FEDERAL UNIVERSITY OF SÃO CARLOS

TECHNOLOGY AND EXACT SCIENCES CENTER

COMPUTER SCIENCE GRADUATE PROGRAM

AUTOMATIC COMPUTATIONAL SCHEME FOR SEGMENTATION, VOLUMETRIC ASSESSMENT AND ANALYSIS OF MULTIPLE SCLEROSIS LESIONS IN MAGNETIC RESONANCE IMAGES OF THE HUMAN BRAIN

PAULO GUILHERME DE LIMA FREIRE

SUPERVISOR: RICARDO JOSÉ FERRARI, PHD

São Carlos – SP

September/2019

FEDERAL UNIVERSITY OF SÃO CARLOS

TECHNOLOGY AND EXACT SCIENCES CENTER

COMPUTER SCIENCE GRADUATE PROGRAM

AUTOMATIC COMPUTATIONAL SCHEME FOR SEGMENTATION, VOLUMETRIC ASSESSMENT AND ANALYSIS OF MULTIPLE SCLEROSIS LESIONS IN MAGNETIC RESONANCE IMAGES OF THE HUMAN BRAIN

PAULO GUILHERME DE LIMA FREIRE

Thesis document presented to the Computer Science Graduate Program of the Federal University of São Carlos as part of the requisites to obtain the title of Doctor in Computer Science, concentration area: Computer Techniques and Methodologies
Supervisor: Ricardo José Ferrari, PhD

São Carlos – SP

September/2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Paulo Guilherme de Lima Freire, realizada em 27/09/2019:

Prof. Dr. Ricardo José Ferrari
UFSCar

Prof. Dr. Diego Furtado Silva
UFSCar

Prof. Dr. Alexandre Luis Magalhães Levada
UFSCar

Prof. Dr. Paulo Mazzoncini de Azevedo Marques
USP

Prof. Dr. Adilson Gonzaga
USP

ACKNOWLEDGMENTS

First and foremost, I thank God for my life.

I thank my supervisor, Dr. Ricardo José Ferrari, for all the knowledge shared with me during these years and the always fruitful discussions of ideas.

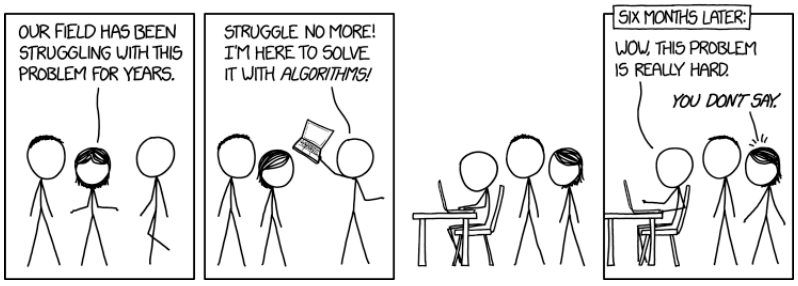
I thank the thorough and competent Doctors from UNIFESP, namely Enedina, Marcos, Nitamar and Carrete, for their valuable contributions. Without their efforts and help we would not have achieved the full potential of this work.

I thank my parents, Paulo and Lúcia, for all their support, dedication and unconditional love throughout my whole life, for pushing me to go further and for showing me by example what righteousness means.

I thank my girlfriend and partner in crime (and life), Isabela, for walking beside me and inspiring me to be a better person every day by being a role model of intelligence and determination.

I also thank my lab friends who have been with me along for the ride – and along for the coffee as well.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. I also thank the São Paulo Research Foundation (FAPESP) for the financial support given to this research (grant number 2016/15661-0).



xkcd.com/1831

RESUMO

Esclerose Múltipla (EM) é uma doença inflamatória e desmielinizante do sistema nervoso central (SNC). É considerada uma doença autoimune na qual o sistema imunológico reconhece erroneamente a bainha de mielina do SNC como um elemento estranho e então a ataca, resultando em inflamação e formação de cicatrizes gliais (escleroses) em múltiplas áreas da substância branca do SNC. O imageamento multi-contraste por ressonância magnética (RM) tem sido usado clinicamente com muito sucesso para o diagnóstico e monitoramento da EM devido às suas excelentes propriedades de alta resolução e boa diferenciação entre tecidos moles. Duas áreas de interesse do ponto de vista computacional podem ser destacadas neste contexto: a segmentação das lesões e identificação de quais estão em estágio inflamatório, também chamadas de lesões realçadas ou ativas. Atualmente, o método utilizado para a segmentação de lesões de EM é o delineamento manual em imagens 3D de RM. Tal procedimento é realizado por especialistas com ajuda limitada do computador. Entretanto, tal procedimento é custoso e propenso à variabilidade inter e intraobservadores. Neste ponto, este projeto propõe o estudo e o desenvolvimento de um esquema computacional automático para a segmentação, medição volumétrica e análise de lesões de EM em imagens de RM utilizando uma abordagem baseada em modelos de mistura finita de distribuições *t-Student* e atlas probabilísticos para a segmentação e medição do volume de lesões de EM em imagens de RM. Quanto à identificação de lesões realçadas, contrastes à base de gadolínio são usados para destacá-las visualmente das demais. Porém, estudos recentes indicam a perda fisiológica gradual da capacidade dos pacientes em eliminar esta substância, sendo proporcional ao número de injeções intravenosas já administradas no paciente. Isto leva ao acúmulo do contraste no organismo. Assim sendo, este projeto faz também o uso de características de texturas para fazer a distinção de lesões realçadas (ativas) e não-realçadas sem a necessidade da aplicação de contraste à base de gadolínio, eliminado o risco de acúmulo intravenoso dessa substância no organismo do paciente e barateando o custo do procedimento de RM.

Palavras-chave: esclerose múltipla, segmentação, classificação, textura, t-Student.

ABSTRACT

Multiple Sclerosis (MS) is an inflammatory demyelinating (that is, with myelin loss) disease of the Central Nervous System (CNS). It is considered an autoimmune disease in which the immune system wrongly recognizes the myelin sheath of the CNS as an external element and attacks it, resulting in inflammation and scarring (sclerosis) of multiple areas of CNS's white matter. Multi-contrast magnetic resonance imaging (MRI) has been successfully used in diagnosing and monitoring MS due to its excellent properties such as high resolution and good differentiation between soft tissues. In this context, from a computational standpoint, two important goals stand out: lesion segmentation and lesion classification, the latter being related to the identification of which lesions are under an inflammatory state, also called active or enhancing lesions. Nowadays, the preferred method to segment MS lesions is manual delineation, made by specialists with limited aid of a computer. However, this approach is tiresome, expensive and prone to error due to inter- and intra-variability between observers caused by low contrast on lesion edges. Here, we propose the development of an automatic computational technique based on Student's t-distribution finite mixture models and probabilistic atlases to segment and measure MS lesions volumes in MR images. Regarding the identification of enhacing lesions, Gadolinium-based contrasts are used to visually highlight them during an MRI procedure. However, recent studies indicate that patients gradually lose their ability to eliminate the contrast substances from their bodies when they undergo many contrast injections throughout their lives, which is the case for MS subjects. In this sense, in this work we used textural features to distinguish enhancing (active) and nonenhancing lesions without the aid of intravenous injection of Gadolinium-based contrast, thus eliminating the risk of accumulation of this substance in one's body and making the MRI procedure faster and cheaper.

Keywords: multiple sclerosis, segmentation, classification, texture, Student's t-distribution.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC – *Area Under the receiver operating haracteristic Curve*

CAD – *Computer-aided diagnostics*

CNN – *Convolutional neural network*

CNS – *Central Nervous System*

CSF – *Cerebral Spinal Fluid*

DSC – *Dice Similarity Coefficient*

DTPA – *Dynamic Texture Parameters Analysis*

EF – *Extra Fraction*

EM – *Expectation Maximization*

FANTASM – *Fuzzy and Noise Tolerant Adaptive Segmentation Method*

FCN – *Fully Convolutional Network*

FLAIR – *Fluid-attenuated Inverse-recovery*

FLICM – *Fuzzy Local Information C-Means*

FPR – *False Positive Rate*

FP – *False Positives*

GLCM – *Gray level Co-Occurrence Matrix*

GMM – *Gaussian Mixture Model*

GM – *Gray Matter*

HI – *Hyperintensity*

HMRF-FCM – *Fuzzy C-Means based Hidden Markov Random Field*

IPD – *Intensity Profile Distinction*

LBP – *Local Binary Pattern*

LDA – *Linear Discriminant Analysis*

LFPR – *Lesion False Positive Rate*

LI – *Lesion Intersection*

LTPR – *Lesion True Positive Rate*

MAP – *Maximum a Posteriori*

MCR – *Misclassification Ratio*

MEANF – *Mean Field Algorithm*

MFCM – *Modified Fuzzy C-Means*

MLE – *Maximum Likelihood Estimate*

MRF – *Markov Random Field*

MRI – *Magnetic Resonance Image*

MR – *Magnetic Resonance*

MSP – *Mid-sagittal Plane*

MS – *Multiple Sclerosis*

NAWM – *Normal Appearing White Matter*

NLM – *Non-Local Means*

NWM – *Normal White Matter*

OF – *Overlap Fraction*

PDF – *Probability Density Function*

PD – *Proton Density*

PPMS – *Primary-Progressive Multiple Sclerosis*

PRMS – *Progressive-Relapsing Multiple Sclerosis*

PR – *Probabilistic Rand*

PVE – *Partial Volume Effect*

RBF – *Radial basis function*

RF – *Radio Frequency*

RLM – *Run Length Matrix*

ROI – *Region Of Interest*

RRMS – *Relapsing-Remitting Multiple Sclerosis*

SMM-SC – *Spatially Constrained Student's t-mixture model*

SMM-iter – *Iterative Student's t-mixture Model*

SMM – *Student's t-mixture model*

SPMS – *Secondary Progressive Multiple Sclerosis*

SP – *Specificity*

SS – *Sensitivity*

SVFMM – *Spatially Variant Finite Mixture Model*

SVM – *Support Vector Machine*

T1-w – *T1-weighted*

T2-w – *T2-weighted*

TA – *Texture Analysis*

TPR – *True Positive Rate*

TP – *True Positives*

TV – *Third Ventricle*

VD – *Volume Difference*

WMH – *White Matter Hyperintensities*

WML – *White Matter Lesions*

WM – *White Matter*

k-NN – *k-Nearest Neighbors*

SUMMARY

Chapter 1

INTRODUCTION

This chapter presents the context of this work, the motivation to provide a solution for the multiple sclerosis automatic segmentation problem and the goals of this research.

1.1 Context and motivation

Multiple sclerosis (MS) is a demyelinating disease that attacks the central nervous system (CNS) and affects more than 2 million people worldwide (BROWNE et al., 2013). It destroys neurons' myelin sheaths, causing many effects on one's body and mind, including dizziness, confusion, memory problems and numbness of arms and legs (World Health Organization, 2008). The cause of MS is still unknown, and the disease itself has a devastating effect on individuals and society because its onset is typically around the age 30, thus affecting subjects at the peak of their productivity in life (WARREN; WARREN, 2001).

According to a study conducted by the World Health Organization (2008), the average prevalence (i.e., number of subjects with a condition in a population in a given time period) of MS around the world is of thirty people for every one hundred thousand. Europe is the continent with the highest prevalence, with eighty people for every one hundred thousand. Comparatively, this same number in the Americas is of only 8.3 people. Besides presenting prevalence numbers, this study also confirmed that MS is not a disease restricted to the most developed countries and can occur in every part of the globe.

Since there is still no cure for MS and patients must undergo regular exams and follow-ups, it is typically an expensive disease to treat. For instance, direct and indirect care costs range from eight thousand dollars to fifty-four thousand dollars per patient per year in the United States. MS ranks second regarding costliness compared to other chronic conditions, topped

only by congestive heart failure(National Multiple Sclerosis Society, 2013).

There are four different types of MS. The most frequent one is called relapsing-remitting MS (RRMS) and is characterized by clearly defined relapses of increased disease activity and worsening symptoms, which are then followed by remissions in which the disease does not progress. Approximately 85% of people with MS are diagnosed with RRMS at onset. The second type is called secondary-progressive MS (SPMS) and often occurs in patients who do not treat RRMS within a decade of the initial diagnosis. The third type is primary-progressive MS (PPMS), which is diagnosed in approximately 10% of MS patients at onset. This type is more severe than the previous ones, and people with PPMS experience a steady progression of the disease with no apparent relapses or remissions. Finally, the fourth type is called progressive-relapsing MS (PRMS) and is the rarest form of MS with only 5% of patients being diagnosed with it. It causes clear relapses combined with a steady progression of the disease (National Multiple Sclerosis Society, 2013).

