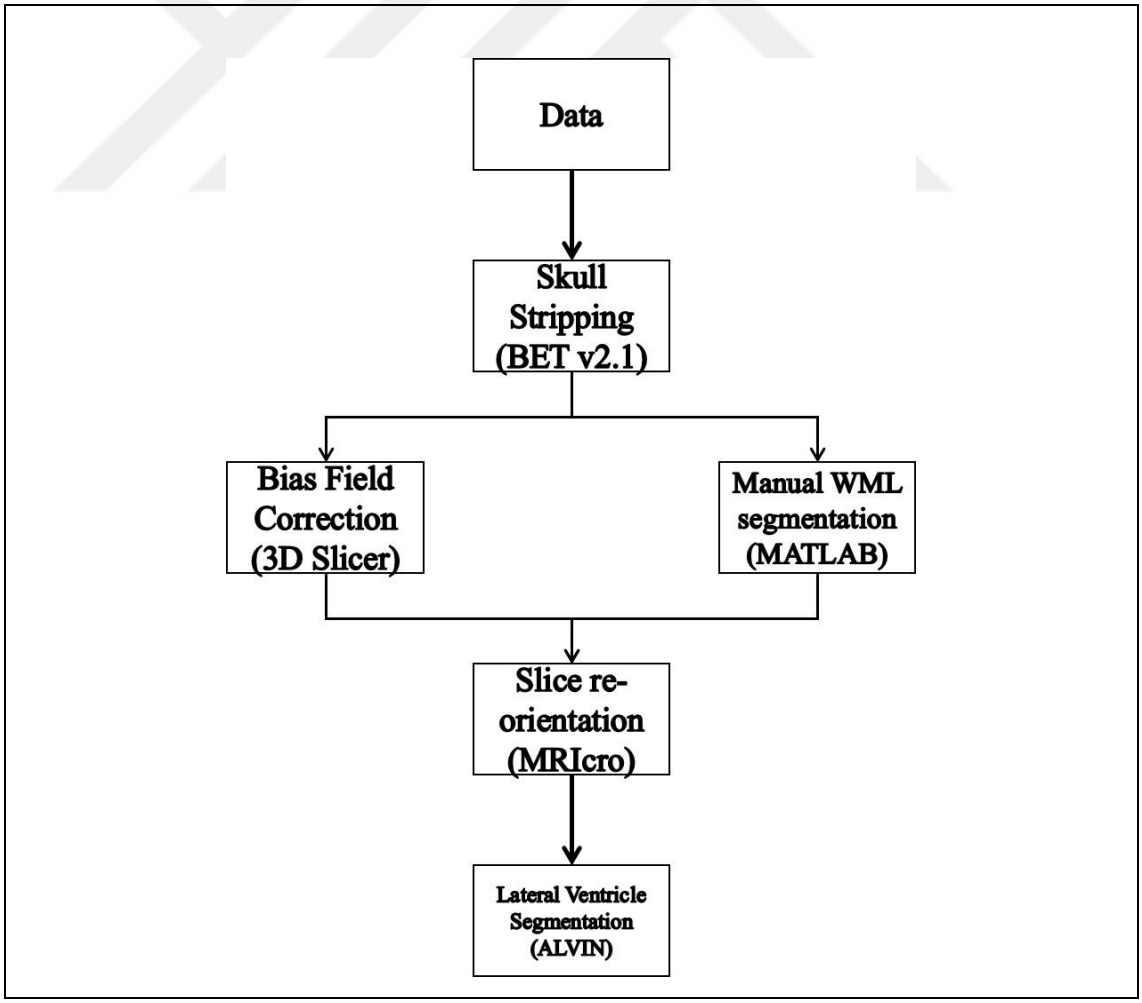


Figure 3.4: Data pre-processing flowchart.



3.3.2 Lateral Ventricle Segmentation with Alvin (Automatic Lateral Ventricle delIneatioN)

ALVIN is a fully automated algorithm that works within SPM8 to segment the lateral ventricles from structural MRI images. The algorithm has been tested with the Brain Segmentation Testing Protocol (Kempton, et al., 2011) and is valid for use in patients with Alzheimer's disease. ALVIN was designed to process T1- and T2-weighted images however, the authors outlined that the results obtained from processing FLAIR MRIs will require some post-processing.

The flow chart in Figure 3.4 shows a summary of all the pre-processing steps taken. After skull stripping, either bias field correction or manual segmentation could be done and the slice re-orientation was necessary because some of our images did not have the same orientation as the SPM templates. The details of the pre-processing pipeline for the MS Challenge dataset can be found in (Carass A. , et al., 2017).

4. AUTOMATIC PERIVENTRICULAR WHITE MATTER HYPERINTENSITY SEGMENTATION

4.1 INTRODUCTION

In this chapter, the methodology as well as the results of periventricular white matter hyperintensity segmentation are presented. Three segmentation algorithms were used in order to demonstrate the superiority of CRF over the others. The three algorithms used in this work are listed below.

i Thresholding (THR)

ii Support Vector Machines (SVM)

iii Conditional Random Field (CRF)

To minimize false positives, before and after segmentation, inter-hemispheric fissure (IHF) detection and 3D connectivity analysis, respectively, were integrated into each algorithm. In addition to these, it was necessary to segment the lateral ventricle since we are interested in the periventricular WMH (PWMH). A summary of the process of segmentation is depicted in Figure 4.1. The final result of segmentation serves as input to the periventricular white matter hyperintensity signature which is discussed in Chapter 5.

4.2 THRESHOLDING & 3D CONNECTIVITY ANALYSIS

Part of the work presented in this section has been reviewed and published in (Carass A. , et al., 2017) and (Iheme, et al., 2012). We work on the bias field corrected and skull-stripped FLAIR images of the subjects to compute a 256-bin histogram of the brain parenchyma. The peak of the histogram is then identified in order to compute the 7 dB-drop value (I_{7dB}) to the right of the peak. The peak is assumed to be that of a normal distribution so that its 7 dB drop value is more than twice its Full Width at Half Maximum (FWHM) value. I_{7dB} was empirically set, (Iheme, et al., 2013) its intensity value is guaranteed to be amongst the highest of the image as shown in Figure 4.2. With this value as a minimum threshold for WMH, we define the threshold as:

$$T = I_{peak}(1 - W) + I_{7dB}W \quad \textbf{(4.1)}$$

where I_{peak} is the intensity value at the peak of the histogram and W is a parameter which can be adjusted in order to obtain an optimum result. It was empirically set at 3.5

Figure 4.1: Flowchart showing the progression of segmentation and automatic grading of periventricular white matter hyperintensity

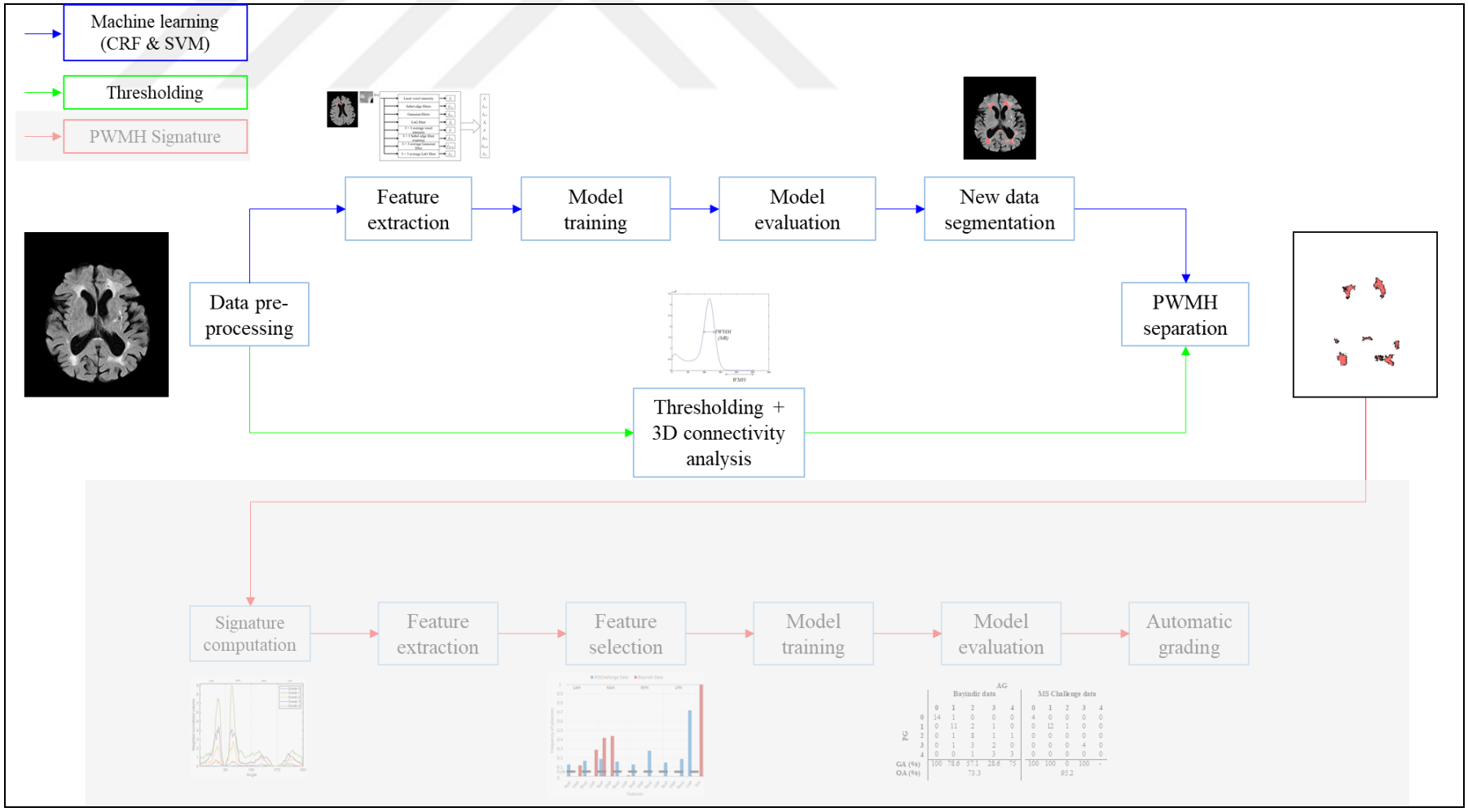
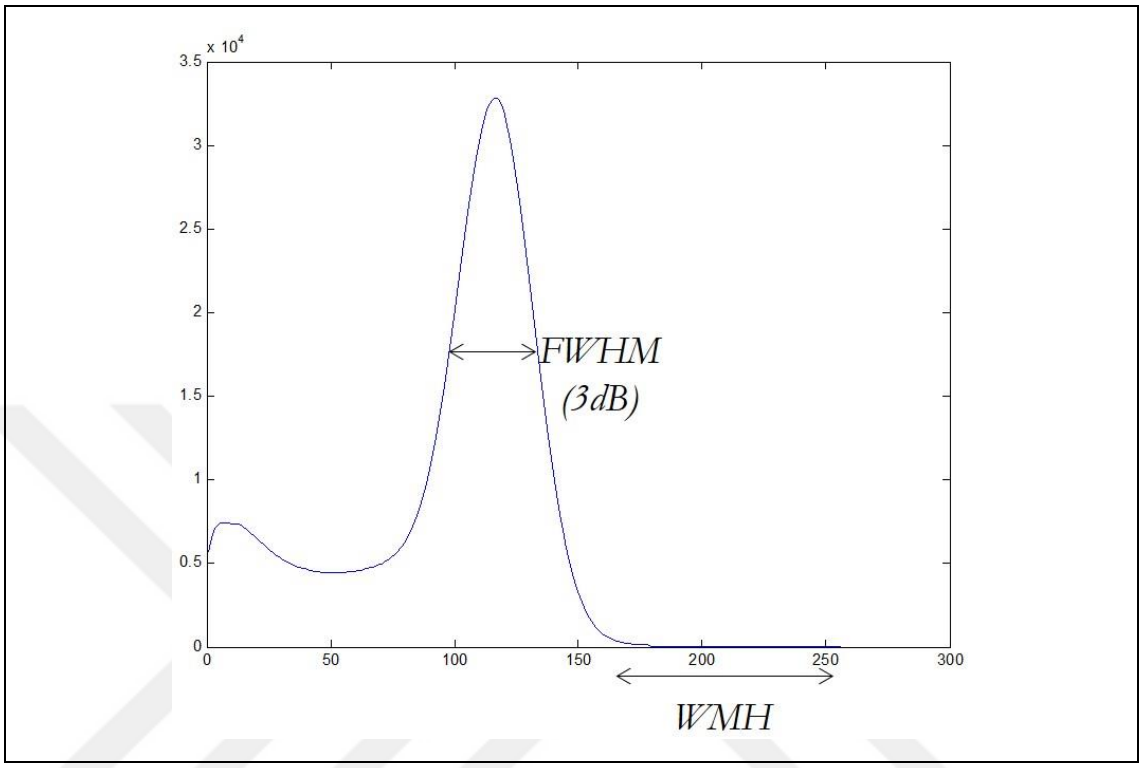


Figure 4.2: Example (smoothed) histogram of a subject's brain parenchyma showing the FWHM and WMH region

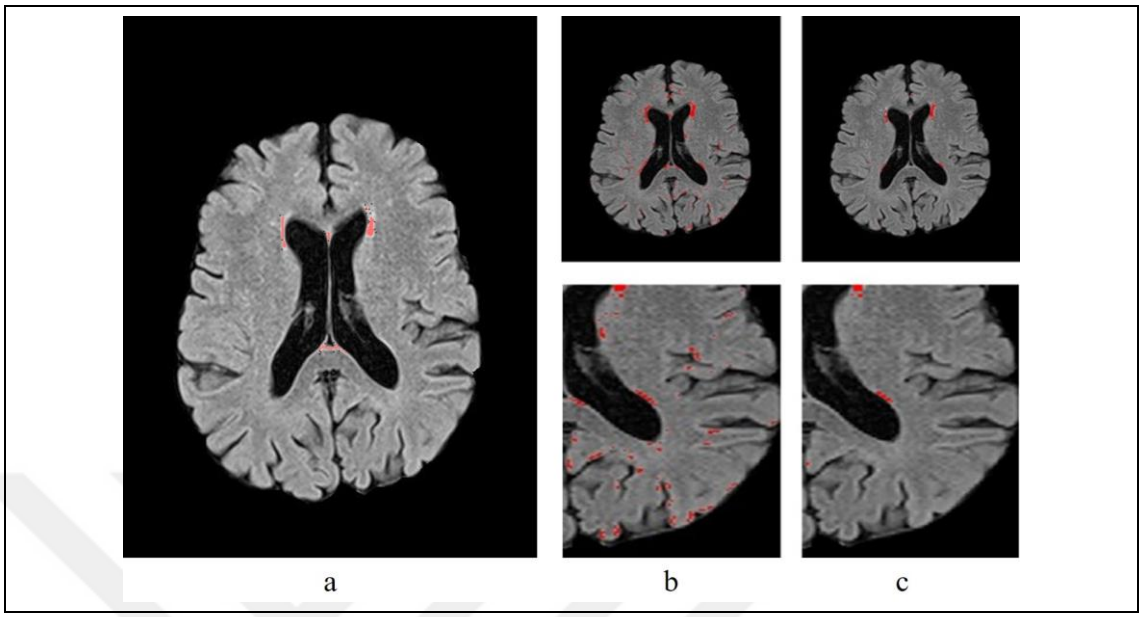


4.2.1 3D Connectivity Analysis

As seen in Figure 4.3a, thresholding alone is not sufficient to accurately segment WMLs since it does not take the 3 dimensional nature of the images into account. Furthermore, anatomical and contextual information are also not considered. Since voxels at the cerebrospinal fluid-brain interfaces also appear hyperintense on FLAIR images, even though they are not members of WMLs, the thresholding-based segmentation algorithm detects them.