The most used medical imaging technique to help diagnose MS is the magnetic resonance imaging (MRI). Since it provides excellent differentiation between soft tissues (BUSHBERG et al., 2012; COMPSTON; COLES, 2008), MRI allows physicians to identify lesions and follow the disease progression over time. By tuning pulse sequence parameters appropriately, different types of image weights, namely T1-w, T2-w, Fluid Attenuated Inversion Recovery (FLAIR) and Proton Density (PD), can be generated. Each image weight provides different responses for brain tissues and pathologies. Regarding MS imaging diagnosis, FLAIR images are often used because lesions appear hyperintense in this particular image weight, thus making it easier for physicians to identify them (HASHEMI et al., 1995).

Up to this day, the gold standard for MS lesion segmentation is an expert's manually annotated ground truth. However, this is an expensive and time-consuming procedure, and is inherently biased due to subjective perceptions that vary from expert to expert (ZIDJDENBOS; FORGHANI; EVAN, 2002). Moreover, the very nature of MS lesions poses challenges to their segmentation since they do not have a defined shape, size or location (BROSCH et al., 2016). In this sense, many automatic segmentation techniques have been proposed over the years (BAKSHI et al., 2008; GARCÍA-LORENZO et al., 2013; X. et al., 2012; VOVK; PERNUS; LIKAR, 2007) in order to provide a more consistent output and decrease the cost and time spent on the segmentation task itself. However, due to the challenging characteristics of lesion segmentation, no automatic technique has been adopted as a reference standard so far, indicating that this is still an open field of research to be further explored.

From a tissue clustering point of view, this work is focused only on images that have MS

lesions in them, since they represent the cluster we want to accurately segment. Our approach makes use of finite mixture models (MCLACHLAN; KRISHNAN, 1997; DUDA; HART; STORK, 2000) and information from image gray level intensities to provide a reliable segmentation. By achieving that, quantitative data on lesions can be extracted and used to give a better grasp on how the disease is progressing and help physicians understand if a given treatment prescribed to a patient is being effective, especially in clinical trials scenarios.

Another important aspect of this work is the distinction between enhancing (or active) and non-enhancing (or nonactive) lesions (LEWIS; SPILLANE, 2019). To identify which lesions are active, a Gadolinium-based contrast is injected in the patient prior to the procedure itself. The properties of the contrast allow it to enhance active lesions (LEWIS; SPILLANE, 2019), making them distinguishable from other kinds of lesions. However, a number of studies from various research groups in the last few years (BURKE et al., 2016; HU et al., 2016; ZOBEL et al., 2016) indicate that Gadolinium-based contrasts tend to accumulate in one's brain, bones, skin and other parts of the body after a number of injections.

A comprehensive understanding of this accumulation is not yet available, which has made several agencies, such as the U.S. Food and Drug Administration and the European Medicines Agency, issue statements (European Medicines Agency, 2016; AGENCY, 2017; U.S. Food and Drug Administration, 2015; FOOD; ADMINISTRATION, 2017, 2018) restricting the usage of Gadolinium-based contrasts only to cases where it is absolutely necessary. Though there are indications that infrequent contrast administrations pose no threat (KROMREY et al., 2016), this is not the case for MS patients, who must undergo an MRI procedure with contrast injections from time to time for the rest of their lives in order to assess how the disease is progressing.

Given this scenario, in this work we also proposed a supervised algorithm to distinguish active from nonactive lesions in Gadolinium-free FLAIR images, thus eliminating the need for contrast injections altogether. The benefits of a correct classification span from improving patients' health - by avoiding their exposure to a heavy metal-based contrast and its effects on well-being - to reducing the time and costs of the MRI procedure itself.

1.2 Goals

The primary goal of this work was to study and develop an automatic computational pipeline to segment and classify MS lesions in MR images. For this purpose, a finite mixture model based on the Student's t-distribution (ZHANG; WU; NGUYEN, 2013; NGUYEN; WU, 2012; SFIKAS; NIKOU; GALATSANOS, 2007) will be used as a clustering technique. We chose this model due to

its unsupervised nature. Also, the Student's t-distribution has a heavy tail, making it less sensitive to outliers when compared to other widely used distributions, i.e., the Gaussian distributions (CHATZIS; KOSMOPOULOS; VARVARIGOU, 2009).

There is, however, a portion of lesions that will inherently have either a significant intensity overlap with other tissues, such as white matter (WM) and gray matter (GM) or be too small (< 10 voxels) to be correctly segmented. This is a drawback that affects automatic and manual segmentation approaches alike (BENTO et al., 2017) and must be taken into consideration.

Moreover, we classified lesions according to their inflammatory state (active and nonactive) using the XGBoost classifier (CHEN; GUESTRIN, 2016), since the active ones serve as a proxy for the progression of the disease. As mentioned in Section 1.1, most active lesions are usually identified using a Gadolinium-based contrast agent injected into the subject a few minutes prior to the MRI exam, but recent studies have shown that the more often a subject is injected with the contrast, the harder it becomes to be eliminated. In this sense, identifying active lesions from the segmentation output without using any contrast agent can help with the patients' health and decrease the cost of the MRI procedure. As important as the technique used in classification is the quality of the data to be used as input. Regarding this matter, information from gray level intensities, hyperintensity enhancement and textures extracted from run length matrices(CASTELLANO et al., 2004)and gray level co-occurrence matrices (HARALICK; SHANMUGAM; DINSTEIN, 1973) were combined to form a discriminative feature space efficient to distinguish active and nonactive MS lesions.

Byproducts of this work involved developing a technique to improve the distinction between hyperintensities in FLAIR images and other brain tissues and to compare different techniques to detect the midsaggital plane (MSP), which is used to halve the brain into its two hemispheres in order to analyze symmetry information.

1.3 Publications

The goals described in Section 1.2 resulted in the following publications:

- (Qualis CAPES A2) Freire, P.G.L. and Ferrari, R.J. "Automatic iterative segmentation of multiple sclerosis lesion using Student's t mixture models and probabilistic anatomical atlases in FLAIR images". Computers in Biology and Medicine (73): 10-23, 2016 (FREIRE; FERRARI, 2016)

- (Qualis CAPES B1) Freire, P.G.L., Silva, B.C.G., Pinto, C.H.V., Moreira, C. and Ferrari,

R.J. “Midsaggital plane detection in magnetic resonance images using phase congruency, Hessian matrix and symmetry information: a comparative study”. 18th International Conference on Computational Science and its Applications (FREIRE et al., 2018).

- (Qualis CAPES A1) Freire, P.G.L. and Ferrari, R.J. “Multiple sclerosis lesion enhancement and white matter region estimation using hyperintensities in FLAIR images”. Biomedical Signal Processing and Control (49): 338-348, 2019 (FREIRE; FERRARI, 2019)

And the following paper is currently under review (Qualis CAPES A1):

- (Qualis CAPES A1) Freire, P.G.L., Idagawa, M., Oliveira, E.M.L., Abdala, N. Carrete, H. and Ferrari, R.J. “Classification of active multiple sclerosis lesions without the aid of Gadolinium-based contrast using textural and enhanced features from FLAIR images”. (Medical Image Analysis - under review) .

All of the accepted paper are attached to Appendix B.

1.4 Thesis organization

This thesis is divided into five more chapters and an appendix. The content in each of them is the following.

- Chapter 2: details the image formation process regarding the physics of MRI and how images are generated. It also showcases different image weights and how tissues and lesions appear on each one of them.
- Chapter 3: background knowledge on the Student’s t-mixture model and Expectation-Maximization algorithm and how they are combined in the context of image segmentation. We also present theoretical information about textures and the XGBoost classifier. This chapter also presents the metrics used to assess the segmentation and classification accuracies.
- Chapter 4: literature review with works related to MS lesion segmentation, textures and symmetry.
- Chapter 5: details of our methodology, databases and pipeline.
- Chapter 6: results from MSP detection, hyperintensity enhancement and MS lesion segmentation and classification.

- Chapter 7: conclusions drawn from the results and next steps of this work.
- Appendix A: details of the Expectation-Maximization used to estimate parameters in a Student's t-distribution.
- Appendix B: full papers mentioned in Section 1.3.

Chapter 2

IMAGE FORMATION IN MRI

This chapter aims to explain how magnetic resonance images are generated and the response of different brain tissues and lesions have on different image weights.

2.1 Physics of MRI

The interaction of different kinds of materials with magnetic fields is a typical situation in daily life. Taking a compass as an example, its needle always points to the direction of lines that come out of the north pole to the south pole of a magnetic field (HALLIDAY; RESNICK; WALKER, 2010). In this scenario, if a second magnetic field is positioned perpendicular to the first one and starts oscillating in a specific resonance frequency, it can disturb the needle, which in turn will align instead to this second field. If we cease the oscillation and remove the second field from the system, the needle will return to its initial alignment state. The vibration of the needle when returning to its original state emits magnetic waves according to the intensity of the field it is currently under the influence of. These waves can be captured, measured and used to provide information about the surroundings of the needle.

Similarly, the same process happens to the hydrogen atoms of one's body undergoing a MRI scan. Hydrogen is the element of choice for MRI for three main reasons. First, it is the most abundant element in the human body. Second, hydrogen characteristics in magnetic resonance differ between normal tissues and pathologies. And third, the proton of hydrogen atoms have a significant magnetic moment, making them more sensitive to the magnetic field of an MRI scanner (MAZZOLA, 2009).

An important aspect related to MRI is the precession phenomenon, which is intrinsic to atoms with nuclei that have an odd number of protons and/or neutrons. It is characterized by a

spinning movement around the atom's axis, as shown in Figure 2.1. This kind of movement has two relevant peculiarities. The first one is the atoms will precess around that field's direction when under a magnetic field. This behavior is known as the Larmor precession. Its frequency is proportional to the intensity of the magnetic field and is given by

$$\omega_0 = \gamma B_0, \tag{2.1}$$

where γ is the gyromagnetic ratio (which is a constant and unique value for each chemical element) and B_0 is the intensity of the field (also constant).

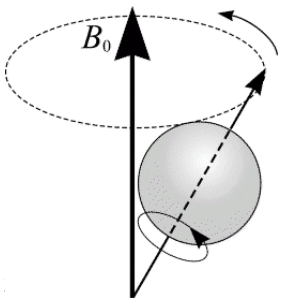


Figure 2.1: Precession movement.

The second peculiarity is that atoms emit magnetic waves in the precession frequency. Suppose a 3D coordinate system (x,y,z), where z is related to the longitudinal axis and the (x,y) plane is associated with the transversal plane. In this sense, to disturb the alignment of atoms to an initial field B_0, parallel to the z-axis, a perpendicular magnetic field B_1 must oscillate in the precession movement; in other words, it must oscillate in the Larmor frequency. By doing so, atoms will start to precess around B_1, on the transversal plane (x,y). If B_1 is removed from the system, atoms return to precess around B_0, in a movement called relaxation. During relaxation, atoms lose energy and emit radio frequency signals through magnetic waves.

Initially, the hydrogen atoms in one's body are not aligned to any particular direction. When undergoing a MRI scan, a first and constant magnetic field B_0 is applied to the subject's body, making his/her hydrogen atoms align parallel or anti-parallel to it, as shown in Figure 2.2. Then a second magnetic field B_1, perpendicular to B_0, is applied with the same oscillation frequency as the atoms' Larmor frequency, thus creating a resonance effect (and hence the name of the technique itself). At this moment, the hydrogen atoms start to leave their state of alignment to B_0. After a brief period of time, the second magnetic field B_1 is turned off, making the atoms return to their previous alignment state to B_0. As mentioned before, they release energy in this process, which is then captured and analyzed by a scanner. This procedure can be used to get information about the neighborhood of each atom. And since the response of each atom is directly associated with the tissue they are in, an image with tissue nuances can be generated

and provide a way to visualize the scanned body part in a 3D image (BUSHBERG et al., 2012).