To minimize the false positives (FP) resulting from thresholding, a three dimensional connectivity analysis was carried out wherein every voxel detected as hyperintense was examined for its 3D connectivity. For every such voxel, its twenty-six neighbors (eight 2D neighbors and eighteen others on adjacent slices) are examined. If a set minimum

Figure 4.3: (a) Ground truth segmentation (b) Result of intensity thresholding and (c) Application of the 3D connectivity rule.



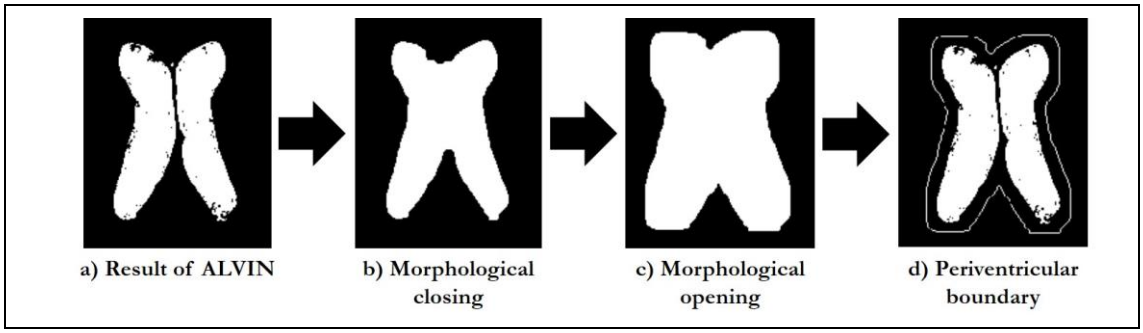
number of neighboring voxels (lesion volume) are also hyperintense in this 3D neighborhood, then that voxel is labeled hyperintense else, it is labeled otherwise. Figure 4.3c shows an improvement in the segmentation result after the 3D connectivity analysis, it is more visually similar to the ground truth shown in Figure 4.3a.

4.3 PERIVENTRICULAR WHITE MATTER HYPERINTENSITY SEGMENTATION

We have achieved PWMH segmentation by extending the perimeter of the lateral ventricle by approximately 4 mm (Nyquist, et al., 2015) so that every hyperintense voxel that falls within or on the perimeter is classified as a periventricular WML. Specifically, we work on the segmented lateral ventricle image (binary mask).

At first, we perform morphological closing to erode the segmented lateral ventricle of any small *holes* that were left over after segmentation. Morphological opening follows the closing operation. This operation extends the perimeter of the lateral ventricle so that the new perimeter corresponds to the periventricular boundary of the lateral ventricle as seen in Figure 4.4. The sizes of the structuring elements used in the morphological operations are adaptive and are chosen based on the dimensions of the image voxels and the results of lateral ventricle segmentation. Specifically, the closing structuring element is disk-shaped and has a radius twice the length of the largest hole in the image. The opening

Figure 4.4: (a) Result of lateral ventricle segmentation (ALVIN) (b) Result of morphological closing (c) Result of morphological dilation (d) Periventricular region of the lateral ventricle



structuring element is square-shaped; its size is adaptive and dependent on the voxel size of the image. However, its size, which is based on the voxel dimensions, is chosen so that the lateral ventricle mask is dilated to approximately 4 mm. The 3D connectivity analysis is applied to ensure connectivity across adjacent slices.

In Figure 4.5, the WMLs labelled in red are the PWMH; the red solid line indicates the periventricular region and the green labelled WMLs represent the "other" WMLs.

4.3.1 Inter-Hemispheric Fissure Detection

The need to detect the IHF was brought to light when we observed that the MRIs of some subjects where slightly rotated. A reference vertical axis was required for keeping a fairly constant angular location of the lateral ventricle horns with respect to the geometric center of the brain across all slices and subjects. In addition to that, detecting the IHF would also facilitate the minimization of false positives, especially those located on the corpus callosum.

Our method is based on the intuition that the IHF, which is filled with CSF, will appear as dark, low intensity or background voxels on FLAIR images. Accordingly, it takes root in the RANdom SAmple Consensus (RANSAC) algorithm (Fischler & Bolles, 1981) and is adapted from (Ekin, 2006). Briefly, we begin by cropping the axial image in order to get rid of any background border voxels. Candidate IHF points are then collected by scanning all the rows of the cropped image and selecting the voxels with the least intensity value (per row). The candidate points are then fed into the RANSAC algorithm as hypothetical in-liers of a straight-line model. As expected, most of the candidate points

were aligned on the IHF; the estimated model parameters then ensure that the detected line passes through the IHF.

Figure 4.5: Periventricular white matter hyperintensity (PWMH) segmentation. The solid red line indicates the periventricular region. The red lesions are PWMH and the green lesions are "other" WMLs

Figure 4.6 summarizes the process of automatic IHF detection. Voxels falling on or around the IHF line, such as those on the corpus callosum, are unmarked as hyperintense as need be. This step was based on our experts' instructions.

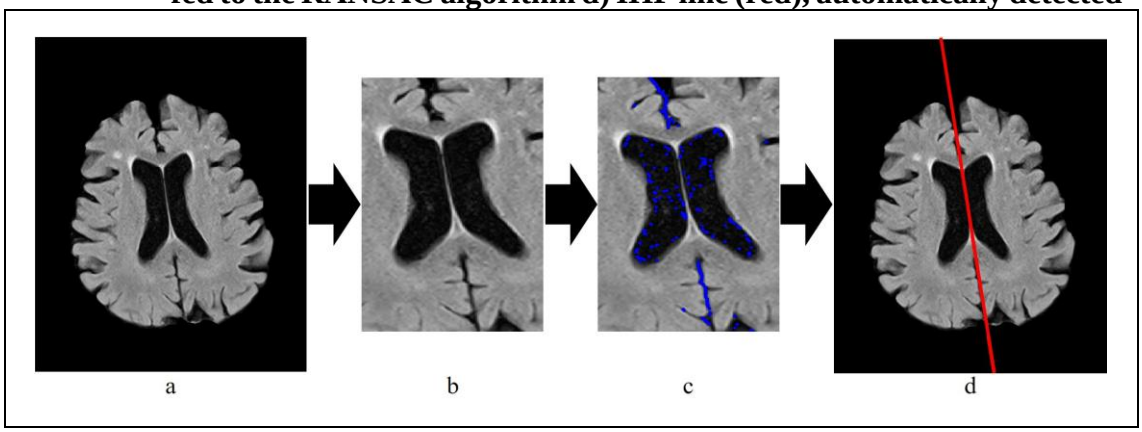
4.4 MACHINE LEARNING BASED PERIVENTRICULAR WHITE MATTER HYPERINTENSITY SEGMENTATION

To achieve more accurate segmentation results we turn to machine learning for its robustness and efficiency. However, there are prerequisites to applying machine learning and among the most important are feature extraction and feature selection. Accordingly, in this section we present the steps taken as well as the methods applied in order to take advantage of the powers of machine learning algorithms for WML segmentation.

4.4.1 Periventricular White Matter Hyperintensity Feature Extraction

Before feature extraction, to maintain homogeneous filter responses across all subjects, it was necessary to re-size each image to a chosen dimension and scale the intensity range of each image to a common set range. The features that were extracted were texture as well as intensity based. While the features were voxel-centered, neighborhood information was also accounted for.

Figure 4.6: Inter-hemispheric fissure (IHF) detection. a) Original FLAIR image b) Cropped image c) The blue dots are candidate IHF points which are fed to the RANSAC algorithm d) IHF line (red), automatically detected



Precisely, we computed edge (vertical and horizontal), Gaussian (high- and low- pass), as well as Laplacian of Gaussian (LoG) filter responses of each voxel; the raw intensity value was also appended as a feature. In addition, we included the in-plane 3×3 mean of each previously extracted feature. In total, 12 features were extracted for each voxel as depicted in Figure 4.7.

4.4.2 Conditional Random Fields (CRF)

CRFs are advantageous to structured segmentation because unlike ordinary classifiers, they are able to predict labels with regard to their "neighboring" samples. Specifically, a conditional probability distribution over label sequences **y** is defined given a particular observation sequence **x**, rather than a joint distribution over both label and observation sequences. Unlike other Markov models, CRF is a discriminative model that relaxes the independence assumptions in order to ensure a tractable inference. Concretely, given an observation sequence **x** and a label sequence **y**, a CRF is formulated as follows:

$$p(\mathbf{y}|\mathbf{x},\lambda) = \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{j}\lambda_j F_j(\mathbf{y},\mathbf{x})\right) \quad \textbf{(4.2)}$$

where

$$F_j(y,x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)$$

is a set of feature functions of the entire observation sequence and the labels at positions i and $i-1$ in the label sequence. λ_j are the parameters that are to be estimated from the training data; and $Z(\mathbf{x})$ is a normalization factor.

Each $F_j(y, x)$ is either a state feature function $s(y_{i-1}, x, i)$ or a transition feature function $t(y_{i-1}, y_i, x, i)$. The set of feature functions is formed by each function taking on the value of one of the features if the current state or previous and current states take on particular values.

4.4.2.1 CRF Training

The objective of training is to estimate the parameters of the CRF model. Since CRFs define the conditional probability $p(\mathbf{y}|\mathbf{x})$, the objective for parameter learning is to maximize the conditional likelihood of the training data

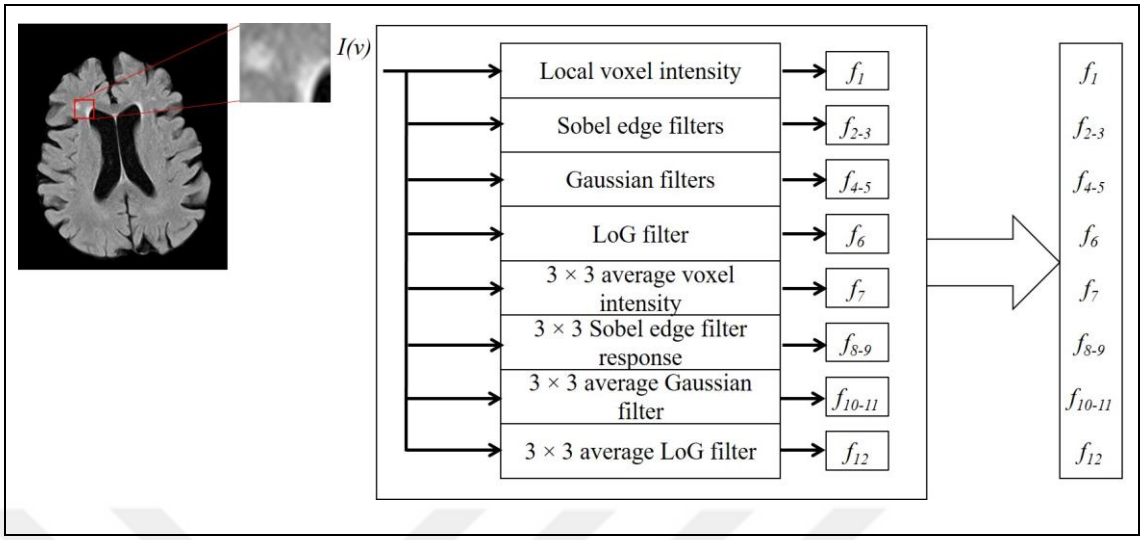
$$\sum_{j=1}^{m} \log p(\mathbf{y}^{(j)}|\mathbf{x}^{(j)}) \quad \textbf{(4.3)}$$

If we impose a Gaussian prior on the λ's to regularize training, the objective function becomes

$$\sum_{j=1}^{m} \log p(\mathbf{y}^{(j)}|\mathbf{x}^{(j)}) - \sum_{i}^{F} \frac{\lambda_i^2}{2\sigma^2} \quad \textbf{(4.4)}$$

for $\lambda \sim N(0, \sigma^2)$. The gradient of the objective function can be relatively easily computed and used in an optimization algorithm such as L-BFGS. It can be computed by finding the partial determinant of the objective function with respect to λ_k.

Figure 4.7: Feature extraction for machine learning-based periventricular white



The result is of the form:

$$\sum_{j=1}^{m}\sum_{i} f_k(y_{i-1}^{(j)}, y_i^{(j)}, \mathbf{x}^{(j)}, i) - \sum_{j=1}^{m}\sum_{i}\sum_{y,y'} E_{y_{i-1}',y_i'}[f_k(y_{i-1}', y_i', \mathbf{x}^{(j)}, i)] - \frac{\lambda_k}{\sigma^2} \qquad \textbf{(4.5)}$$

The first term of Equation **(4.5)** is the expected value of f_k and the second term which arises from the derivative of $Z(x)$ is the expectation of f_k. This means that for the unregularized maximum likelihood solution, when the gradient is zero, the two expectations are equal. There are far more normal tissue than there are WMLs in our training data. To prevent training a model that would suffer from overfitting, we balanced the data by selecting an equal number of samples from the positive (WML) and negative (normal brain tissue) classes.

CRF inference can be performed efficiently and exactly by slight variants of the standard dynamic-programing algorithms for HMMs. We have made use of the HCRF library (Hugues & Yale, 2015) implementation of CRF in our study.

The final result of CRF is a probability map. Samples of CRF segmented images are shown in Figure 4.8.

Figure 4.8: Results of CRF segmentation (probability maps) overlaid on the original image.

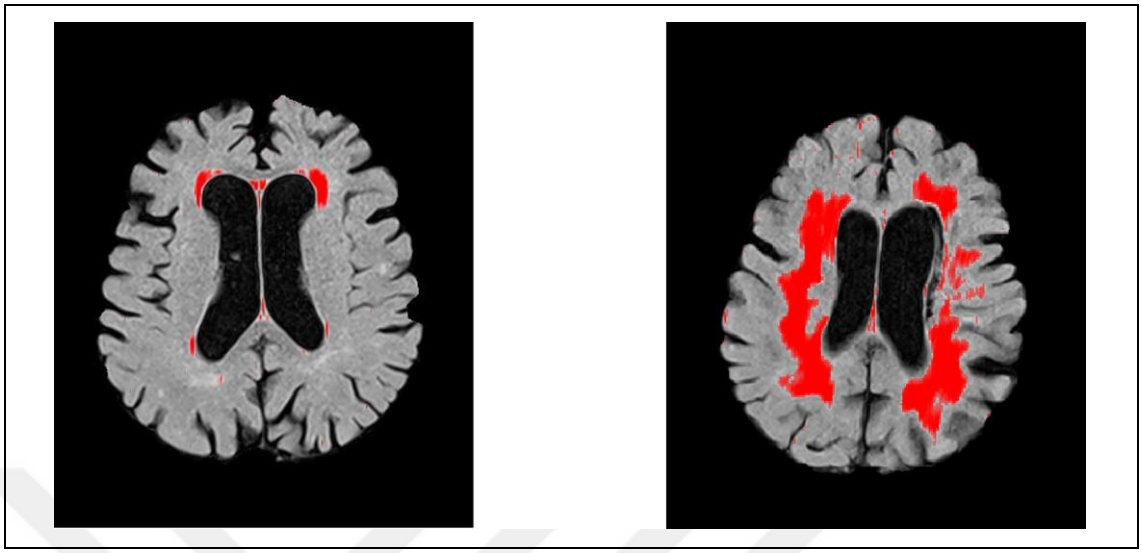
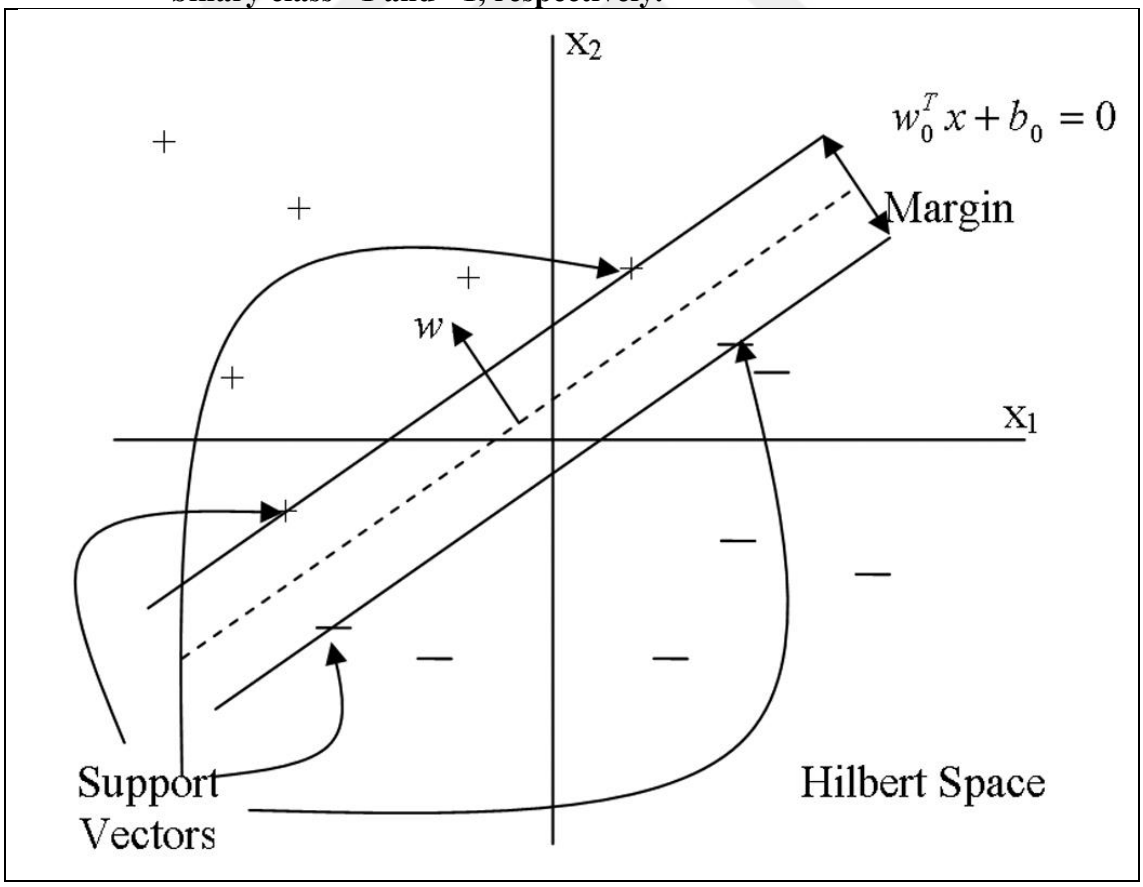


Figure 4.9: An example of two-class (+ and −) problem showing optimal separating hyperplane (dotted line) that SVM uses to divide two groups' data, and the associated Support Vectors. Data shown by '+' and '−' represent binary class +1 and −1, respectively.



Source: (Lao, et al., 2008)

4.4.3 Support Vector Machines (SVM)

It is no suprize that SVM has become a household name in the machine learning sphere given its advantageous features such as its ability to effectively solve binary classification problems.

SVM originate from Vapnik's statistical learning theory (Vapnik, Statistical Learning Theory, 1998),where the learning problem is formulated as a quadratic optimization problem whose error surface is free of local minima and has global optimum. In a binary classification task like the one in our study (normal tissue/lesion tissue), the aim is to find an optimal separating hyperplane (OSH) between the two datasets. Figure 4.9 illustrates a two-class problem with a hyperplane separating the two groups. SVM finds the OSH by maximizing the margin (minimum distance) between the classes. The main concepts of SVM are to first transform input data into a higher dimensional space (Hilbert Space) by means of a kernel function, and then construct an OSH between the two classes in the transformed space (Hilbert Space). Those data vectors nearest to the constructed hyperplane in the transformed space are called the support vectors (Figure 4.9) that contain valuable information regarding the OSH. SVM are an approximate implementation of the method of "structural risk minimization" aiming to attain low probability of generalization error. Briefly, the theory of SVM can be referenced in (Vapnik, 2000).

The kernel function used in our application is the Gaussian radial basis function kernel, defined as

$$K(x, y) = exp\left(\frac{\|x - y\|^2}{2\alpha^2}\right) \quad \textbf{(4.6)}$$

Where *x* and *y* are two feature vectors, and α controls the size of the Gaussian kernel.

The fitness of a hyperplane in feature space is usually measured by the distance between the hyperplane and those training points lying closest to it (the support vectors). A consequence of this is that we can completely specify our decision surface in terms of these support vectors. An overview of SVM pattern recognition techniques may be found in (Burges, 1998).

In this study, we employ the LiBSVM package of (Chang & Lin, 2011) with a (Gaussian) radial basis function kernel; the parameters C and γ were optimized via a five-fold cross validation accuracy as the criterion. Using the optimized parameters, a model was trained to predict future, unseen samples. The same features highlighted in Section 4.4.1 above were used to train the SVM model.

4.5 RESULTS

Our first objective was to segment white matter lesions using conditional random fields but for the sake of comparison and to demonstrate the superiority of CRF over SVM and thresholding, we present the segmentation results from the three algorithms. Figure 4.10 and
Figure 4.11 show the average dice scores and respective standard deviations obtained from the segmentation of the Bayindir and the MS Challenge datasets.

To investigate the performance of the respective algorithms based on how much PWMH load was present, we computed the mean dice scores and standard deviations per grade in each dataset. These results are shown in Figure 4.10 and
Figure 4.11. The dice score is a similarity metric that varies between 0 (no similarity) and 1 (full similarity) and it can be computed from Equation (4.7)

$$D = \frac{2|A \cap M|}{|A| + |M|} \tag{4.7}$$

Where A and M are the results of automatic and manual segmentations respectively. A dice score of 0.7 or more is considered to be a good performance in the literature (Maria, et al., 2015)

4.6 DISCUSSION

From Figure 4.10 and
Figure 4.11, we can see that the CRF segmentation algorithm clearly outperforms the thresholding- and SVM-based algorithms. For both datasets, there is not a significant difference in performance of the CRF algorithm but the converse is the case for SVM and thresholding-based algorithms. Precisely, there is a 2.22 percent, 8.33 percent and a 27.85 percent difference between the CRF, SVM and thresholding-based results respectively.

To analyze the performance of the algorithms on both datasets further, a pairwise percentage difference matrix was computed from the mean dice scores. The matrix revealed a similar trend in both datasets. However, the percentage improvement (between

CRF and the other algorithms) in the MS Challenge dataset was higher than that in the Bayindir dataset as depicted in Figure 4.12.

For further insight, refer to Figure 4.14 where sample visual results of the segmentation algorithms are depicted alongside the ground truth segmentations and corresponding overall dice scores.

Figure 4.10: Bayindir dataset: mean dice scores and standard deviations of 0.1, 0.14 and 0.18 obtained from CRF, SVM and THR algorithms respectively.

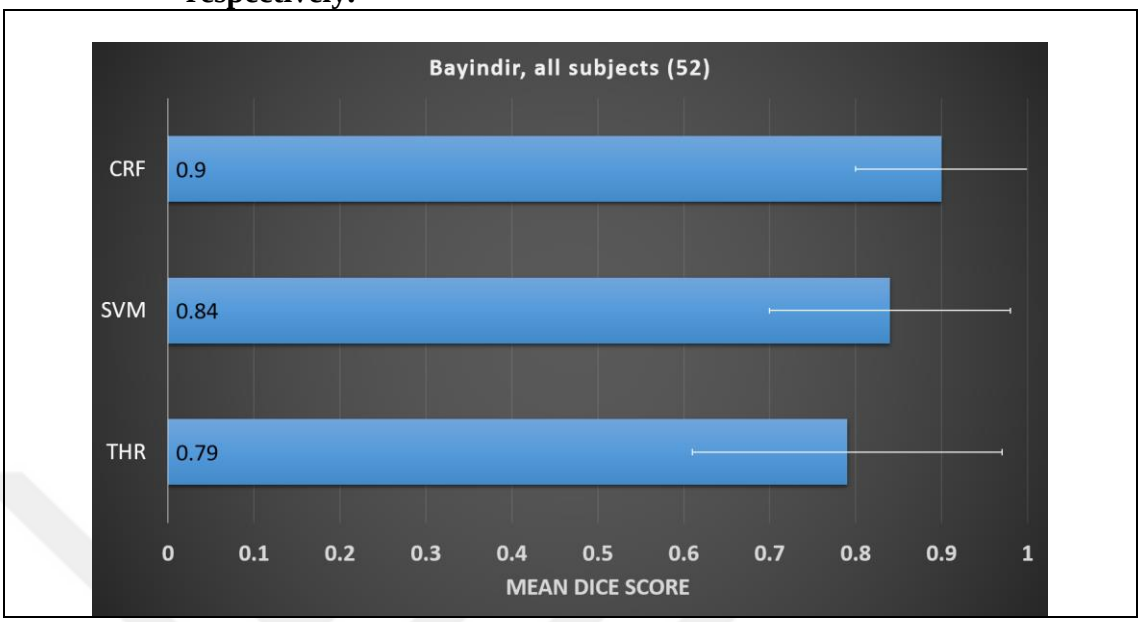


Figure 4.11: MS Challenge dataset: mean dice scores and standard deviations of 0.06, 0.1 and 0.24 obtained from CRF, SVM and THR algorithms respectively.