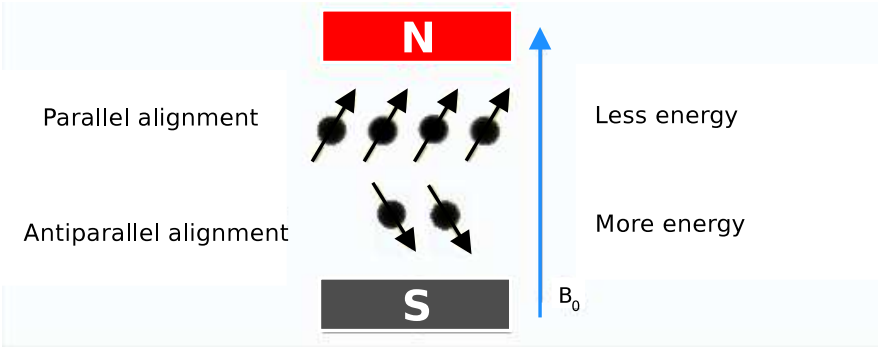


Figure 2.2: Atoms aligned parallel or antiparallel to B_0.

2.1.1 Image weights

Image weights in MRI can be defined as a combination of radio frequency pulses and gradients that generate images with different tissue responses. The gray level intensity value of each voxel in MRI varies according to the proton density (PD) of each tissue. The higher the proton density is, the more intense the response of the radio frequency signal in that area (HANSON, 2009). The contrast between tissues in MRI also depends on two other parameters: T_1, the longitudinal relaxation time, and T_2, the transversal relaxation time. More precisely, T_1 is the time it takes for atoms under the influence of B_1 to return to balance after this field is removed from the system. In other words, it is the time required to get them aligned back to B_0. On the other hand, T_2 is related to the magnetization reduction time in the transversal plane. In practice, this is a situation where the magnetization vector, initially aligned to B_1 in the transversal plane, gets back to being aligned with B_0 in the z-axis, as shown in Figure 2.3.

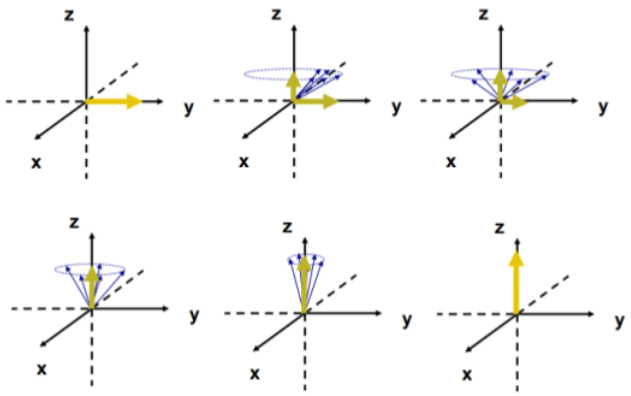


Figure 2.3: Return of magnetization when B_1 is removed from the system.

The return of magnetization to the longitudinal axis is given by

$$\mathbf{M}_z = \mathbf{M}_0(1 - e^{\frac{-t}{T1}}), \tag{2.2}$$

where $\mathbf{M}_z$ is the magnetization vector in the z axis, $\mathbf{M}_0$ is the initial magnetization vector, t is the time and $T1$ is the longitudinal relaxation constant.

Similarly, the magnetization decay in the transversal plane is given by

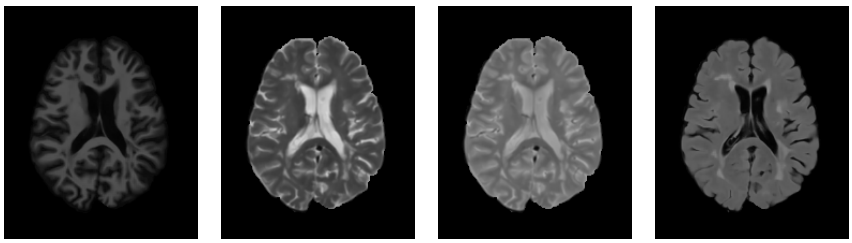
$$\mathbf{M}_{xy} = \mathbf{M}_0 e^{\frac{-t}{T2}}, \tag{2.3}$$

where $\mathbf{M}_{xy}$ is the magnetization vector in the xy plane and $T2$ is the transversal time.

The radio frequency (RF) pulse emitted by $\mathbf{B}_1$ is repeated at a predefined rate. The period of an RF pulse is called repetition time (TR). The time between a RF pulse is emitted and the atom's response is called echo time (TE). By tuning these two parameters, one can generate different image weights. Overall,

- Long TR and short TE generate PD images.
- Short TR and short TE generate T1-w images.
- Long TR and long TE generate T2-w images.

There are also Fluid Attenuated Inversion-Recovery (FLAIR) images. The main feature of this particular weight is the suppression of the effects of fluids (OKUDA et al., 1999). The $T1$ time in this image weight is adjusted in a way to be equal to the relaxation time of the component to be suppressed. By doing so, some structures, like MS lesions, become more evident in regions where they would otherwise be barely visible. These four different kinds of image weights are shown in Figure 2.4.



(a) (b) (c) (d)

Figure 2.4: Different MR image weights. (a) T1-w, (b) T2-w, (c) PD and (d) FLAIR.

2.1.2 Image composition

An MR image can be seen as a stack of bi-dimensional images, as shown in Figure 2.5. Each slice must be identified and, within it, a spatial mapping must be done between a point

in the image and its gray level intensity value. The way to achieve this is by using gradients (MAZZOLA, 2009).

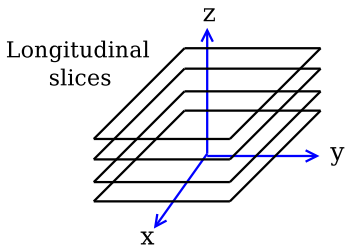


Figure 2.5: Representation of longitudinal image slices in MRI.

Gradients are used to identify each slice by linearly varying the intensity of a magnetic field in a specific direction. In this case, the new field can be calculated as

$$\mathbf{B}_z = \mathbf{B}_0 + \mathbf{G}_z, \tag{2.4}$$

where $\mathbf{B}_z$ is the new intensity of a magnetic field in a given z position, $\mathbf{B}_0$ is the value of a static magnetic field and $\mathbf{G}_z$ is the intensity of the gradient applied to the z-direction. This way, the new field is created locally by using a slice gradient that changes the precession frequency of the atoms in that specific region. In other words, each spatial position in that particular region can be identified thanks to its new specific precession frequency.

The localization of each element (voxel) in a given slice is achieved by using two other gradients that incorporate the phase and frequency components of the signal, respectively. Applying a phase gradient alters the spin phase of the atoms proportionally to their location. This way, one of the axes (x or y) gets phase mapped. Then the frequency gradient is applied to the remaining axis, and the mapping is made in the same manner as with the phase. These elements are then organized into a matrix to be processed (EDELMAN; HESSELINK; ZLATKIN, 2006).

2.1.3 K-space

The K-space is a representation of the frequency domain related to the image acquisition process (MAZZOLA, 2009). The K-space can be seen as a gray level matrix in which each entry corresponds to signal intensity and a point in time, thus representing the amplitude of the signal measured in a given moment. The coordinates (x, y) represent the frequency and phase gradients, respectively. It is important to highlight that there is no correspondence between an entry in the K-space and a voxel in the actual MRI output. In each entry, there is information about the whole slice.

Knowing that the signal of each slice is mapped in terms of phase and frequency (meaning it is a signal that varies over time and has information about all the tissues), the bi-dimensional Fourier transform (BRACEWELL, 2000) is used to calculate the spatial position of the magnetic resonance signal. So the Fourier transform is applied to each matrix in the K-space and the final slice image is finally generated, as shown in Figure 2.6. The coordinates k_x and k_y in the K-space matrix correspond to the frequency and phase gradients, respectively. It is also important to note that the K space can be three-dimensional depending on the acquisition protocol. In these scenarios, a 3D Fourier transform is applied to get a volumetric image instead.

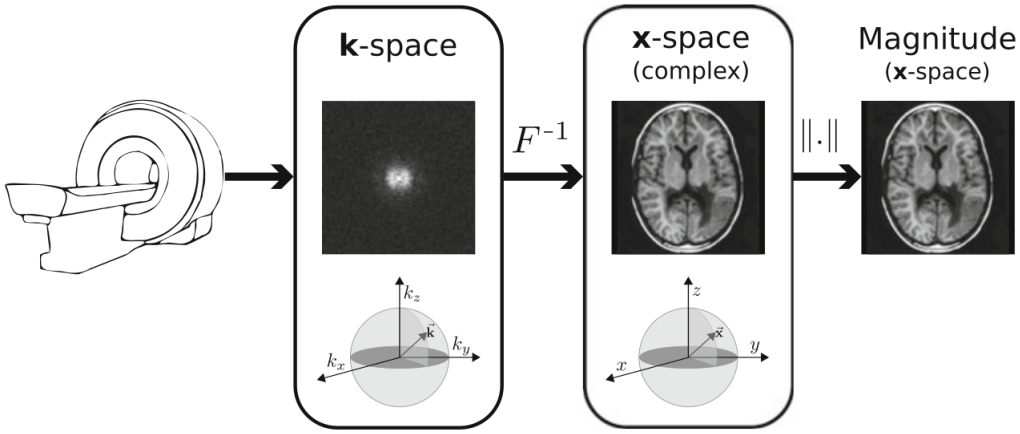


Figure 2.6: K-space and its correspondence to an MRI slice in the image domain (x-space). The signal obtained by the inverse 2D Fourier transform is a complex signal defined over the image domain. In order to generate real data, phase information is discarded, thus yielding the magnitude image. Extracted from (AJA-FERNÁNDEZ; VEGAS-SÁNCHEZ-FERRERO, 2016), Chapter 2.

2.2 Image noise and bias field in MRI

In general terms, noise can be seen as an unwanted fluctuation in the pixel (or voxel) values of an image, degrading image quality. It is a random and stochastic process, making it very difficultto precisely predict its values. However, it is possible to determine its statistical properties by quantifying the noise in terms of an average criteria such as mean value, mean squared value and probability density functions (SUETENS, 2009).

In imaging systems, noise can be either dependent or independent of the signal, being the former labeled as multiplicative noise and the latter, additive noise. Let $g(x,y,z)$ be the voxel value of an image. It can be described as (DOUGHERTY, 2009)

$$g(x,y,z) = f(x,y,z) * h(x,y,z) + n(x,y,z), \tag{2.5}$$

where $f(x,y,z)$ is the "pure" voxel value, $h(x,y,z)$ is the multiplicative noise and $n(x,y,z)$ is the

additive noise.

The most common noise in MR images is called thermal noise, which is additive (DOUGHERTY, 2009). The main sources of this noise come from both the patient and the receiver part of the MR scanner. Though Rician noise can be observed in MR images, in a high signal-to-noise ratio (SNR) scenario, which is common in most scanners nowadays, the noise tends to a Gaussian distribution (NOWAK, 1999), with a probability density function given by

$$p(a) = \sqrt{(1/2\pi\sigma^2)}\exp\left(-(a-\mu)^2/2\sigma^2\right), \tag{2.6}$$

where a is the gray level intensity, μ is the average gray level intensity and σ is its standard deviation. This particular type of noise is convenient for a few reasons, since it can be analytically integrated and has the same spectral shape in the frequency domain. For this reason, there are many noise reduction filters (VAISHALI; RAO; RAO, 2015; MACOVSKI, 1996) that can be applied to mitigate this problem.

Regarding the multiplicative part $h(x,y,z)$, its most common occurrence in MRI is known as bias field, which is a low frequency and very smooth signal that corrupts MR images (JUNTU et al., 2005). This issue is mainly caused by inhomogeneities in the RF field and causes intensity discrepancies within the same tissue; i.e., similar regions from the same tissue end up presenting different gray level intensities. More precisely, bias field is characterized as a gradual change in intensity within segmentation classes across the entire image that cannot be attributed to random noise (RAJAPAKSE; KRUGGEL, 1998). These inhomogeneities can have a significant impact on segmentation and quantitative analysis of MR images, since intensity-based segmentation techniques assume spatial invariance between tissues of the same class across the entire image (DOUGHERTY, 2011).

There are mainly two categories of approaches that can be used to deal with the inhomogeneity problem: prospective and retrospective modeling. The first one uses prior knowledge of the bias field obtained by imaging a homogeneous phantom. However, this phantom only provides a good estimate for objects of the same size as the phantom itself. Since there are many parts of one's body that can be scanned with different sizes, this approach becomes unfeasible. The second approach, on the other hand, is more practical, since it does not depend on phantoms or prior knowledge. Rather, in retrospective approaches, the bias field is modeled based on the low frequency components of the original image (BEHRENBRUCH et al., 2004), thus being able to mitigate the inhomogeneity problem.

It is important to note that adequately treating noise and bias field in MR images may improve segmentation and classification accuracy, since these preprocessing steps are paramount

to reduce intensity variability across tissues and images.