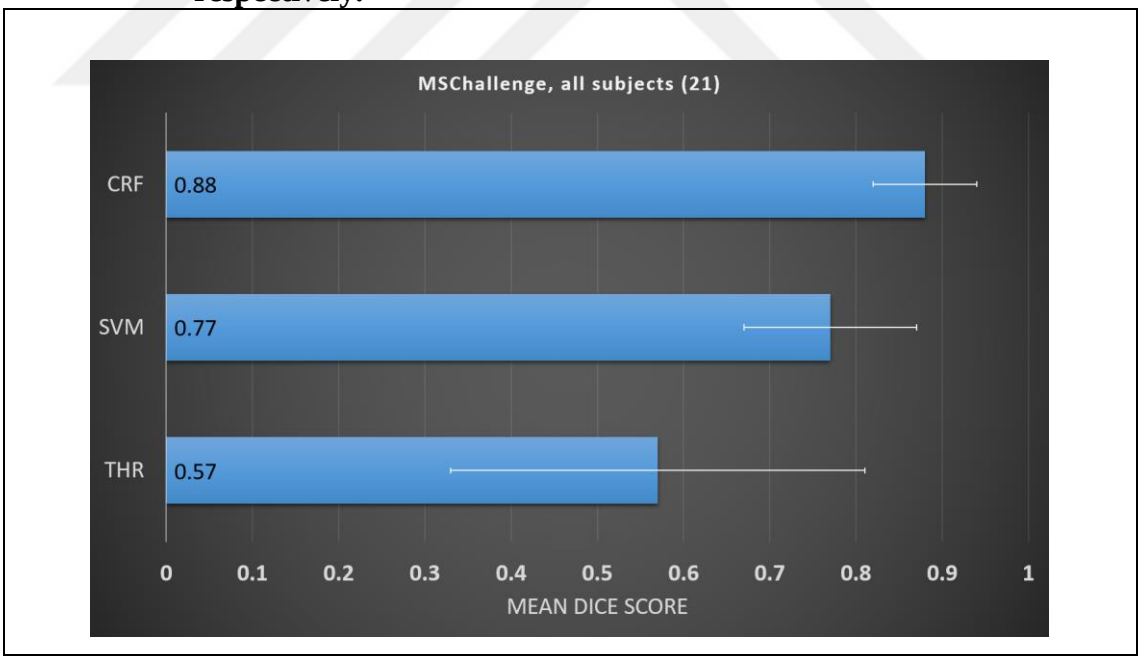
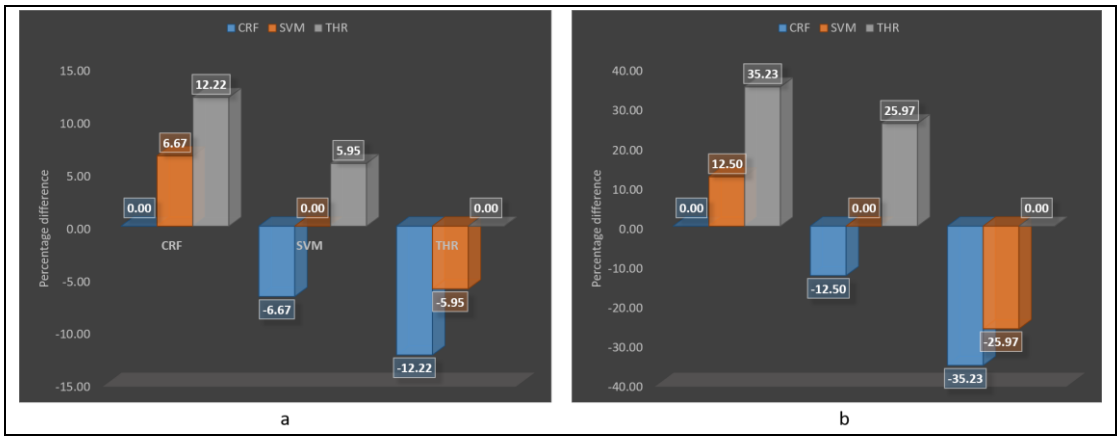


Table 4.1: Mean Dice score and standard deviation of each grade obtained for CRF, SVM and THR

Algorithm	Bayindir Dice(SD)	MS Challenge Dice(SD)
	Grade 0	
CRF	0.9(0.10)	0.88(0.06)
SVM	0.84(0.14)	0.77(0.10)
THR	0.79(0.18)	1.00(0.00)
	Grade 1	
CRF	0.85(0.11)	0.87(0.05)
SVM	0.76(0.10)	0.74(0.06)
THR	0.77(0.16)	0.44(0.14)
	Grade 2	
CRF	0.88(0.07)	0.90(0.00)
SVM	0.84(0.09)	0.86(0.00)
THR	0.80(0.22)	0.44(0.64)
	Grade 3	
CRF	0.98(0.01)	0.95(0.02)
SVM	0.97(0.01)	0.94(0.03)
THR	0.85(0.04)	0.52(0.03)
	Grade 4	
CRF	0.99(0.01)	–
SVM	0.98(0.01)	–
THR	0.88(0.05)	–

Figure 4.12: Pairwise percentage change in algorithm performance. (a) Bayindir dataset (b) MS Challenge dataset



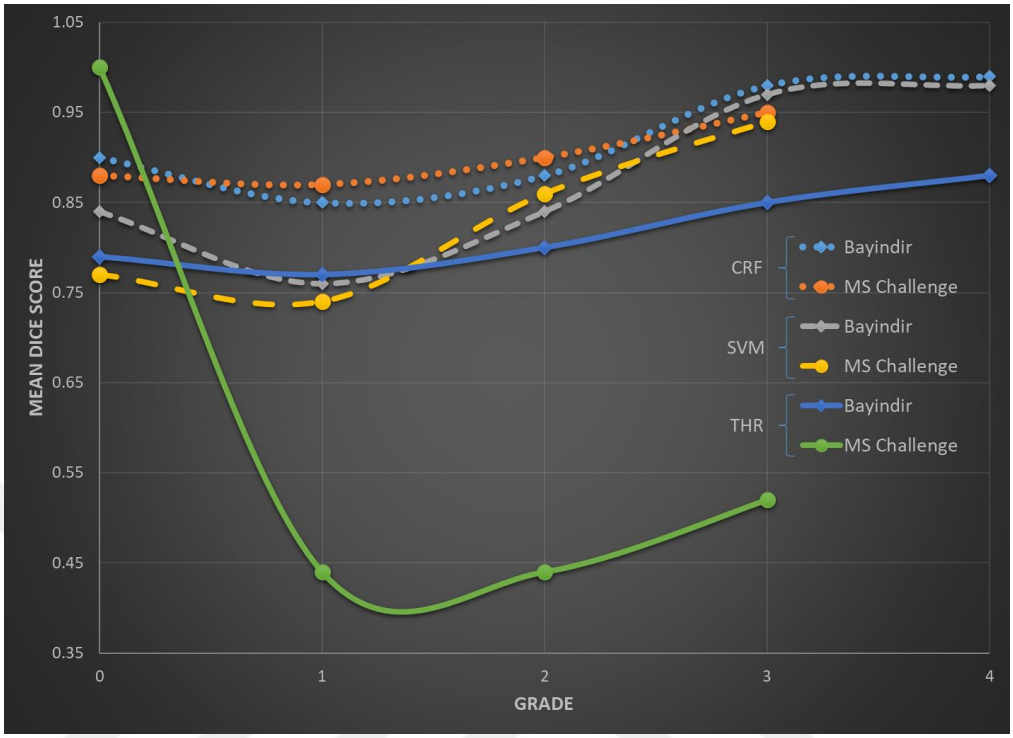
An in-depth look at how the CRF algorithm performs is shown in Table 4.1. We see that CRF outperforms SVM and the thresholding-based algorithms for all the grades. All three algorithms follow a similar trend, having the best performance each in the higher grades (3 & 4), as can be seen in Figure 4.13. Notice that CRF and SVM, in general show significant improvements over the thresholding algorithm in both datasets. However, the results from SVM and CRF are similar for grade 3 and 4 patients. This could be because with larger lesions, the need for neighborhood information becomes less apparent. For the lower grades, on the other hand, the significance of CRF performance is more since the need for neighborhood information with smaller lesions increases. There were no grade 4 patients in the MS Challenge dataset to make an algorithm performance comparison.

4.7 CONCLUSION

In this chapter, we have explained the procedure as well as shown the results of automatic periventricular white matter hyperintensity segmentation using three algorithms, namely: Thresholding and 3D analysis, Support Vector Machines and Conditional Random Fields.

All three algorithms were tested on two separate datasets. We showed that CRF is a superior algorithm for our purpose by comparing the results of the three algorithms. The results, presented in terms of the Dice score, were visually as well as quantitatively assessed. The assessments consistently showed a trend of improvement in both datasets

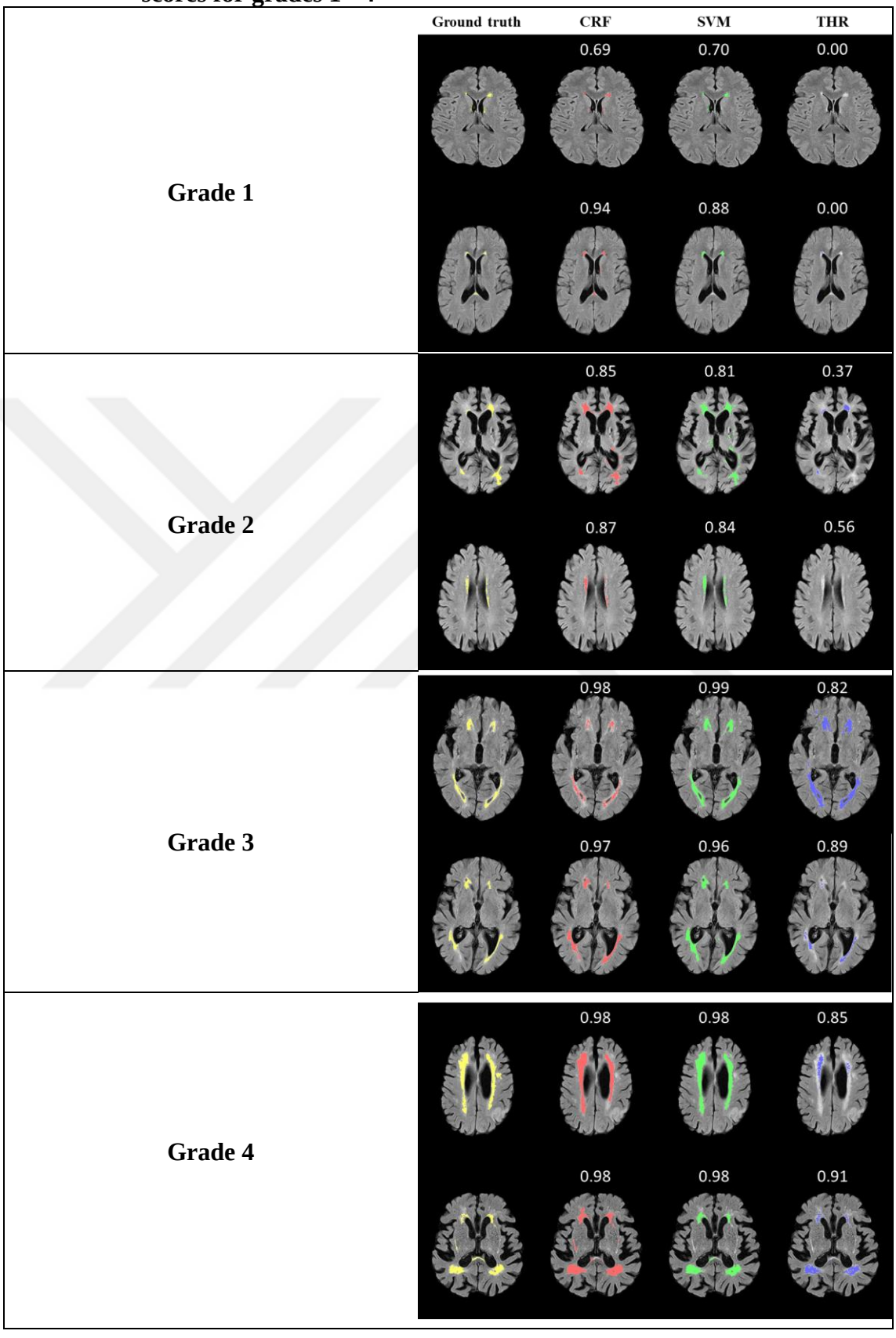
Figure 4.13: Mean Dice scores per grade for the CRF, SVM and THR algorithms on the Bayindir and MS Challenge datasets



from the thresholding-based algorithm to SVM and CRF as the most superior algorithm for periventricular white matter hyperintensity segmentation.

Furthermore, the fact that subjects with smaller lesions (grades 1 and 2) present greater challenges for visual as well as automatic segmentations was buttressed. Nevertheless, CRF significantly outperformed the other algorithms when the comparison was made. In conclusion, the lack of neighborhood dependencies in the thresholding- and SVM-based algorithms gives CRF the edge over them and this is demonstrated in the obtained results.

Figure 4.14: Sample visual PWMH segmentation results with corresponding dice scores for grades 1 - 4



5. AUTOMATIC PERIVENTRICULAR WHITE MATTER HYPERINTENSITY RATING

5.1 INTRODUCTION

In this chapter, we will discuss and outline the methods and details of our automatic periventricular white matter hyperintensity rating. Two schemes will be presented, namely:

i Lesion-based automatic PWMH rating &

ii Signature-based PWMH rating

Both methods take advantage of machine learning, specifically multi-class SVM, to achieve the set goals.

5.2 LESION-BASED PWMH RATING

The work detailed in this section was presented at the ESMRMB 2012 congress ((Iheme, et al., 2012)). Our cohort consisted of elderly patients with memory complaints (n=12) whose T2-weighted FLAIR MRI scans were acquired on a 1.5T scanner. Segmentation of supratentorial PWMHs was performed using the adaptive thresholding and 3D connectivity analysis described in Section 4.2. From the ground truth PWMH segmentations (realized manually by experts), we were able to supply descriptive features of the segmented lesions to a supervised machine learning (ML) model that could predict the grades of future PWMH lesions.

From the twelve subjects included in this experiment, 475 2D lesions ($grade1 = 171, grade2 = 176, grade3 = 88\ and\ grade4 = 40$) were examined. Shape as well as size descriptive features were extracted from the lesions to train the classifier. Precisely, Table 5.1 summarizes the features that were extracted.

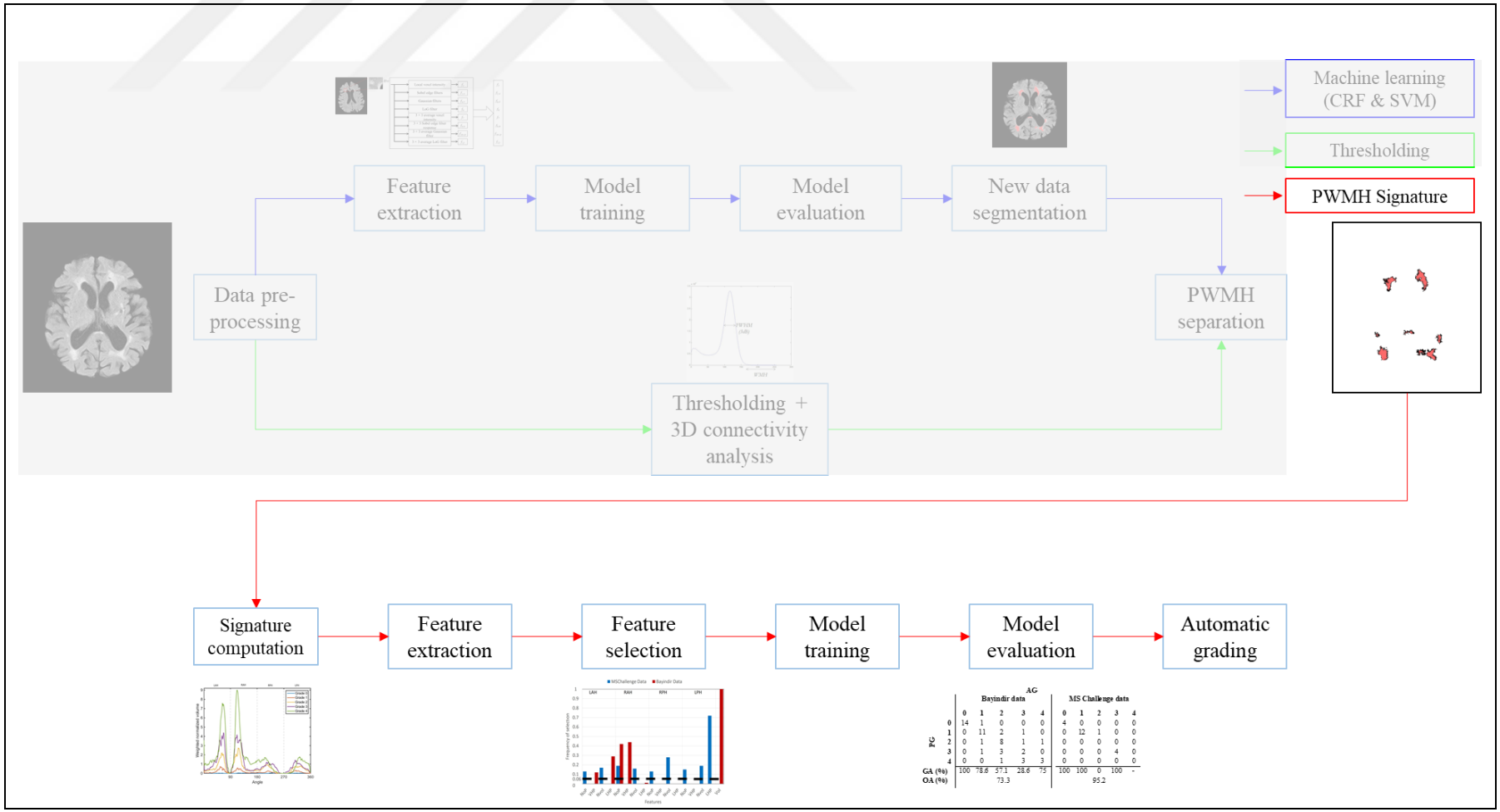
Table 5.1: Descriptive features for lesion-based automatic rating of PWMH

Feature	Sub-feature	Description
Area		Actual number of pixels in the lesion
Bounding box	Length	Smallest rectangle containing the lesion
	Width	
Equivalent diameter		Diameter of a circle with the same area as the region $\left[4 \times \frac{Area}{\pi}\right]$
Distance from centroid		The distance of the centroid of the lesion to the geometric centroid of the brain
Extent		Ratio of pixels in the lesion to pixels in the total bounding box,
Major axis length		Length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the lesion
Orientation		Angle between the x-axis and the major axis of the ellipse that has the same second-moments as the lesion
Eccentricity		Eccentricity of the ellipse that has the same second-moments as the region

5.2.1 Automatic Lesion Rating via SVM

After extracting and scaling the features, they were fed into a multi-class SVM classifier with an RBF kernel, were parameter optimization was performed. The criterion was the cross validation accuracy. To assess the performance of the model, we performed a 5-fold cross validation and obtained the results outlined in the confusion matrix of Table 5.2

Figure 5.1: Flowchart showing the progression of automatic grading of periventricular white matter hyperintensity and segmentation



5.3 SIGNATURE-BASED PWMH RATING

Table 5.2: Confusion matrix of automatic PWMH lesion rating.

		Predicted Grade				Accuracy (%)
		1	**2**	**3**	**4**	
Actual Grade	**1**	142	15	12	2	83
	2	34	133	9	0	75.6
	3	12	5	68	3	77.3
	4	5	0	3	32	80
Accuracy (%)		73.6	86.9	73.9	86.5	79

The PWMH signature (or simply the signature) is a visual as well as a statistical representation of the PWMH voxels present in a subject; it accounts for the volume and spatial location of PWM lesions. Specifically, the signature is a 360-bin histogram of PWMH where each bin represents the angle in degrees between the weighted geometric center of the brain and the location of each PWMH voxel. The lateral ventricle is divided vertically (left and right side) by the IHF and horizontally (anterior and posterior) by a line (perpendicular to the IHF line) that passes through the weighted geometric center of the brain. The two perpendicular lines form an axis with an origin at their intersection. The signature is then computed slice-by-slice by an anticlockwise scan of all PWMH voxels present on the slice in question. The result is a histogram with the angles (0-359) in degrees on the horizontal axis and the number of voxels on the vertical axis. Since the voxel dimension and the total WM volume is known, the vertical axis is converted to a normalized volume so as to be consistent with the varying WM volumes across all subjects. This computation has been summarized in a flowchart shown in Figure 5.1.

Figure 5.2 illustrates the construction of the PWMH signature. The vertical and horizontal axes represent the volume in milliliters and the angle in degrees respectively. Notice that the signature is divided into quadrants such that each quadrant corresponds to a half of the lateral ventricle horn starting from the left anterior (0-90) moving counter--clockwise to the left posterior (270-359). To compute the overall PWMH signature, we simply sum up all the signatures of the individual slices of the image volume. Since the IHF line

detection is performed for every slice that reveals the lateral ventricle, it ensures that the anatomical locations of connected lesions across adjacent slices are preserved.

Figure 5.2: Periventricular White Matter Hyperintensity Signature construction. LAH=Left Anterior Horn; RAH=Right Anterior Horn; RPH=Right Posterior Horn and LPH=Left Posterior Horn.

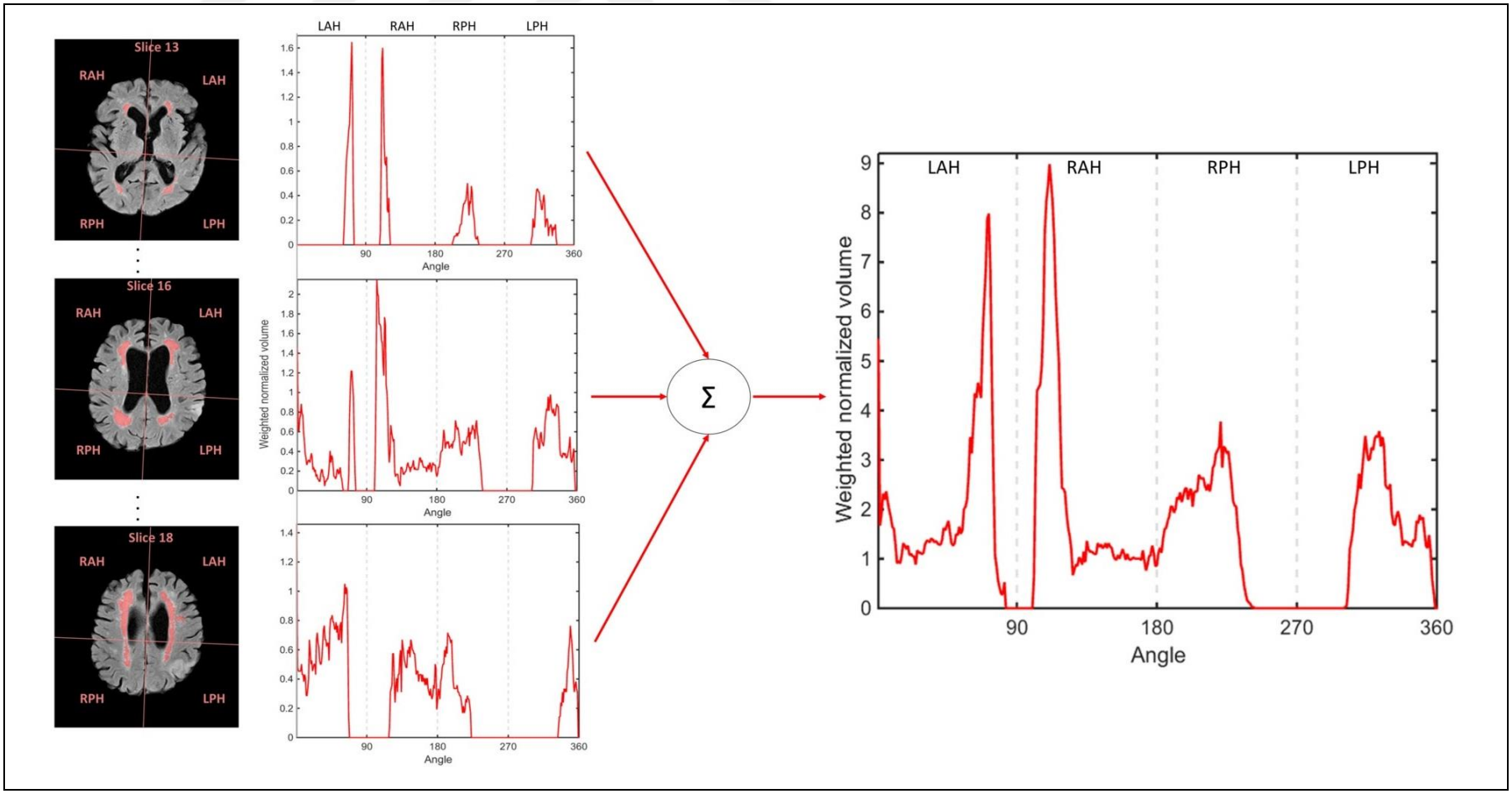
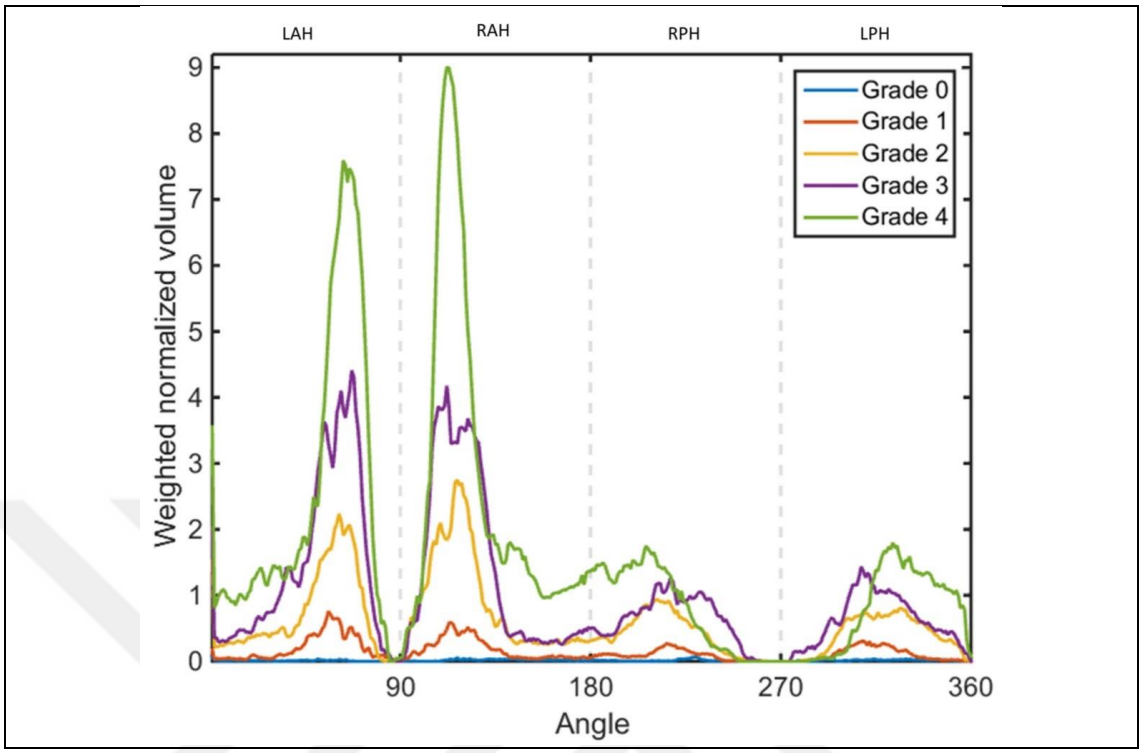


Figure 5.3: Average PWMH signatures of each grade.



5.3.1 PWMH Signature Analysis

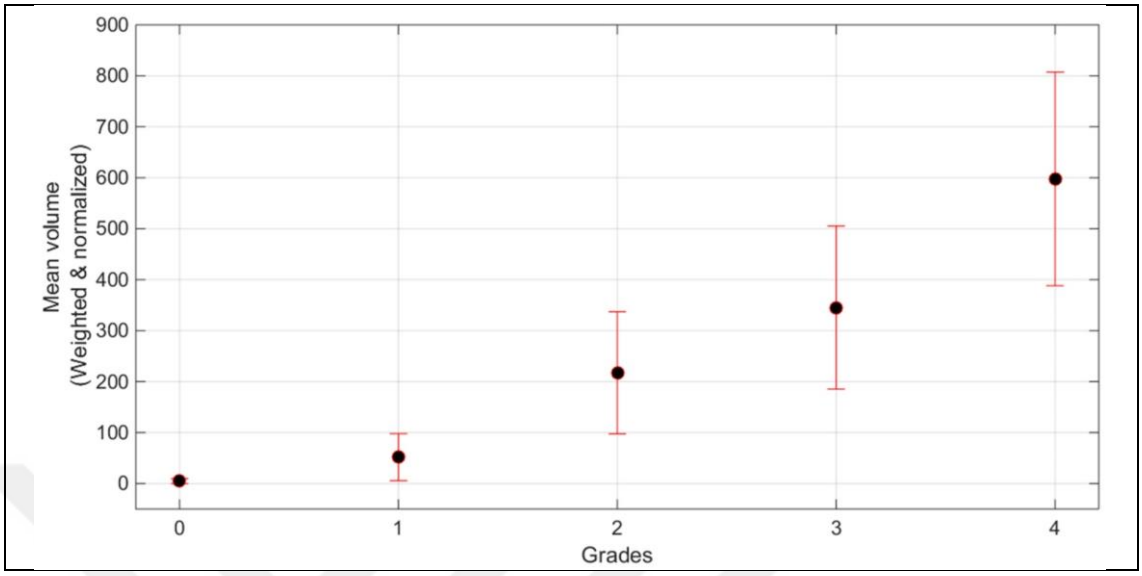
Our analyses were carried out on the Bayindir dataset of which all images are in the native space. In order to automatically grade a subject; we investigated the inter– and intra–grade similarities between the PWMH signature of patients. Though the signature of every subject is unique, we expected subjects of the same grade to exhibit a higher similarity when compared to signatures from subjects of other grades. To measure the degree of similarity between two given signatures, **p** and **q**, we compute the Euclidean distance in-between as

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i} (q_i - p_i)^2} \qquad \textbf{(5.1)}$$

where i runs from 0 to 359.