2.3 Tissues appearance in MRI

One of the main difficulties in the context of MS is the intensity profile overlap of tissues and lesions. Though the three main brain tissues (cerebrospinal fluid - CSF, GM and WM) are quite distinguishable intensity-wise, they suffer from partial volume effect (PVE). In other words, transition regions between one tissue and the other often present a mixture of intensities, as shown in Figure 2.7. This effect is amplified if the image resolution is low.

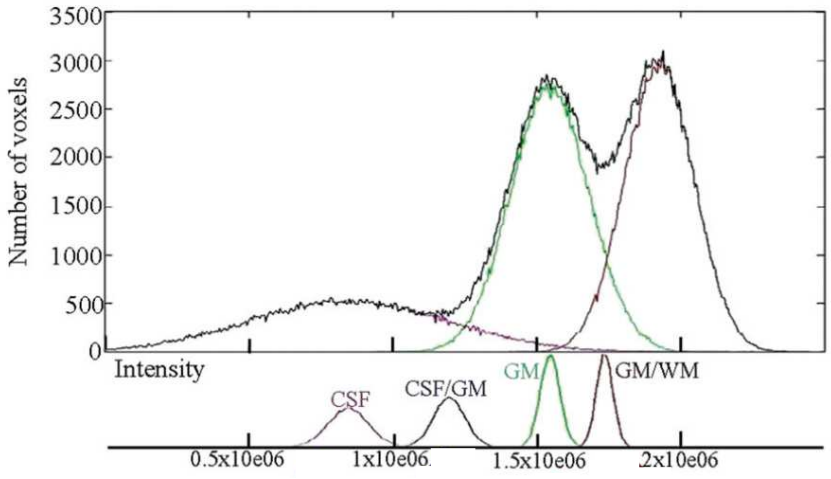


Figure 2.7: Histogram of a MR image depicting distributions of CSF, GM, WM and transition regions. Extracted from (BATTAGLINI; JENKINSON; STEFANO, 2012).

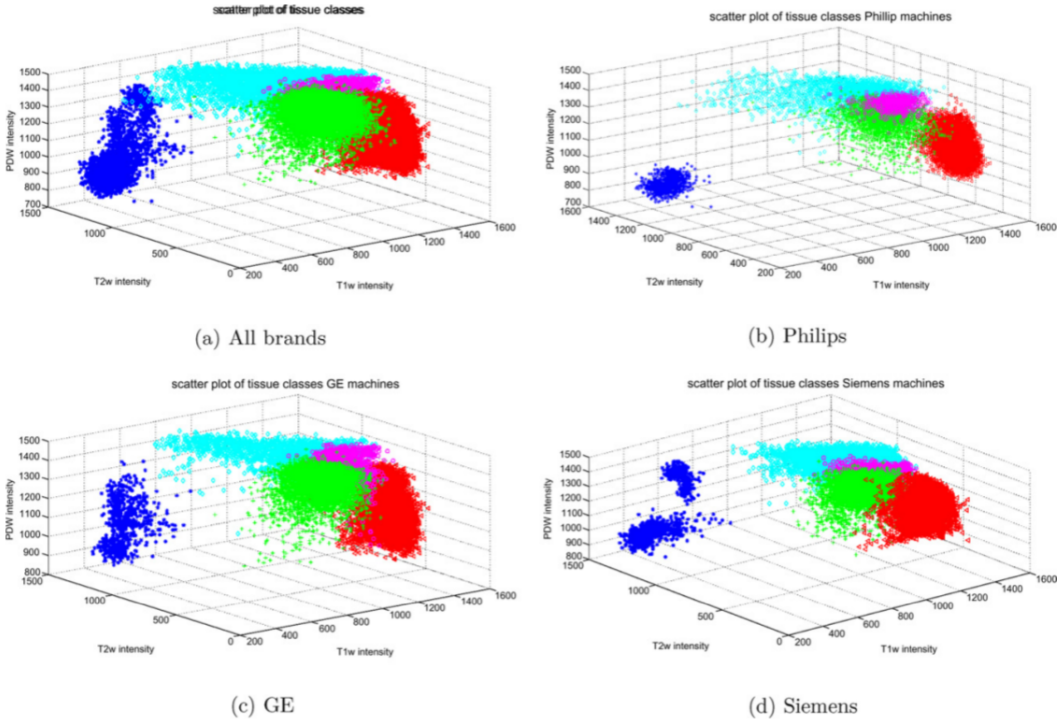
In this sense, multiple sclerosis lesions also have a significant intensity overlap with other brain tissues, mainly WM and GM, as shown in Figure 2.8.

Moreover, they do not have precise shape, size or location in the brain (BORSCH et al., 2015), as it can be seen in Figure 2.9.

However, lesions fall under the category of pathologies that have a rather common appearance, but different intensities in different image weights. In general, they appear hyperintense in T2-w images and hypointense in T1-w (ALI; BUCKLE, 2009). In this regard, it is important to note that not all hyperintense points in T2-w have counterparts in T1-w, as shown in Figure 2.10.

The tissue intensities for each image weight are summarized in Table 2.1 and follow the definition presented in (EDELMAN; HESSELINK; ZLATKIN, 2006).

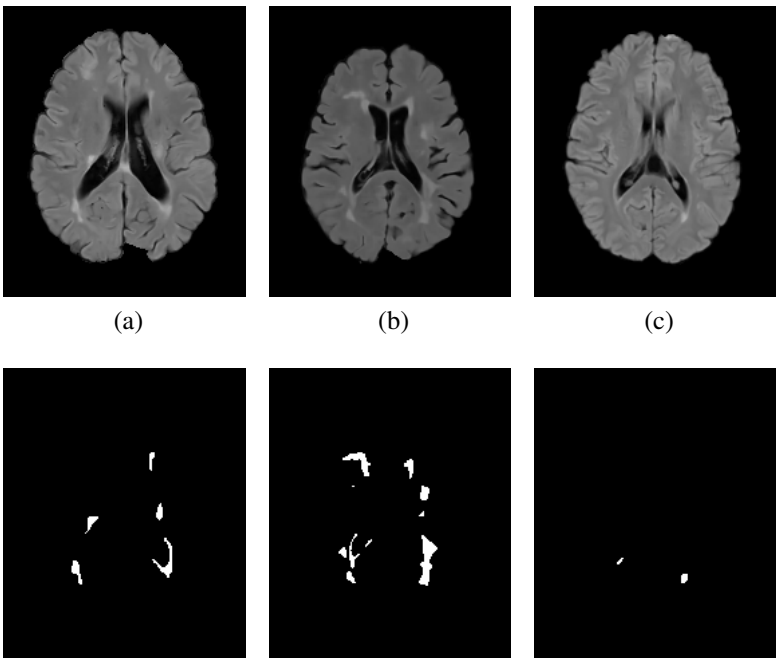
As mentioned before, MS lesions belong to the pathology category, making them appear



(a) All brands (b) Philips

(c) GE (d) Siemens

Figure 2.8: Multiple sclerosis intensity profile overlap with other brain tissues. Red denotes WM, green denotes cortical GM, magenta denotes deep GM, cyan denotes lesions and blue denotes CSF. Note the significant intensity overlap of lesions with GM, regardless of scanner brand. Extracted from (SHAH et al., 2011).



(a) (b) (c)

(d) (e) (f)

Figure 2.9: Different lesion sizes, shapes and locations. (a)-(c) FLAIR images from three different subjects, (d)-(f) lesion masks for each subject, in order of appearance from left to right.

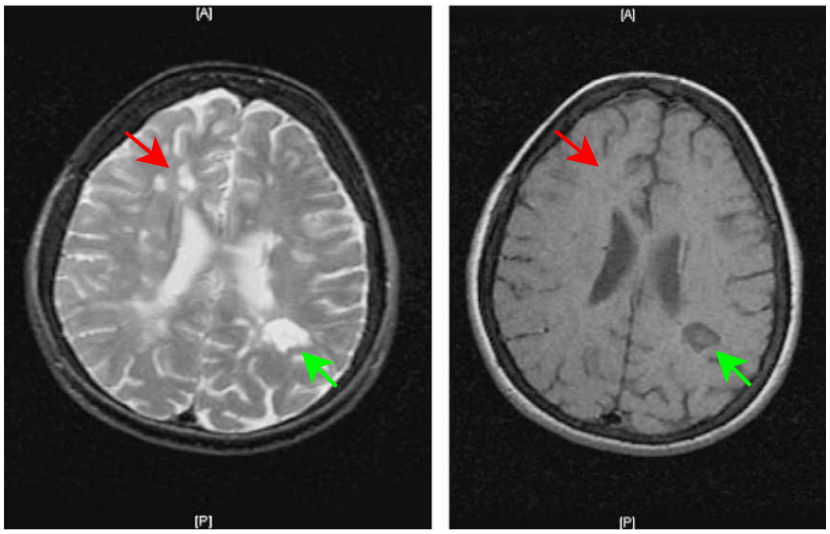


Figure 2.10: MS lesions in T2-w (left) and T1-w (right). Green arrows indicate lesion correspondence between the two image weights, whereas red arrows indicate a lesion that appears only in T2-w.

	T1-w	T2-w	PD
Solid mass	Dark	Bright	Bright
Cyst/pathology	Dark	Bright	Bright
Sub-acute blood	Bright	Bright	Bright
Chronic and acute blood	Mild	Dark	Dark
Fat	Bright	Dark	Bright

Table 2.1: Tissue intensities for different image weights in MRI.

dark (or hypointense) in T1-w and bright (or hyperintense) in T2-w and PD. Regarding FLAIR, hyperintense points (mainly lesions) in this particular image weight are characterized by an intensity profile that partially overlaps with other brain regions, but it is different enough to make it adequate for proper pathology segmentation (HERSKOVITS; ITOH; E.R., 2001).

2.4 Final considerations

This chapter presented information about the magnetic resonance image formation process, image weights and MS lesion appearance in MRI. We can state that image formation and acquisition are complex procedures and it is essential to understand how they work to comprehend the intrinsic characteristics of MR images. It was also mentioned that MS lesions have a plethora of shapes, sizes and locations in the brain. They also have a gray level intensity profile that overlaps with healthy tissues. All these points make the estimation of the model parameters and lesion segmentation and classification even more challenging and indicate why this field of research is still open to new contributions.

Chapter 3

BACKGROUND KNOWLEDGE

This chapter presents an overview of the Student's t mixture model, the Expectation-Maximization algorithm used to estimate parameters of the model, the XGBoost classifier used to distinguish active and nonactive lesions and the texture features used in classification. The metrics used to assess the segmentation and classification accuracies are also detailed.

3.1 Student's t mixture model

A statistical mixture model is a stochastic model commonly used to represent subpopulations included in a general population (TITTERINGTON; SMITH, 1985). In an MRI scan of the human brain, the subpopulations are represented by different kinds of tissues that are part of the brain, i.e., gray matter, white matter and cerebral spinal fluid, while the whole image itself is considered to be the general population. A commonly used feature to distinguish one tissue from another is the gray level intensity from different image weights such as T1-w, T2-w, PD or FLAIR (detailed in Chapter 2). These different gray level intensities are usually put together to create a feature vector for each spatial image location.

In this work, we chose the Student's t distribution because of its "heavy tail," which makes it less susceptible to outliers[1](GEROGIANNIS; NIKOU; LIKAS, 2009). Also, the finite mixture of Student's t-distributions is a model that falls under the unsupervised clustering category for image segmentation. This characteristic is important because it relinquishes the need for a great number of annotations of healthy tissues and pathologies, which is an expensive and time consuming task. A corollary of this fact is that it makes atraining stage unnecessary, thus avoiding any bias that could be artificially introduced by image normalization. Also, a training set may not take into account anatomical and physiological variabilities between different subjects

[1] An outlier means a sample that is quite distant from every cluster in the mixture model.