Figure 5.3 shows a plot of the average signatures of the subjects per grade (0 to 4) on the same axis. It is clear from Figure 5.3 that there are more discrepancies between the lower grades (grades 1 and 2) especially in the posterior region than in the anterior region. It would be

Figure 5.4: Summation of average PWMH signatures showing the standard deviations and the overlap that exists between grades.



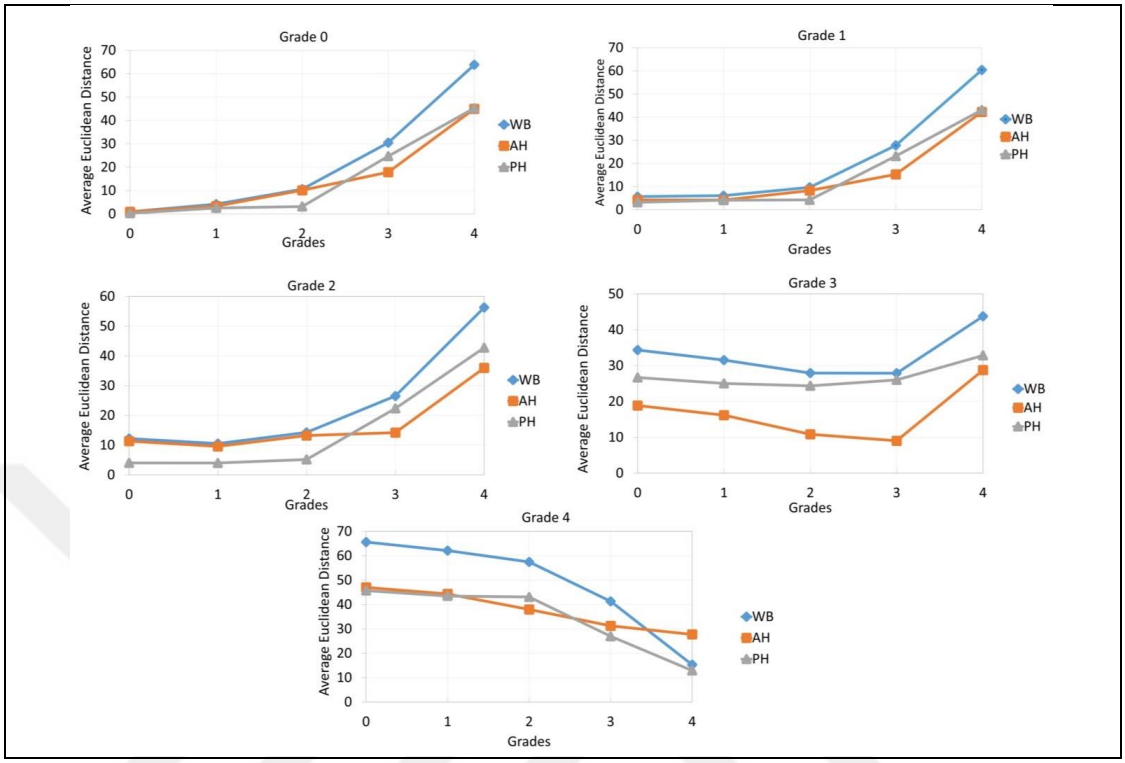
difficult to predict the grade of new unseen data based on a volumetric threshold alone because some grades overlap heavily in that respect. Furthermore, a plot of the summation of the average signatures and the respective standard deviations (Figure 5.4) reveals the overlap between the volumes of the various grades.

5.3.2 Automatic Grading of PWMH via Machine Learning

In order to train a model that can be used to accurately predict the PWMH grade of future subjects, it is necessary to intelligently select quantifiable features that best describe the given label. Based on the results of our analysis, we gathered that feature extraction should be carried out on each quadrant of the signature.

Salient and distinctive features that we extracted include: the number and location of peaks that are present in each quadrant. The peaks represent the extent of the spread of lesions as observed on an image. However, given the spiky nature of the signature, we narrowed down the number of peaks so that only significant peaks (those above 40 percent of the height of the highest peak) were located. The 40 percent value was determined empirically and it refers to the visibly *obvious* lesions in the image. We computed three peak-related features namely the value of the highest peak, its radial location and the most significant peak(s) to include in our feature matrix. The feature matrix was made up of seventeen ($4 \times 4 + 1$) features, the last one being the summation of all signature values.

Figure 5.5: Inter-grade and intra-grade mean Euclidean distances for WB, AH and PH. WB=Whole Brain, AH=Anterior Horn, PH=Posterior Horn



To address the multi-class classification problem, we employed the use of multi-class Support Vector Machine (SVM) learning algorithm with a Gaussian kernel. The parameters C and γ were optimized with a 3 fold cross validation accuracy as the criterion. Feature selection was run so that only the features that best predicted the correct grades were selected.

5.4 RESULTS

5.4.1 Lesion-Based PWMH Rating

The confusion matrix presented in Table 5.2 provides an insight to the detailed performance of the SVM model for the four PWMH grades.

5.4.2 Distance-based Similarity Analysis of PWMH Signatures

By investigating the Euclidean distance measure between PWMH signatures, we were able to ascertain the inter-grade and intra-grade similarities. We expected to observe higher distance measures between subjects of different grades and lower values for those from the same grade.

Table 5.3: Average three-fold cross validation accuracies of Bayindir & MS Challenge datasets with and without feature selection.

Dataset	Average CV Accuracy (%)	
	All features	Selected features
Bayindir	58.8	73.5
MSChallenge	95.2	95.2

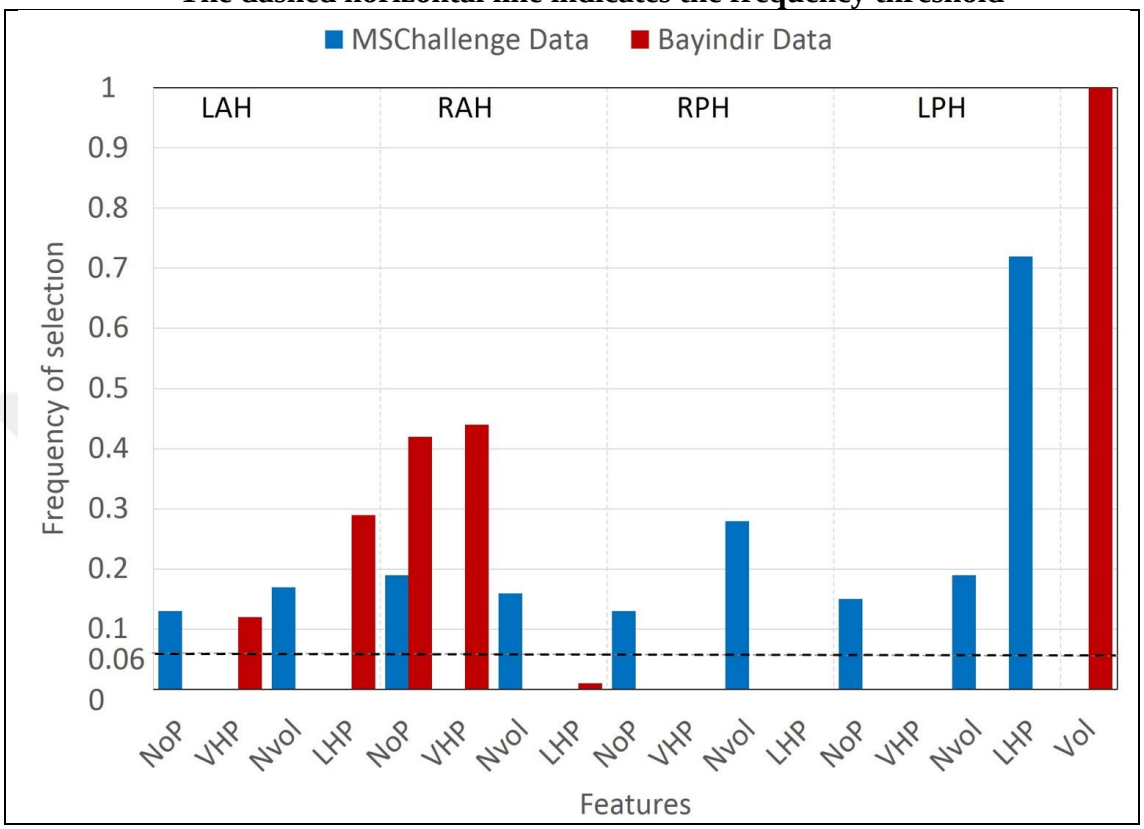
A pairwise comparison between the distance of each subject's signature and the average of the rest of the subjects within each grade was made. The results depicted in Figure 5.5 shows the outcome of the analysis that was performed on the whole brain (WB), the anterior horn (AH) and the posterior horn (PH) signatures. Ideally, for every curve, the lowest distance value should be that of the grade in question. Curves with non-ideal characteristics imply confusion between grades.

To shed more light on Figure 5.5 we present an example. The graph titled Grade 0 shows the mean Euclidean distance between the mean signature (Whole Brain, Anterior and Posterior Horn) of Grade 0 subjects and the mean signature (WB, AH and PH) of subjects from each respective grade; the same goes for the other graphs in the figure. The revelations of Figure 5.5 are as follows: the distance curves from grade 0 subjects exhibited ideal characteristics. Subjects of grade 1 however revealed confusion with grade 0 subjects since the value of the distance at grade 0 is lower than it is at grade 1. For grade 2 subjects, the confusion exists with the lower grades, especially for WB and AH. The grade 3 curves exhibited ideal characteristics except in the PH where the confusion exists with the lower grades. Grade 4 subjects did not exhibit any unexpected characteristics in terms of the distance measure analysis.

5.4.3 Signature Similarity Based Grading of PWMHs

Subject grades were predicted based on the minimum distance between a subject's signature and the average signatures of the five grades. The analysis was carried out for WB, AH and PH. The prediction accuracies were poor; with the highest accuracy observed for WB and AH at 51.9%. Given the foregone analysis, automatic classification of PWMH from the PWMH signature cannot be successfully achieved by such a simple classification model. The analysis further reveals the intricate nature of the problem thus more complex classification algorithms are required to tackle it.

Figure 5.6: Bar graph showing the frequency of features selected after running sequential feature selection for 100 Monte Carlo simulations. VHP=Value of Highest Peak, LHP=Location of Highest Peak, NoP=Number of significant Peaks, Vol=Total normalized Volume. The dashed horizontal line indicates the frequency threshold



5.4.4 SVM Based PWMH Grading

Due to the inability of the distance measure to express the degree of similarity between signatures sufficiently, we turned our attention to machine learning. The results presented here show the prediction capability of a model trained and tested on both datasets. We employ SVMs with a Radial Basis Function kernel and then feature selection with an SVM wrapper function.

The correlation between the total PWMH load per subject in the Bayindir dataset and the grades was found to be 83.5% thus it led us to begin the feature selection by training the SVM wrapper function with the total load as the only feature and then adding features that were selected recursively until the optimum grade prediction accuracy was obtained. On top of that, we performed 100 Monte Carlo simulations and made use of the features that were selected the most number of times. The same procedure was carried out for the MS Challenge dataset.

Table 5.4: Sample confusion matrices and grade prediction accuracy of the SVM model. AG = Actual Grade, PG = Predicted Grade, GA = Grade Accuracy, and OA = Overall Accuracy.

		AG									
		Bayindir data					MS Challenge data				
		0	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**
PG	**0**	14	1	0	0	0	4	0	0	0	0
	1	0	11	2	1	0	0	12	1	0	0
	2	0	1	8	1	1	0	0	0	0	0
	3	0	1	3	2	0	0	0	0	4	0
	4	0	0	1	3	3	0	0	0	0	0
GA (%)		100	78.6	57.1	28.6	75	100	100	0	100	-
OA (%)		73.3					95.2				

It is obvious from Figure 5.6, which features were to be used to train the model. By empirically setting a frequency threshold of 0.06, 5 of the seventeen features were selected from the Bayindir dataset while 9 of the seventeen features were selected for the MS Challenge dataset. According to the feature selection routine, the features that best characterize the grades for the Bayindir dataset are: the value of the highest peak in the left anterior horn and in the right anterior horn; the location of the highest peak in the left anterior horn and in the right posterior horn; the number of significant peaks in the right anterior horn and the total volume of PWMH. Though the feature selection routine did not affect the cross validation accuracy of the MS Challenge dataset, it refined the features so that only the most descriptive features were selected. The mean prediction accuracy obtained for both datasets for the 100 Monte Carlo simulations are presented in Table 5.6.

Table 5.4 shows how the SVM model performed on one of the Monte Carlo simulations to yield an overall prediction accuracy of 73.3% and 95.2% on the Bayindir and MS Challenge datasets respectively. The performance of the model is further evaluated through the precision, sensitivity and specificity of the individual grades as shown in Table 5.5.

Table 5.5: Precision, sensitivity and specificity of the individual grades (Bayindir data).