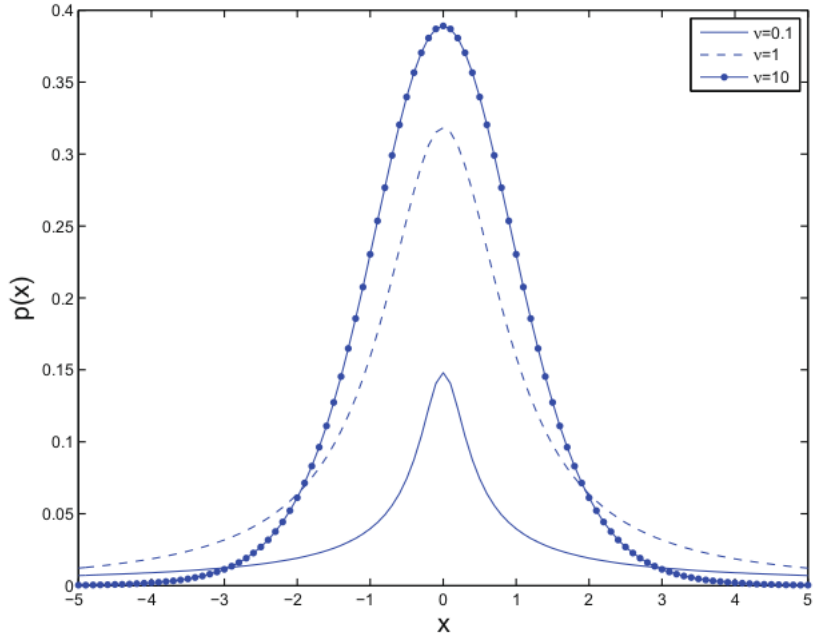


Figure 3.1: Student's t-distribution behavior for different degrees of freedom.

(DESPOTOVIC; GOOSSENS; PHILIPS, 2015).

Formally, a random variable X that follows a multivariate Student's t-distribution has a mean vector μ, a positive, definite and real covariance matrix Σ with dimensions $d \times d$ and degrees of freedom $\nu \in [0, \infty)$ that can be expressed by (MURPHY, 2012):

$$p(X = \mathbf{x}; \mu, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right) |\Sigma|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{d}{2}} \Gamma(\frac{\nu}{2}) \left[1 + \nu^{-1} \delta\left(\mathbf{x}; \mu, \Sigma\right)\right]^{\frac{\nu+d}{2}}}, \tag{3.1}$$

where $\delta(\mathbf{x}; \mu, \Sigma) = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$ is the quadratic Mahalanobis distance and Γ is the Gamma function.

For $\nu \to \infty$, the Student's t-distribution tends to a Gaussian distribution with covariance Σ, as shown in Figure 3.1. Moreover, if $\nu > 1$, then μ is the mean vector of $\mathbf{x}$ and if $\nu > 2$, the covariance matrix of $\mathbf{x}$ is given by $\nu(\nu - 2)^{-1}\Sigma$.

It can be shown that a Student's t-distribution is equivalent to a Normal distribution with a stochastic covariance matrix (GEROGIANNIS; NIKOU; LIKAS, 2009). Formally, given a weight u that follows a Gamma distribution parameterized by ν, we have

$$u \sim \Gamma\left(\nu/2, \nu/2\right). \tag{3.2}$$

Given Equation 3.2, a random variable $\mathbf{x}$ can then follow a multivariate Normal distribution

with a mean vector μ and covariance Σ/u as

$$\mathbf{x}|\mu,\Sigma,\nu,u \sim N(\mu,\Sigma/u). \tag{3.3}$$

The mixture of K Student's t components can be written as

$$\phi(\mathbf{x},\Psi) = \Sigma_{i=1}^{K}\pi_i p(\mathbf{x};\mu_i,\Sigma_i,\nu_i), \tag{3.4}$$

where $\pi_1,\ldots,\pi_k$ are the weights of each component, $\mathbf{x} = (x_1,\ldots x_N)^T$ is the vector of observed data and

$$\Psi = (\pi_1,\ldots,\pi_K;\mu_1,\ldots,\mu_K;\Sigma_1,\ldots,\Sigma_K;\nu_1,\ldots,\nu_K)^T \tag{3.5}$$

represents the complete parameters set of the mixture model components.

3.2 Expectation–Maximization for parameters estimation

A common approach for estimating unknown parameters of a Student's t-mixture model is the Expectation-Maximization (EM) algorithm (PEEL; MCLACHLAN, 2000). Before formally defining it, consider the following example to intuitively understand how it works. Two biased coins, A and B are flipped with bias θ_A and θ_B. The goal is to estimate θ_A and θ_B by repeating five times the rationale of randomly choosing (with equal probability) one of the coins, flipping it, independently, ten times and writing down the number of "heads" observed. In the end, there will be five sets of ten flips each. In this sense, the probability of getting k "heads" in ten flips of coin i ($i \in \{A,B\}$) is given by a binomial

$$p_i(k) = \binom{10}{k}\theta_i^k(1-\theta_i)^{10-k}. \tag{3.6}$$

An iterative way to solve this problem is to guess initial values for θ_A and θ_B and then estimate - for each one of the five sets - which coin is most likely to have been picked for that particular set. From this point on, we can use maximum likelihood estimation to get new values for θ_A and θ_B until convergence (DO; BATZOGLOU, 2008).

To visualize this process, consider Figure 3.2. In it, we have ten observations made for each one of the five sets (box on the left), where H stands for "heads" and T for "tails".

In the first stage, indicated by number 1, we guess two initial values for θ_A and θ_B ($\theta_A = 0.60$, $\theta_B = 0.50$). In stage number 2, the expectation step (or E step) is used to calculate the probability of each coin having generated the observations for each set. To make it clearer,

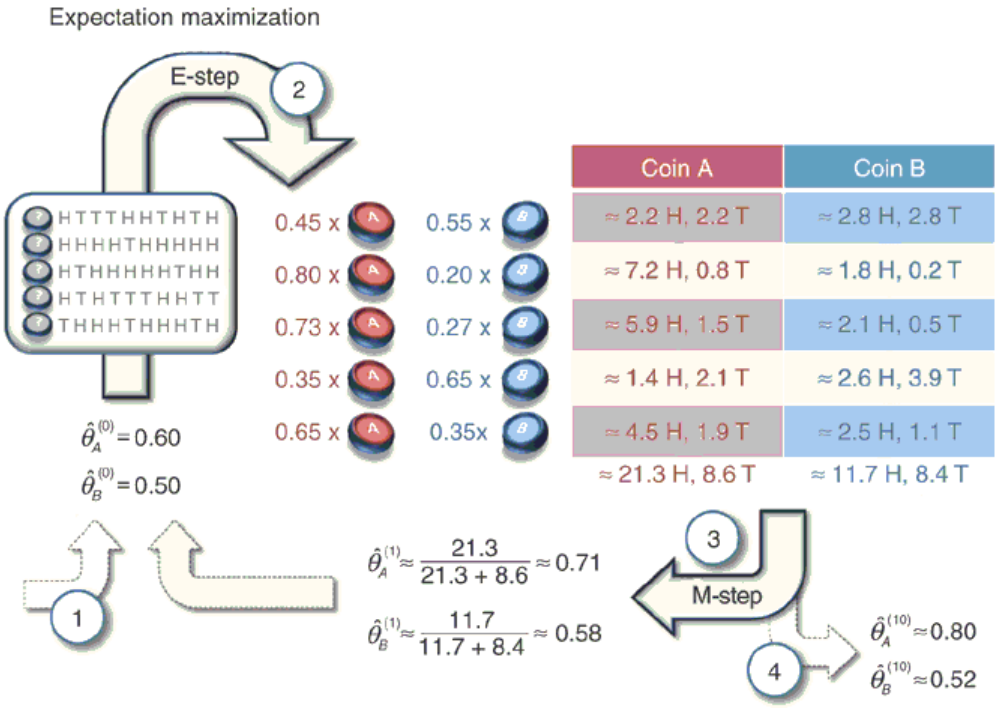


Figure 3.2: Example of using the EM algorithm to estimate the probabilities of each coin generating the observations of heads and tails. Extracted from (DO; BATZOGLOU, 2008).

consider the second row of the table on the right, where 9 "heads" and 1 "tails" were observed. Using Equation 3.6, we get the following probability values for coins A and B:

$$p_A(9) = \binom{10}{9} \theta_A^9 (1-\theta_A)^{10-9} \approx 0.04, \tag{3.7}$$

$$p_B(9) = \binom{10}{9} \theta_B^9 (1-\theta_B)^{10-9} \approx 0.01. \tag{3.8}$$

Normalizing it,

$$p_A = \frac{0.4}{0.04+0.01} = 0.80, \tag{3.9}$$

$$p_B = \frac{0.01}{0.04+0.01} = 0.20. \tag{3.10}$$

These two values (0.80 and 0.20) are the ones that appear on the left side of the second row in stage 2. Using the observations made in this particular set (9 "heads" and 1 "tails"), we can calculate the values of the pair ("heads," "tails") for coins A and B as ($9 \times 0.8 = 7.2$, $1 \times 0.8 = 0.8$) and ($9 \times 0.2 = 1.8$, $1 \times 0.2 = 0.2$). This same rationale is repeated for the other sets.

In stage 3, the maximization step (or M step) consists of calculating the maximum likeli-

hood estimation of parameters θ_A and θ_B in order to update their values. This is done based on the proportion of "heads" observed when flipping the coins. So for coin A, the new value of θ_A (say $\hat{\theta_A}$) is

$$\hat{\theta_A} = \frac{21.3}{21.3+8.6} = 0.71. \tag{3.11}$$

Similarly, the new value for θ_B (say $\hat{\theta_B}$) is

$$\hat{\theta_B} = \frac{11.7}{11.7+8.4} = 0.58. \tag{3.12}$$

These updated values are then used as input for another iteration of the EM algorithm. This process continues until convergence, i.e., when the difference between θ_A and $\hat{\theta_A}$ and θ_B and $\hat{\theta_B}$ in two consecutive iterations is smaller than ε. In Figure 3.2, the final values were estimated as $\theta_A \approx 0.80$ and $\theta_B \approx 0.52$, as shown in stage 4. It is important to note that in this example the values of θ_A and θ_B were initially unknown, whereas the observations of each flip were not.

Formally, consider a full set of samples $X = \{x_1, \ldots, x_n\}$ extracted from a single distribution. Suppose, however, that part of the data is missing. This way, each individual sample can be described as $x_k = \{x_{kp}, x_{km}\}$, where x_{kp} are the present data and x_{km} are the missing data. For ease of notation, we can separate these two kinds of data (present and missing) into two sets as X_p for the present and X_m for the missing in a way that the complete data is $X = X_p \cup X_m$. From this point, we can write the following function (DUDA; HART; STORK, 2000)

$$Q(\theta; \theta^t) = E_{X_m}[\ln p(X_p, X_m; \theta) | X_p; \theta^t], \tag{3.13}$$

where t is the number of the current iteration and the semi-colon (";") is used on the left side of Equation 3.13 to indicate that $Q(\theta; \theta^t)$ is a function of θ with θ^t fixed. On the right-hand side of Equation 3.13, the semi-colon indicates that the expected value is calculated regarding the missing data, assuming parameters θ^t fully describe the distribution.

A simple way to interpret Equation 3.13 is the following. The parameter vector θ^t is taken as the best estimation for the full distribution as far as iteration t, and θ is a candidate vector that may offer an even better estimation. Given that, the data likelihood is calculated on the right-hand side of Equation 3.13, including the missing data X_m, marginalized to the best description of the distribution at the moment (given by θ^t). Different candidate vectors θ will output different likelihood values. The EM algorithm selects the best one, which then becomes θ^{t+1}. This new vector is related to the maximum value of function $Q(\theta; \theta^t)$. Letting C be a convergence criterion, this process will be repeated until $Q(\theta^{t+1}; \theta^t) - Q(\theta^t; \theta^{t-1}) \leq C$ or a given number of iterations is reached. More information about the EM algorithm is described

in Appendix A.

In this work, the notation of the complete data vector is given by

$$\mathbf{C}_j = (\mathbf{x}_1, \ldots, \mathbf{x}_M, z_1, \ldots, z_N, u_1, \ldots, u_N)^T, \tag{3.14}$$

where $z_1, \ldots, z_N$ are the labels of each sample vector $\mathbf{x}_j$ and z_{ij} can assume a value between 0 and 1, depending on whether or not sample $\mathbf{x}_j$ was generated by the i-th mixture component. Given the definition of the Student's t-distribution in Equations 3.1, 3.2 and 3.3, it is important to note that the augmented data z_i, with $i = 1, \ldots N$, are still incomplete, since the covariance matrix of each component depends on the degrees of freedom (GEROGIANNIS; NIKOU; LIKAS, 2009). For this reason, the complete data vector also includes $u_1, \ldots, u_N$.