Measure	Grades				
	0	1	2	3	4
Precision (%)	93	79	73	33	43
Sensitivity (%)	100	79	57	29	75
Specificity (%)	97	92	92	91	92

5.5 DISCUSSION

5.5.1 Automatic Lesion-based PWMH rating

As the automatic PWMH lesion rating was preliminary, we did not proceed with it. However, the overall prediction accuracy of almost 80 percent was encouraging. The work fell short because:

i Our rating scale was modified after the work was completed

ii The automatic rating was performed for individual lesions and not for the whole brain

iii The features extracted did not include any texture characteristics

iv The 3D nature of lesions were not taken into account

v The data was highly imbalanced

5.5.2 Automatic Signature-based PWMH Rating

We obtained a threefold cross validation accuracy of over 73% and 95% on the Bayindir and MS Challenge datasets respectively. With regard to the more challenging dataset (Bayindir), the distance measure analysis guided us to divide the signature into sub-signatures in order to achieve better classification accuracy. The results were as expected for the intra-grade distance measures, while unexpected results were obtained for the inter-grade case, which can be attributed to the confusion in the total PWMH load of different grades (notice the overlapping standard deviations in Figure 5.4). Grade 4 subjects showed a clear distinction from the others especially in the posterior horn, which may be a result of the limited number of grade 4 subjects available.

Table 5.6: Prediction accuracies of the two methods employed for automatic PWMH classification using the signature (Bayindir data)

Method	Accuracy
SVM	73.4
Signature similarity	51.9

The overall accuracies presented in Table 5.6 show a clear improvement in prediction accuracies between the two models however further analysis reveal that the SVM model is highly specific with the lowest specificity being 91%. The low sensitivity of grade 3,

as seen in Table 5.5, where 3 of the subjects were confused to be in grade 4 and 2 others in grades 1 and 2 each (Table 5.4); can be attributed to the high similarity between grades 3 and 4 subjects' loads. It is clear from the confusion matrix that the model performed better on the grades with larger number of subjects (grades 0, 1 and 2).

The imbalanced structure of our data is worth noting because it plays an important role in the predictive power of the trained model. This problem of imbalanced datasets in machine learning comes naturally in real world tasks and even more so in medical data analysis. It is further demonstrated in the MS Challenge dataset where patients with high grades were significantly less than those with the low grades.

5.6 CONCLUSION

The introduction of the automation of periventricular white matter hyperintensity rating should not be discounted as visual ratings are highly subjective (Iheme, Kandemir, Yalciner, & Unay, 2015).

In this chapter, we have presented the development as well as the implementation of the periventricular white matter hyperintensity signature. The quantitative as well as graphical model opened up the possibility of automating the process by use of machine learning algorithms. While researchers have, until this point, focused on automatic lesion segmentation, the issue of rating seems to have been ignored.

The in-depth analysis of the signature exposed the complexity of the problem and called for a robust solution since an attempt to automatically rate using the distance measure was not sufficient. The division of the signature into sub-signatures created an opportunity for a more detailed examination of the problem. After extracting features

from the signature, it was imperative to carry out feature selection so as to determine the optimal features that best described the signature.

The results obtained after testing the algorithm on both datasets were acceptable and promising. Nevertheless, more importantly, the results were not subjective and could lead to the standardization of periventricular white matter hyperintensity rating.

6. CONCLUSION & FUTURE WORK

Neuroimaging has long depended on engineering for the manufacturing and maintenance of imaging hardware devices. In recent years however, findings in the field of digital image processing and neuroimaging have necessitated collaborative studies between medical experts and engineers. Thus, the importance of the application of computer vision to medical imaging has been on the rise. Our work, the automation of a diagnostic process, is one of many that bridge the gap between the two fields of study.

In this thesis, we have demonstrated the capability of computer vision to fill the void and exceed the limitation of human interference in a diagnostic process. Specifically, our work has automated the process of segmenting and rating demented patients' white matter lesions from their MRIs. We have employed as well as developed algorithms to automate otherwise subjective processes. In addition, the data provided would serve as an invaluable resource to the research community.

The major contributions of this work are as follows: 1. White matter hyperintensity dataset consisting of expert rated real MRIs obtained from routine practice. 2. Automatic segmentation of periventricular white matter hyperintensity using Conditional Random Field and comparison of the results with the state-of-the-art algorithms. 3. Automatic rating of periventricular white matter heyperintensity via our novel signature based approach.

6.1 WHITE MATTER HYPERINTENSITY DATABASE

Our database consists of MRIs of patients who have been medically diagnosed with dementia. The white matter lesions have been manually segmented and rated by experts. The rating scale can be easily replicated in order to reproduce the results. In fact, with adequate training, non-medical personnel may manually segment and rate since our segmentation tool is easy to use and the rating scale is not complicated.

In addition to the FLAIR images that were used in this work, the dataset includes T1, and T2 images which may be useful to other researchers in their work. Since the images are in their native space, it gives users the privilege of processing the data from scratch.

6.2 PERIVENTRICULAR WHITE MATTER HYPERINTENSITY SEGMENTATION

Our segmentation regime involved a light pre-processing stage, lateral ventricle segmentation, interhemispheric fissure detection and 3D connectivity analysis. The core segmentation algorithm used were 1. Thresholding 2. Support vector machines and 3. Conditional random fields. The three algorithms were compared and tested on our dataset and another dataset obtained from a segmentation challenge (Carass A. , et al., 2017).

6.2.1 Thresholding-Based Segmentation

Thresholding is the simplest algorithm we applied. It involved obtaining an optimum threshold with which images could be segmented. The performance of the algorithm was unsurprisingly below par for medical image segmentation purposes. It is however noteworthy that it performed better on our dataset than on the challenge data. The results of thresholding alone left many unacceptable false positives that required post-processing to get rid of.

The 3D connectivity analysis, as its name denotes, examines the image for lesions which were too small and weren't connected across adjacent slices. This step was necessary since the core segmentation algorithm only examines one voxel at a time and does not take neighborhood information into account. The overall result drastically improved but still was not satisfactory.

6.2.2 Machine Learning-Based Segmentation

The main prerequisite for applying any machine learning-based segmentation is feature extraction. Our features were voxel-centered but included neighborhood information by means of spatial filtering masks. Our feature matrix consisted of 17 features which were common to both machine learning based algorithms. As expected, the machine learning based algorithms outperformed the thresholding-based one.

6.2.2.1 Periventricular White Matter Hyperintensity Segmentation via Support Vector Machines

Our expectation before applying this algorithm to our data was to see an improvement over the thresholding based algorithm and it was met. After the model was optimized via cross validation accuracy, the results of prediction were acceptable but needed improvement. Due to the fact that, SVM does not consider neighborhood samples when

making predictions, it could not accurately segment small lesions. Such segmentations were usually riddled with misclassifications. Furthermore, as a result of the pre-processing regime of the MS Challenge data, the SVM result deteriorated. The finer voxel resolution required neighborhood information in order to make predictions that are more meaningful.

6.2.2.2 Periventricular White Matter Hyperintensity Segmentation via Conditional Random Fields

The application of conditional random fields was highly advantageous mostly because the algorithm takes neighboring samples into account while making predictions. The segmentation results we obtained were superior to the results from the other algorithms. Again, the algorithm performed better on the Bayindir data than on the MS Challenge data. Nevertheless, both results are considered to be excellent - approximately 90% mean Dice score each.

6.3 PERIVENTRICULAR WHITE MATTER HYPERINTENSITY SIGNATURE

Visualizing periventricular lesions has been innate right from the onset of its discovery. Researchers have depended on visual spatial representations for this purpose. However, a graphical representation has the potential to uncover new ground one of which is the potential to automate some diagnostic processes.

In our novel work, we made use of the power of the *signature* to predict the periventricular white matter hyperintensity grade of patients. Based on manually rated data and our rating scale, we trained a classifier that predicted patients' grades. After rigorous analyses and testing, the prediction accuracies obtained were acceptable but more importantly, the results open up a new horizon of research. On the MS Challenge data, the result was quite impressive, reaching a prediction accuracy of over 95%.

The main factor that might have affected the obtained result is the imbalanced nature of the datasets. Signatures of patients with high grades (3 & 4) were sometimes unavailable.

6.4 FUTURE WORK

The work presented in this study is inexhaustible; it opens up new areas of research. We intend to extend the different facets of the work in the following ways:

6.4.1 Automatic White Matter Hyperintensity Segmentation

This thesis focused on the segmentation of periventricular white matter hyperintensity. An extension to this is the segmentation of deep white matter hyperintjensity as well. Lesions located outside the periventricular region also play a part in the diagnosis of not only Alzheimer's disease but in other neurological disorders as well.

In our work, we did not perform any parameter optimization for CRF. The default parameters were used. However, perhaps with parameter optimization, better results might be obtained.

We have compared the performance of three algorithms, thresholding based, SVM and CRF. It would be interesting to see how the performance of the relatively newer machine learning based algorithms compare to CRF and SVM. For instance, convolutional networks have recently been reported to perform excellently when applied to medical image segmentation (Bernal, et al., 2017).

6.4.2 Automatic White Matter Hyperintensity Rating via the Signature

So far, we have extracted descriptive features from the signature and trained an SVM classifier to predict grades. In the future, we hope to be able to extract features that are even more descriptive as well as test other machine leaning algorithms for the purpose of grade prediction.

The problem of data imbalance is commonplace in real world machine learning problems. In our experiments, we tried a couple of strategies to overcome the challenge, these included over-sampling/under-sampling and Synthetic Minority Over-sampling TEchnique SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2011). Neither of these strategies improved our results. SMOTE was designed and optimized for decision trees wherein the minority class was oversampled by up to 500% of its original size and an improvement was reported. We are interested in testing our algorithm on a more balanced to see the effect on the grade prediction accuracy.

We are curious to see how our algorithm performs on datasets from other sites. We are confident that with a balanced training dataset, our algorithm will perform much better. Though we tried to consolidate our dataset, the challenge we faced was that we did not have enough samples to generate synthetic data from. The signature can be adapted to

other rating scales and in the future, we hope to test the prediction capability on other rating scales.

Finally, we are interested in automating the entire diagnostic process, from segmentation to rating whole brain white matter lesions not just the periventricular lesions.

REFERENCES

Books

Vapnik, V. (1998). *Statistical Learning Theory.* New York: Wiley.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory* (2 ed.). New York: Springer-Verlag . doi:10.1007/978-1-4757-3264-1

Periodicals

Admiraal-Behloul, F., Van Den Heuvel, D. M., Olofsen, H., Van Osch, M., Van der Grond, J., Van Buchem, M. A., & Johan H C, J. H. (2005). Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage, 28*, 607-617.

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004* (pp. 39-50). Springer.

Anbeek, P., Vincken, K. L., Osch, M. J., Bisschops, R. H., & Grond, J. (2004). Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Medical image analysis, 8 3*, 205-15.

Anurag, A., Shuai, Z., Sadeep, J., Bernardino, R.-P., Måns, L., Alexander, K., . . . Philip, H. T. (2018, January 11). Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation: Combining Probabilistic Graphical Models with Deep Learning for Structured Prediction. *IEEE Signal Processing Magazine*, pp. 37 - 52.

Arnold, A. S., Wilson, J. S., Boshier, M. G., & Smith, J. (1998, 3). A Simple Extended-Cavity Diode Laser. *Review of Scientific Instruments, 69*, 1236-1239. Retrieved from http://link.aip.org/link/?RSI/69/1236/1

Bartko, J. J. (1991). Measurement and reliability: statistical thinking considerations. *Schizophrenia Bulletin, 17*, 483-489.

Bento, M., Souza, R., Lotufo, R., Frayne, R., & Rittner, L. (2018). WMH Segmentation Challenge: A Texture-Based Classification Approach. In A. Crimi, S. Bakas, H. Kuijf, B. Menze, & M. Reyes (Ed.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (pp. 489-500). Cham: Springer International Publishing.

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Marti, R., & Lladó, X. (2017). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *CoRR, abs/1712.03747*.

Besson, J. A., Ebmeier, K. P., Best, P. V., & Smith, F. W. (1988). Do white matter changes on MRI and CT differentiate vascular dementia from Alzheimers disease? *Journal of Neurology, Neurosurgery & Psychiatry, 51*, 318-319. doi:10.1136/jnnp.51.2.318-a

Beyer, M. K., Aarsland, D., Greve, O. J., & Larsen, J. P. (2006, 2). Visual rating of white matter hyperintensities in Parkinson's disease. *Mov. Disord., 21*, 223-229. doi:10.1002/mds.20704

Breteler, M. M., Van Swieten, J. C., Bots, M. L., Grobbee, D. E., Claus, J. J., Van Den Hout, & others. (1994). Cerebral white matter lesions, vascular risk factors, and cognitive function in a population-based study The Rotterdam Study. *Neurology, 44*, 1246-1246.

Brickman, A. M., Muraskin, J., & Zimmerman, M. E. (2009). Structural neuroimaging in Alzheimer's disease: do white matter hyperintensities matter? *Dialogues in clinical neuroscience, 11*, 181.

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 121–167.

Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., & Cherubini, A. (2015, 7 01). Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. *Neuroinformatics, 13*, 261-276. doi:10.1007/s12021-015-9260-y

Capizzano, A. A., Ación, L., Bekinschtein, T., Furman, M., Gomila, H., Martínez, A., . . . Starkstein, S. E. (2004). White matter hyperintensities are significantly associated with cortical atrophy in Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry, 75*, 822-827.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., . . . Pham, D. L. (2017). Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *Neuroimage, 148*. doi:10.1016/j.neuroimage.2016.12.064

Cash, D. M., Frost, C., Iheme, L. O., Ünay, D., Kandemir, M., Fripp, J., . . . Ourselin, S. (2015). Assessing atrophy measurement techniques in dementia: Results from

the MIRIAD atrophy challenge. *NeuroImage, 123*, 149-164. doi:https://doi.org/10.1016/j.neuroimage.2015.07.087

Cerasa, A., Bilotta, E., Augimeri, A., Cherubini, A., Pantano, P., Zito, G., . . . Quattrone, A. (2012, 1). A Cellular Neural Network methodology for the automated segmentation of multiple sclerosis lesions. *Journal of neuroscience methods, 203*, 193-9.