For the $t+1$-th iteration of the algorithm, the E step is used to calculate the posterior probability of sample $\mathbf{x}_j$ belonging to the i-th component of the mixture model

$$z_{ij}^{t+1} = \frac{w_i^t p\left(\mathbf{x}_j; \mu_i^t, \Sigma_i^t, \nu_i^t\right)}{\Sigma_{m=1}^K w_m^t p\left(\mathbf{x}_j; \mu_m^t, \Sigma_m^t, \nu_m^t\right)}, \tag{3.15}$$

and the expectation of the weights for each observation as

$$u_{ij}^{t+1} = \frac{\nu_i^t + d}{\nu_i^t + \delta\left(x; \mu_i^t, \Sigma_i^t\right)}. \tag{3.16}$$

We incorporated neighborhood information in our model as proposed by (NGUYEN; WU, 2012; ZHANG; WU; NGUYEN, 2013; GREBOL, 2013). Let v be a voxel and let N_v be the neighborhood of v with radius r and size $s = (2 \times r + 1)^{dim}$, where dim is the number of dimensions of the image (in the context of 3D images, $dim = 3$). Then, in Equation 3.15, w_i^t is calculated as

$$w_i^t = \frac{\pi_i^t + n_{iq}^t}{2}, \tag{3.17}$$

where

$$n_{iq}^t = \frac{\pi_i^t p\left(\mathbf{x}_j; \mu_i^t, \Sigma_i^t, \nu_i^t\right)}{\Sigma_{m=1}^K \pi_m^t p\left(\mathbf{x}_j; \mu_m^t, \Sigma_m^t, \nu_m^t\right)} \tag{3.18}$$

and $\mathbf{X}_q \in N_v$. This is the mean field approximation of a Markov Random Field (ZHANG, 1992). In other words, for each sample $\mathbf{x}_j$, w_i^t was calculated as the average between the ith component weight and the posterior probability of the neighborhood of $\mathbf{x}_j$ belonging to the ith component as well.

In the M step, the maximization of the log-likelihood of the complete data provides the

update equations of the mixture model parameters as

$$\pi_i^{t+1} = \frac{1}{N}\Sigma_{j=1}^{N} z_{ij}^t, \tag{3.19}$$

$$\mu_i^{t+1} = \frac{\Sigma_j^N z_{ij}^t u_{ij}^t \mathbf{x}_j}{\Sigma_j^N z_{ij}^t u_{ij}^t}, \tag{3.20}$$

$$\Sigma_i^{t+1} = \frac{\Sigma_{j=1}^N z_{ij}^t u_{ij}^t \left(\mathbf{x}_j - \mu_i^{t+1}\right)\left(\mathbf{x}_j - \mu_i^{t+1}\right)^T}{\Sigma_{j=1}^N z_{ij}^t}. \tag{3.21}$$

The degrees of freedom ν_i^{t+1} for the i-th component in iteration $t+1$ are calculated in an iterative manner as the solution for the following equation

$$\begin{aligned} \log\left(\frac{\nu_i^{t+1}}{2}\right) - \psi\left(\frac{\nu_i^{t+1}}{2}\right) + 1 - \log\left(\frac{\nu_i^t + d}{2}\right) + \\ \frac{\Sigma_{j=1}^N z_{ij}^t \left(\log u_{ij}^t - u_{ij}^t\right)}{\Sigma_{j=1}^N z_{ij}^t} + \psi\left(\frac{\nu_i^t + d}{2}\right) \quad = \quad 0, \end{aligned} \tag{3.22}$$

where $\psi(x) = \frac{\partial(\ln\Gamma(x))}{\partial x}$ is the Digamma function.

At the end of the estimation we have the parameters for the mixture model, thus allowing us to move to the clustering step.

3.2.1 Tissue clustering

After estimating the parameters of the Student's t-mixture model, the clustering stage is used to determine the probabilistic mapping between samples and a set of labels in a way that similar samples can be grouped together (DUDA; HART; STORK, 2000).

The process of clustering each sample to its label z (tissue cluster) is done by associating the samples to the cluster that maximizes the posterior probability $P(z_k|\mathbf{x})$. In other words, given a sample vector $\mathbf{x} = (x_1, \ldots, x_N)$, a cluster z_k with the highest posterior probability for $\mathbf{x}$ will be attributed to the sample. This clustering procedure follows the Bayesian Maximum a Posterior (MAP) decision rule, given by

$$P(z_i|\mathbf{x}) \underset{z_j}{\overset{z_i}{\gtrless}} P(z_j|\mathbf{x}), \tag{3.23}$$

where $i = 1, \ldots, K,\ j = 1, \ldots, K$ and $i \neq j$.

Both sides of the decision rule in Equation 3.23 come from the Bayes theorem (DUDA; HART; STORK, 2000)

$$P(z_k|\mathbf{x}) = \frac{p(\mathbf{x}|z_k;\mu,\Sigma,\nu)p(z_k)}{\sum_{j=1}^{K} p(\mathbf{x}|z_j;\mu,\Sigma,\nu)p(z_j)}, \tag{3.24}$$

where $p(z_k)$ is the prior probability and $p(\mathbf{x}|z_k;\mu,\Sigma,\nu)$ is the the Student's t-probability density function (PDF) for cluster k. Parameter K indicates the number of components in the mixture model.

By the end of this process, each voxel in the image has a cluster associated with it, thus generating a segmented image.

3.3 Supervised learning

The goal of identifying which lesions are active or not can be seen as a classification task. There are many approaches that can be used in this scenario, such as probabilistic classifiers, logistic regression, support vector machines and decision trees (FATIMA; PASHA, 2017). There are also deep learning techniques that have been readily adopted in the medical imaging field with promising results (LITJENS et al., 2017), but their caveat is that they require a large dataset to be trained on.

When working on classification tasks, two important aspects have to be taken into account: bias and variance. We did not have access to a large number of patients, which made deep learning techniques, which often have thousands of parameters to estimate, unfeasible. The model would overfit the problem, leading to low bias and high variance. On the other hand, given the nature of our problem, non-deep learning techniques could potentially miss intrinsic relationships among features, leading to high bias and low variance - underfitting. To circumvent this tradeoff, we decided to use an ensemble tree-based classifier (DUDA; HART; STORK, 2000). Ensemble classifiers are known for using a combination of learners to improve the classification outcome. There are two main ensemble approaches regarding model construction: bagging and boosting.

A well-known ensemble tree-based classifier is random forest (DUDA; HART; STORK, 2000). It uses random sampling - also known as bootstrapping - with replacement, meaning some observations may be repeated in each new training dataset, to independently build N learners with various depths. Each tree then outputs a class for new samples, and the class with the majority of votes is assigned to the sample itself. **B**ootstrapping the data plus using the **agg**regate to make a decision is an example of bagging. An example of a random forest classifier is shown

in Figure 3.3.

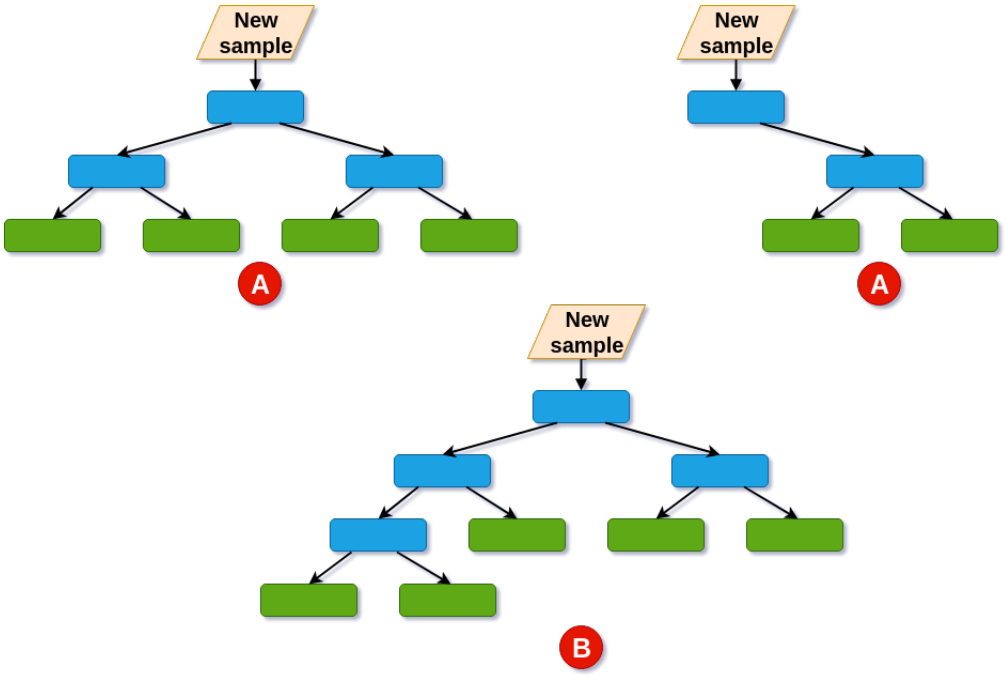


Figure 3.3: Random forest example. Each leaf is colored in green. Assuming the training dataset was comprised of classes A and B and we wanted to classify a new sample, the majority class for this example would be A. Note that each tree is independently built and all trees have different depths.

On the other hand, gradient tree-based boosting algorithms start with a very simple learner - just a leaf - and iterates using the residual errors of previous steps as targets for training. They also make use of bootstrapping, except that each new training dataset is not randomly built: weights are assigned to every sample and updated on every iteration as a way of choosing the observations that were not correctly classified in the previous steps. An example of a gradient tree-based boosting classifier is shown in Figure 3.4.

Both approaches reduce variance and provide more stable models. However, boosting inherently reduces bias as well (DUDA; HART; STORK, 2000). That is why we decided to use XGBoost (CHEN; GUESTRIN, 2016), a boosting-based algorithm which works by reducing both variance and bias to classify active and nonactive lesions. Details of this classifier are given in the next section.

3.3.1 XGBoost classifier

The classifier used in this work was XGBoost (CHEN; GUESTRIN, 2016), a gradient tree boosting algorithm. The main goal of boosting is to improve the accuracy of a given learning algorithm by using an ensemble of weak learners whose joint decision rule has an arbitrarily

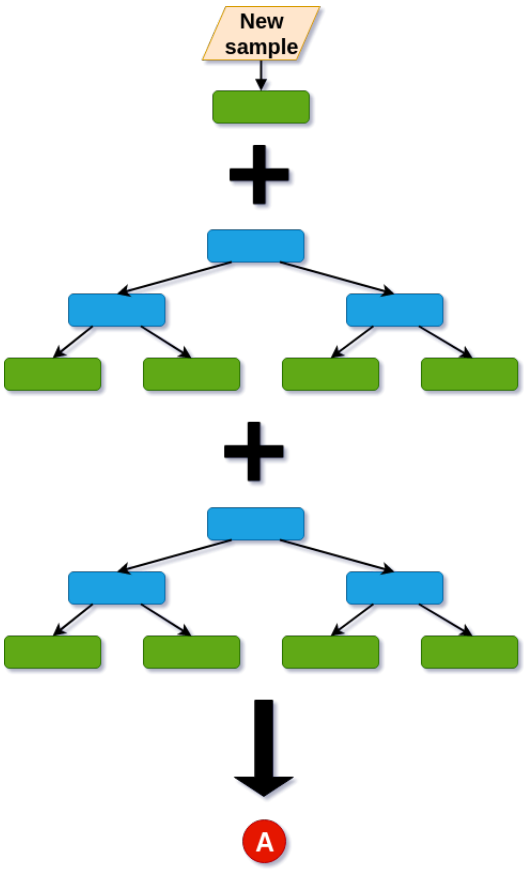


Figure 3.4: Gradient boosting example. Each leaf is colored in green. Assuming the training dataset was comprised of classes A and B and we wanted to classify a new sample, the class for this example would be the combination of the trees. Apart from the first one, note that every other tree has the same depth and they are combined to get the actual class output.

high accuracy on the training set (DUDA; HART; STORK, 2000). We chose XGBoost due to its remarkable performance in classification contests (BENNETT; LANNING, 2007) and optimized structure that makes it scalable.

Mathematically, given a dataset with n samples, indexed by $i = 1 \ldots n$, and m features $D = \{(\mathbf{x}_i, y_i))\}$, where $|D| = n$, $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$, a tree ensemble model is comprised of the sum of K functions to predict the output

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \tag{3.25}$$

where $f_k \in \mathscr{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}$ with $q : \mathbb{R}^m \to T$ and $w \in \mathbb{R}^T$ is the space of trees. In this context, q is related to the arrangement of each tree that maps a sample to its corresponding leaf index and T is the number of leaves in a tree. Therefore, each f_k is related to one particular tree q and leaf weights w.