DeCarli, C., Murphy, D. G., Tranh, M., Grady, C. L., Haxby, J. V., Gillette, J. A., . . . others. (1995). The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults. *Neurology, 45*, 2077-2084.

Dijk, E. J., Prins, N. D., Vrooman, H. A., Hofman, A., Koudstaal, P. J., & Breteler, M. M. (2008). Progression of Cerebral Small Vessel Disease in Relation to Risk Factors and Cognitive Consequences Rotterdam Scan Study. *Stroke, 39*, 2712-2719.

Erkinjuntti, T., Gao, F., Lee, D. H., Eliasziw, M., Merskey, H., & Hachinski, V. C. (1994). Lack of difference in brain hyperintensities between patients with early Alzheimer's disease and control subjects. *Archives of neurology, 51*, 260-268.

Fayao, L., Guosheng, L., & Chunhua, S. (2015). CRF learning with CNN features for image segmentation. *Pattern Recognition*, 2983-2992.

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., & Zimmerman, R. A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Roentgenology, 149*, 351-356.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*, 381-395.

Gerard, G., & Weisberg, L. A. (1986). MRI periventricular lesions in adults. *Neurology, 36*, 998-998.

Gibson, E., Gao, F., Black, S. E., & Lobaugh, N. J. (2010). Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T. *Journal of Magnetic Resonance Imaging, 31*, 1311-1322.

Hawthorn, C. J., Weber, K. P., & Scholten, R. E. (2001, 12). Littrow Configuration Tunable External Cavity Diode Laser with Fixed Direction Output Beam. *Review of Scientific Instruments, 72*, 4477-4479. Retrieved from http://link.aip.org/link/?RSI/72/4477/1

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on, 21*, 1263-1284.

Jeon, S., Yoon, U., Park, J.-S., Seo, S. W., Kim, J.-H., Kim, S. T., . . . Lee, J.-M. (2011). Fully automated pipeline for quantification and localization of white matter hyperintensity in brain magnetic resonance image. *International Journal of Imaging Systems and Technology, 21*, 193-200.

Jokinen, H., Kalska, H., Mäntylä, R., Ylikoski, R., Hietanen, M., Pohjasvaara, T., . . . Erkinjuntti, T. (2005). White matter hyperintensities as a predictor of neuropsychological deficits post-stroke. *Journal of Neurology, Neurosurgery & Psychiatry, 76*, 1229-1233. doi:10.1136/jnnp.2004.055657

Kapeller, P., Barber, R., Vermeulen, R. J., Adèr, H., Scheltens, P., Freidl, W., . . . Fazekas, F. (2003, 1). Visual rating of age-related white matter changes on magnetic resonance imaging: Scale comparison, interrater agreement, and correlations with quantitative measurements. *Stroke, 34*, 441-445. doi:10.1161/01.STR.0000049766.26453.E9

Karimaghaloo, Z., Shah, M., Francis, S. J., Arnold, D. L., Collins, D. L., & Arbel, T. (2012, 6). Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields. *IEEE transactions on medical imaging, 31*, 1181-94.

Kempton, M. J., Underwood, T. S., Brunton, S., Stylios, F., Schmechtig, A., Ettinger, U., . . . others. (2011). A comprehensive testing protocol for MRI neuroanatomical segmentation techniques: Evaluation of a novel lateral ventricle segmentation method. *Neuroimage, 58*, 1051-1059.

Kim, K. W., MacFall, J. R., & Payne, M. E. (2008). Classification of White Matter Lesions on Magnetic Resonance Imaging in Elderly Persons. *Biological Psychiatry, 64*, 273-280. doi:http://dx.doi.org/10.1016/j.biopsych.2008.03.024

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., . . . Davatzikos, C. (2007). Computer-Assisted Segmentation of White Matter Lesions in 3D MR images, Using Support Vector Machine. *Academic Radiology*, 300 - 313.

Maillard, P., Delcroix, N., Crivello, F., Dufouil, C., Gicquel, S., Joliot, M., . . . Mazoyer, B. (2008). An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology, 50*, 31-42.

Maniega, S. M., Valdés Hernández, M. C., Clayden, J. D., Royle, N. A., Murray, C., Morris, Z., . . . Wardlaw, J. M. (2015, 2). White matter hyperintensities and normal-appearing white matter integrity in the aging brain. *Neurobiol. Aging, 36*, 909-918. doi:10.1016/j.neurobiolaging.2014.07.048

Mäntylä, R., Erkinjuntti, T., Salonen, O., Aronen, H. J., Peltonen, T., Pohjasvaara, T., & Standertskjöld-Nordenstam, C. G. (1997, 8). Variable agreement between visual rating scales for white matter hyperintensities on MRI. Comparison of 13 rating scales in a poststroke cohort. *Stroke., 28*, 1614-1623.

Mease, D., Wyner, A. J., & Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *The Journal of Machine Learning Research, 8*, 409-439.

Mohsen, G., Nico, K., Tom, H., Inge, W. M., Clara, I. S., Geert, L., . . . and Bram, P. (2017). Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities. *Scientific Reports*.

Nyquist, P. A., Bilgel, M., Gottesman, R., Yanek, L. R., Moy, T. F., Becker, L. C., . . . Vaidya, D. (2015, 4). Age differences in periventricular and deep white matter lesions. *Neurobiol. Aging, 36*, 1653-1658. doi:10.1016/j.neurobiolaging.2015.01.005

Pantoni, L., Simoni, M., Pracucci, G., Schmidt, R., Barkhof, F., & Inzitari, D. (2002). Visual rating scales for age-related white matter changes (leukoaraiosis): Can the heterogeneity be reduced? *Stroke, 33*, 2827-2833.

Pol, L. A., Verhey, F., Frisoni, G. B., Tsolaki, M., Papapostolou, P., Nobili, F., . . . others. (2009). White matter hyperintensities and medial temporal lobe atrophy in clinical subtypes of mild cognitive impairment: the DESCRIPA study. *Journal of Neurology, Neurosurgery & Psychiatry, 80*, 1069-1074.

Popescu, V., Battaglini, M., Hoogstrate, W. S., Verfaillie, S. C., Sluimer, I. C., Schijndel, R. A., . . . Yousry, T. (2012, 7). Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. *Neuroimage, 61*, 1484-1494. doi:10.1016/j.neuroimage.2012.03.074

Prins, N. D., Dijk, E. J., Heijer, T., Vermeer, S. E., Jolles, J., Koudstaal, P. J., . . . Breteler, M. M. (2005). Cerebral small-vessel disease and decline in information processing speed, executive function and memory. *Brain, 128*, 2034-2041.

Ricardo, G., C., Q., Ozan, O., C., B., L., C., Richard, J., . . . D., R. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, 918-934.

Schmidt, R., Schmidt, H., Haybaeck, J., Loitfelder, M., Weis, S., Cavalieri, M., . . . others. (2011). Heterogeneity in age-related white matter changes. *Acta neuropathologica, 122*, 171-185.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., & Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: voxel-based lesion-symptom mapping evidence from aphasia. *Brain : a journal of neurology, 132 Pt 12*, 3411-27.

Sheline, Y., Price, J., Vaishnavi, S., Mintun, M., Barch, D., Epstein, A., . . . others. (2008). Regional white matter hyperintensity burden in automated segmentation distinguishes late-life depressed subjects from comparison subjects matched for vascular risk factors. *American Journal of Psychiatry, 165*, 524-532.

Shimada, K., Kawamoto, A., Matsubayashi, K., & Ozawa, T. (1990). Silent cerebrovascular disease in the elderly. Correlation with ambulatory pressure. *Hypertension, 16*, 692-699.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping, 17*, 143-155.

Strelkov, V. V. (2008). A new similarity measure for histogram comparison and its application in time series analysis. *Pattern Recognition Letters, 29*, 1768-1774.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *Medical Imaging, IEEE Transactions on, 29*, 1310-1320.

Valdés Hernández, M. C., Piper, R. J., Bastin, M. E., Royle, N. A., Maniega, S. M., Aribisala, B. S., . . . Wardlaw, J. M. (2014, 1). Morphologic, distributional, volumetric, and intensity characterization of periventricular hyperintensities. *AJNR. American journal of neuroradiology, 35*, 55-62.

Van den Heuvel, D. M., Ten Dam, V. H., Craen, A. J., Admiraal-Behloul, F., Olofsen, H., Bollen, E. L., . . . Buchem, M. A. (2006). Increase in periventricular white matter hyperintensities parallels decline in mental processing speed in a non-demented elderly population. *Journal of Neurology, Neurosurgery & Psychiatry, 77*, 149-153.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., & Suetens, P. (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. *Medical Imaging, IEEE Transactions on, 20*, 677-688.

Van Straaten, E. C., Fazekas, F., Rostrup, E., Scheltens, P., Schmidt, R., Pantoni, L., . . . Barkhof, F. (2006). Impact of white matter hyperintensities scoring method on correlations with clinical data: The LADIS study. *Stroke, 37*, 836-840.

Wahlund, L.-O., Agartz, I., Almqvist, O., Basun, H., Forssell, L., Sääf, J., & Wetterberg, L. (1990). The brain in healthy aged individuals: MR imaging. *Radiology, 174*, 675-679.

Wardlaw, J. M., Ferguson, K. J., & Graham, C. (2004, 5). White matter hyperintensities and rating scales - Observer reliability varies with lesion load. *J. Neurol., 251*, 584-590. doi:10.1007/s00415-004-0371-x

Wu, M., Rosano, C., Butters, M., Whyte, E., Nable, M., Crooks, R., . . . Aizenstein, H. J. (2006). A fully automated method for quantifying and localizing white matter hyperintensities on MR images. *Psychiatry Research: Neuroimaging, 148*, 133-142.

Yue, N. C., Arnold, A. M., Longstreth, W. T., Elster, A. D., Jungreis, C. A., O'Leary, D. H., . . . Bryan, R. N. (1997). Sulcal, ventricular, and white matter changes at MR imaging in the aging brain: data from the Cardiovascular Health Study. *Radiology*, *202*, 33-39.

Other Sources

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., . . . others. (2008). SPM8 manual. *Functional Imaging Laboratory, Institute of Neurology*.

Catanese, L., Commowick, O., & Barillot, C. (2015). Automatic graph cut segmentation of multiple sclerosis lesions. *ISBI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge.*

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 27:1--27:27.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*.

Dzung, P., Pierre-Louis, B., Aaron, C., Peter, C., Ciprian, C., Lotta, E., . . . Snehashis, R. (2015). *The 2015 Longitudinal multiple sclerosis lesion segmentation challenge*.

Ekin, A. (2006, 9). Feature-based brain mid-sagittal plane detection by RANSAC. *Signal Processing Conference, 2006 14th European*, (pp. 1-4).

Göçeri, E., Durá, E., & Gunay, M. (2016). Review on Machine Learning Based Lesion Segmentation Methods from Brain MR Images. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 582-587.

Hugues, S., & Yale, S. (2015, January 09). *HCRF library (including CRF and LDCRF)*. Retrieved from sourceforge: https://sourceforge.net/projects/hcrf/

Iheme, L. O., Kandemir, M., Yalciner, Z. B., & Unay, D. (2015). Beyaz Cevher Hiperintensitesinin Görsel Derecelendirmesinde Otomasyon, Standardizasyon ve Tutarlılık Cagrısı A Call for Automation, Standardization and Consistency in White Matter Hyperintensity Visual Grading. *Vogue, 15*.

Iheme, L. O., Ösman, B., Şennaz, A., Tepe, S. M., Kandemir, M., Kahraman, T., . . . Ünay, D. (2012). Automatic Grading of Periventricular White Matter Hyperintensities using FLAIR MRI. *ESMRMB Congress* (pp. 156-157). Lısbon: ESMRMB.

Iheme, L. O., Ünay, D., Başkaya, O., Şennaz, A., Kandemir, M., Yalçıner, Z. B., . . . others. (2013). Concordance between computer-based neuroimaging findings and expert assessments in dementia grading. *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, (pp. 1-4).

Jennings, D., Fennessy, F., Sonka, M., & Buatti, J. (2013). 3D Slicer. *30*, 1323-1341. doi:10.1016/j.mri.2012.05.001.3D

Maria, E. C., Paolo, P., Antonio, A., Federico, R., Aldo, Q., & Andrea, C. (2015). Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. *Download PDF*, 261–276.

Salamin, H., Morency, L.-P., Stratou, G., & Song, Y. (2011). *HCRF library*. Retrieved from https://sourceforge.net/projects/hcrf/

Van Leemput, K., Maes, F., Bello, F., Vandermeulen, D., Colchester, A., & Suetens, P. (1999). Automated segmentation of MS lesions from multi-channel MR images. *Medical Image Computing and Computer-Assisted Intervention--MICCAI'99*, (pp. 11-21).

Wieman, C. E., & Hollberg, L. (1991, 1). Using Diode Lasers for Atomic Physics. *Review of Scientific Instruments, 62*, 1-20. Retrieved from http://link.aip.org/link/?RSI/62/1/1

APPENDICES

Appendix A.1 TABLE 1, VISUAL MRI RATING SCALE

Atrophy	
Sulcal	0–9
Ventricular	0–9

MTA[i]

	Width of choroid fissure	Width of temporal horn	Height of hippocampal formation
0	N	N	N
1	↑	N	N
2	↑↑	↑	↓
3	↑↑↑	↑↑	↓↓
4	↑↑↑	↑↑↑	↓↓↓

WMH[ii]

Periventricular	Subcortical
0 – No lesion 1 – Caps 2 – Thin line 3 – Halo 4– Irregular, extending to the deep white matter	0 – No lesion 1 – < 5 small focal and/or <2 large focal lesions 2 – 5–12 small focal and/or 2–4 large focal lesions 3 – >12 small focal and/or >4 large focal or confluent lesions 4 – predominantly confluent lesions

Infarcts	
Basal ganglia	**Infratentorial**
0 – No lesion	0 – No lesion
1 – Few lesions (1–3)	1 – Few lesions (1–3)
2 – Many lesions (>4)	2 – Many lesions (>4)

[i] Medial temporal lobe atrophy
[ii] White matter hyperintensities