In order to learn the functions that will describe the model itself, the following loss function

is used as the objective

$$\mathscr{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{3.26}$$

where

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{3.27}$$

l is a differentiable loss function and Ω is a penalty term to avoid selecting complex models. The goal is to keep it simple and highly predictive at the same time. We can think of gradient boosting in the following way:

1. Fit a model to the data, $\mathscr{F}_1(\mathbf{x}) = y$.
2. Fit a model to the residuals (i.e., the loss), $h_1(\mathbf{x}) = y - \mathscr{F}_1(\mathbf{x})$.
3. Create a new model, $\mathscr{F}_2(\mathbf{x}) = \mathscr{F}_1(\mathbf{x}) + h_1(\mathbf{x})$.
4. Repeat steps (1)-(3) until a number M of trees or a sufficiently small difference between loss functions of subsequent steps is reached.

In other words, the algorithm trains successive component classifiers and the output for a test sample $\mathbf{x}$ is based on the outputs of these very same components. The scalability of XGBoost is achieved by analyzing and optimizing cache access patterns, data compression and sharding, which are described in more detail in (CHEN; GUESTRIN, 2016). Regarding parameters, we set the maximum depth of each tree to 1 and the number of estimators to 5000 in order to avoid specialization of learners and provide enough of them to get an accurate joint decision rule.

3.4 Textures as features for classification

Texture is an intuitive concept that can be defined as a series of homogeneous visual patterns that are observed in a wide range of materials and surfaces (HALL-BEYER, 2017). Textures are usually described as fine, coarse, granulated or smooth. The quantification of textures has been studied for decades, and there are several different techniques available to extract and describe them in a mathematical manner. In this work, we used two well known texture algorithms to extract relevant features and aid in the distinction of active and nonactive multiple sclerosis lesions. We opted for these techniques due to their wide applicability in medical images and their simple, yet powerful ability to distinguish pathologies (CASTELLANO et al., 2004; ZHANG et al., 2008; ROY; BHUIYAN; RAMAMOHANARAO, 2013). A description of both these techniques are presented in the following sections.

3.4.1 Gray Level Co-Occurrence Matrix

Gray Level Co-Occurrence Matrix (GLCM) was first introduced by (HARALICK; SHANMUGAM; DINSTEIN, 1973) and has been widely used in MR image segmentation and classification problems ever since (LOIZOU et al., 2014; MICHOUX et al., 2015; ARDAKANI et al., 2017; DOYLE et al., 2018). GLCM computes the joint probability of observing two voxels with co-occurring gray level intensities within a given distance d and direction θ. To better understand how GLCM works, a schematic for a 2D example is shown in Figure 3.5, where the outer orange rows and columns represent the i and j indexes of the matrices.

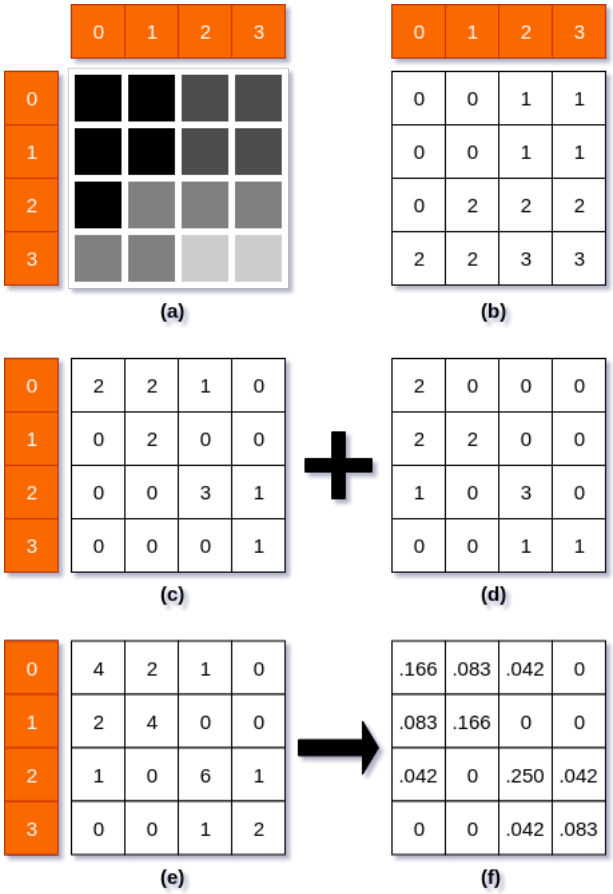


Figure 3.5: Example of how a GLCM matrix is created. (a) Reference image, (b) Gray level intensities from reference image, (c) Eastern GLCM, (d) Western GLCM, (e) Horizontal GLCM, (f) Normalized GLCM.

Figures 3.5 (a)-(b) represent an image and its gray level intensities, respectively. To get textural information, we need to define an ordering (i.e., left to right, right to left, up to down and down to up), a direction (usually 0, 45, 90 and 135 degrees) and look at the spatial relationship between neighboring pixels. In this example we will focus on the horizontal spatial relationship (left to right and right to left - also called eastern and western) within a pixel distance of $d = 1$, meaning we will simply look at the co-occurrence of gray levels that are 1 pixel apart in

the horizontal direction. In this sense, Figures 3.5 (c)-(d) represent the number of times a pair of gray level intensities were observed 1 pixel apart - from left to right and right to left, respectively. For instance, cell $(0,1)$ in Figure 3.5 (c) holds the number of times a gray level of 1 occurred to the right of a gray level of 0 in our reference image, whereas cell $(0,1)$ in Figure 3.5 (d) holds the number of times a gray level of 1 occurred to the left of a gray level of 0 in the same reference image.

To get the full horizontal spatial relationship we sum the eastern and western matrices (which are the transpose of one another), resulting in the matrix shown in Figure 3.5 (e). To get the actual joint probability, we simply normalize this matrix by the sum of all of its elements, finally getting to the normalized GLCM in Figure 3.5 (f). Each cell in the normalized GLCM represents the probability of observing a particular pair of gray level values occurring in a specific combination of direction and distance.

With the normalized GLCM in hand, we can derive a number of features to quantitatively explore the textural characteristics of a given image. In this work, we extracted the following eight features from each GLCM. In the following computations:

- $g(i,j)$ is the element in cell (i,j) of the normalized GLCM;
- $\mu = \sum_{i,j} i \cdot g(i,j) = \sum_{i,j} j \cdot g(i,j)$ is the weighted pixel/voxel average;
- $\sigma = \sum_{i,j} (i-\mu)^2 \cdot g(i,j) = \sum_{i,j} (j-\mu)^2 \cdot g(i,j)$ is the weighted pixel/voxel variance.

1. Energy: measures the local uniformity of texture. The higher the energy value, the bigger the uniformity and organization of the texture.

$$\text{Energy} = \sum_{i,j} g(i,j)^2 \tag{3.28}$$

2. Entropy: expresses the level of organization of a texture. A completely random distribution of gray level intensities in the image volume would have very high entropy, while an image with the same gray level across all pixels/voxels would have very low value of entropy.

$$\text{Entropy} = \begin{cases} \sum_{i,j} g(i,j) \log_2 g(i,j) & \text{if } g(i,j) \neq 0 \\ 0 & \text{if } g(i,j) = 0 \end{cases} \tag{3.29}$$

3. Correlation: measures the linear dependency of gray level values in the co-occurrence matrix.

$$\text{Correlation} = \sum_{i,j} \frac{(i-\mu)(j-\mu) g(i,j)}{\sigma^2} \tag{3.30}$$

4. Inverse difference moment (IDM): measures the homogeneity of the image. IDM will be low for inhomogeneous images and high for homogeneous images.

$$\text{IDM} = \sum_{i,j} \frac{1}{1+(i-j)^2} g(i,j) \tag{3.31}$$

5. Inertia (or contrast): measures local gray level variation in the GLCM matrix. If the neighboring pixels/voxels in the texture are very similar in their gray level values, then the contrast in the image is very low. Contrast is zero for a constant image.

$$\text{Contrast} = \sum_{i,j} (i-j)^2 g(i,j) \tag{3.32}$$

6. Cluster shade: feature of skewness of the matrix and is linked to the perception of uniformity in an image. When this feature is high, the image is asymmetric.

$$\text{Cluster shade} = \sum_{i,j} ((i-\mu)+(j-\mu))^3 g(i,j) \tag{3.33}$$

7. Cluster prominence: this feature is also related to the perceptual symmetry of an image. When the cluster prominence value is high, the image is less symmetric.

$$\text{Cluster Prominence} = \sum_{i,j} ((i-\mu)+(j-\mu))^4 g(i,j) \tag{3.34}$$

8. Haralick correlation: original correlation measure designed by Haralick in 1973. Measures the linear dependence between pixels/voxels relative to each other.

$$\text{Haralick} = \frac{\sum_{i,j} (i,j)\, g(i,j) - \mu^2}{\sigma^2} \tag{3.35}$$

The calculation of the contrast feature is shown in Figure 3.6. The matrix depicted in Figure 3.6 (a) is simply a representation of the $(i-j)^2$ part of Equation 3.32, and the normalized GLCM is shown in Figure 3.6 (b). We perform a cell-wise multiplication of both matrices, getting to the contrast matrix shown in Figure 3.6 (c). Finally, to get the actual image contrast we sum all the contrast matrix cells, thus yielding a value of 0.586.

Of course, the example used in Figure 3.5 (a) is small for the sake of simplification. In real world problems we extract textural features from windows that traverse the whole image, as shown in Figure 3.7. Each center pixel/voxel of a window is then assigned a different value according to each texture, thus creating different texture maps (i.e., contrast map, energy map, entropy map and so on). The way each of these values are calculated, however, follow the exact same rationale shown in Figures 3.5 and 3.6.

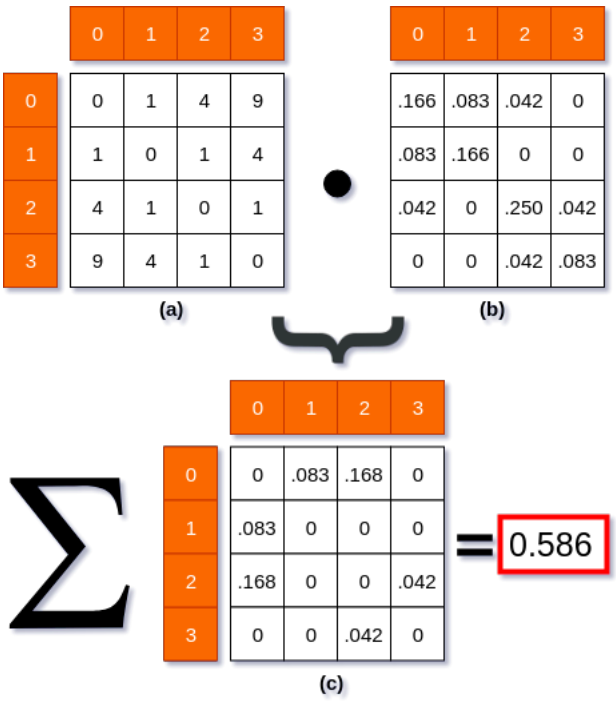


Figure 3.6: Example of how the GLCM contrast feature is calculated. (a) Part of contrast equation, (b) Normalized GLCM, (c) Contrast matrix and sum.

3.4.2 Run Length Matrix

The Run Length Matrix (RLM) (CASTELLANO et al., 2004) counts the number of times two (or more) voxels with the same gray level intensity occur in a given direction and orientation. Features from RLM are mostly related to the fineness and coarseness of a given image represented by long runs and short runs, respectively. To better understand how RLM works, a schematic for a 2D example is shown in Figure 3.8, where the outer orange rows and columns represent the indexes of the matrices and the purple rows represent the length of each run.

Figures 3.8 (a)-(b) represent an image and its gray level intensities, respectively. The same ordering and direction mentioned in the previous section regarding GLCM apply to RLM as well. In this example we will also focus on the horizontal spatial relationship within a pixel distance of $d = 1$. In this sense, Figure 3.8 (c) represents the number of times each run length occurred for a given gray level intensity. For instance, cell $(0,1)$ holds the number of times a gray level of 0 was observed in a run length of 1 (i.e., the number of times the gray level intensity itself occurred in the image), whereas cell $(0,2)$ holds the number of times a gray level of 0 was observed in a run length of 2 (i.e., the number of times a pair of zeros occurred in the image). Note that we no longer refer to right to left or left to right; once we define a direction - in our example, horizontal (or 0 degree) - we solely count the number of times each run length occurs for each gray level intensity in that particular direction. Finally, we normalize

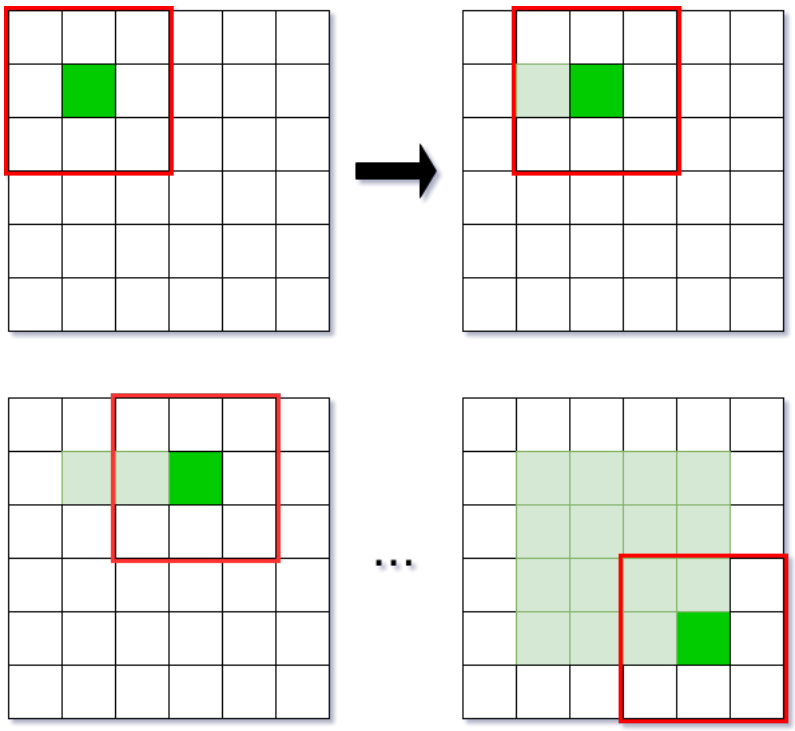


Figure 3.7: Example of a 3×3 moving window in a 2D case.

the RLM matrix by the sun of all of its entries, yielding the matrix shown in Figure 3.8 (d).

We extracted ten features from each RLM. In the following computations, $g(i,j)$ is the element in cell (i,j) of the normalized RLM, i is related to the pixel/voxel intensity and j to the length of the run:

1. Short run emphasis (SRE): measures the distribution of short runs. SRE is expected large for fine textures.

$$\text{SRE} = \sum_{i,j} \frac{g(i,j)}{j^2} \tag{3.36}$$

2. Long run emphasis (LRE): is a feature that measures distribution of long runs. LRE is usually large for coarse structural textures.

$$\text{LRE} = \sum_{i,j} g(i,j)\, j^2 \tag{3.37}$$

3. Gray level non uniformity (GLN): measures the similarity of gray level values throughout the texture. GLN is usually small if the gray level values are alike across the whole texture.

$$\text{GLN} = \sum_{i} \left(\sum_{j} g(i,j) \right)^2 \tag{3.38}$$

4. Run length non uniformity (RLN): measures the similarity of length of runs throughout

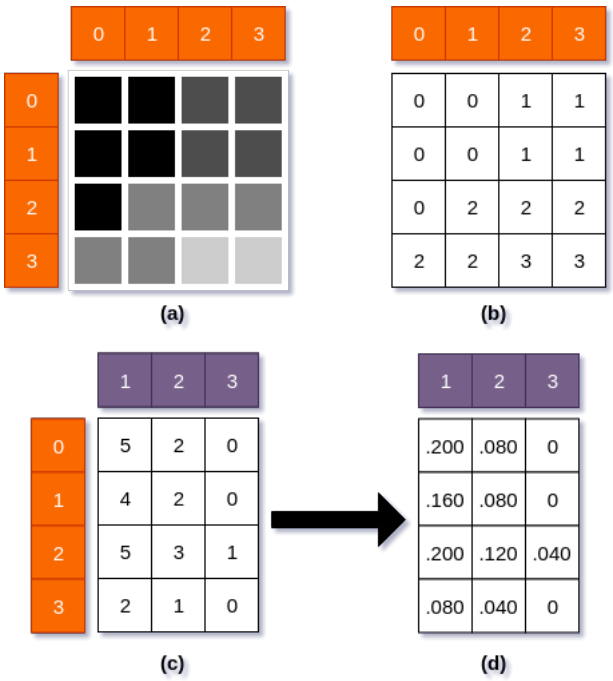


Figure 3.8: Example of how a RLM matrix is created. (a) Reference image, (b) Gray level intensities from reference image, (c) RLM with run lengths 1, 2 and 3 and (d) Normalized RLM.

the image. RLN tends to be small if the run lengths are alike across the image.

$$\mathrm{RLN} = \sum_j \left(\sum_i g(i,j) \right)^2 \tag{3.39}$$

5. Low gray level run emphasis (LGRE): this feature is orthogonal to SRE and its value increases when the texture is dominated by many runs of low gray value.

$$\mathrm{LGRE} = \sum_{i,j} \frac{g(i,j)}{i^2} \tag{3.40}$$

6. High gray level run emphasis (HGRE): this feature is orthogonal to LRE and increases when the texture is dominated by many runs of high gray value.

$$\mathrm{HGRE} = \sum_{i,j} g(i,j)\, i^2 \tag{3.41}$$

7. Short run lowgray level emphasis (SRLGE): this is a diagonal measurement that combines SRE and LGRE. It increases when the texture is dominated by many short runs of low gray value.

$$\mathrm{SRLGE} = \sum_{i,j} \frac{g(i,j)}{i^2 j^2} \tag{3.42}$$

8. Short run high gray level emphasis (SRHGE): this feature is orthogonal to SRLGE and

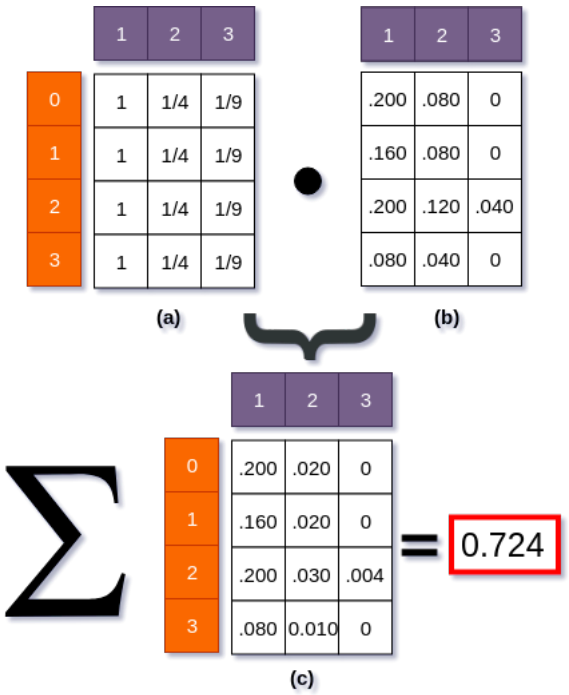


Figure 3.9: Example of how the RLM SRE feature is calculated. (a) Part of SRE equation, (b) Normalized RLM, (c) SRE matrix and sum.

LRHGE and increases when the texture is dominated by short runs with high intensity levels.

$$\text{SRHGE} = \sum_{i,j} \frac{g(i,j)\, i^2}{j^2} \tag{3.43}$$

9. Long run low gray level emphasis (LRLGE): complementary to SRHGE and increases when the texture is dominated by long runs that have low gray levels.

$$\text{LRLGE} = \sum_{i,j} \frac{g(i,j)\, j^2}{i^2} \tag{3.44}$$

10. Long run high gray level emphasis (LRHGE): complementary to SRLGE and increases with a combination of long and high gray level runs.

$$\text{LRHGE} = \sum_{i,j} g(i,j)\, i^2 j^2 \tag{3.45}$$

The calculation of the SRE feature is shown in Figure 3.9. The matrix depicted in Figure 3.9 (a) is simply a representation of the $\frac{1}{(j)^2}$ part of Equation 3.36, and the normalized RLM is shown in Figure 3.9 (b). We perform a cell-wise multiplication of both matrices, getting to the SRE matrix shown in Figure 3.9 (c). Finally, to get the actual image SRE we sum all the SRE matrix cells, thus yielding a value of 0.724.

As mentioned in Section 3.4.1, there are two main parameters regarding textures: distance

and orientation. In 2D, four orientations (or offsets) are commonly used when working with textures (0, 45, 90 and 135 degrees). In 3D, the number of offsets scales analogously to 13. The distance parameter defines the window size that will traverse the image. Smaller windows capture finer details at the expense of global information, and vice versa. Examples of 2D and 3D directions with distance $d = 1$ are shown in Figure 3.10.

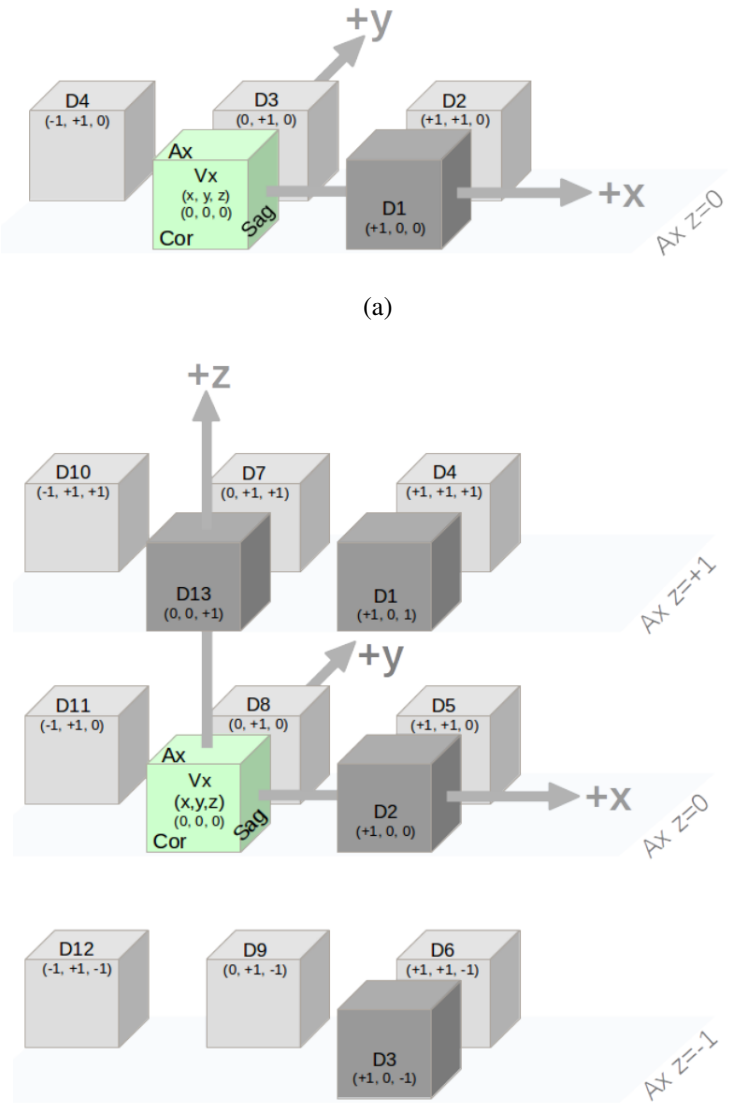


Figure 3.10: Number of directions for (a) 2D and (b) 3D with distance $d = 1$. Extracted from (ORLHAC; NIOCHE; BUVAT, 2019).

3.5 Metrics used for assessment

In this section we present the metrics we used to assess the segmentation, feature enhancement and classification steps.

3.5.1 Segmentation metrics

Five different metrics were used to assess the spatial and volumetric agreement between the segmentation performed by our proposal and the manual delineations done by experts. The metrics were the Dice Similarity Coefficient (DSC) (DICE, 1945), true positive rate (TPR), false positive rate (FPR), volume difference (VD) and Pearson's correlation coefficient (GAYEN, 1951).

The DSC metric is defined as the ratio between the number of voxels assigned as MS lesion by the automatic segmentation and the expert and the average number of voxels assigned as a lesion in both methods. According to Bartko (1991), a DSC ≥ 0.7 indicates a good spatial agreement between images.

More formally, the DSC, TPR, FPR and VD are mathematically described in Table 3.1.

Metric	Acronym	Equation	Interval
Dice Similarity Coefficient	DSC	$\frac{2\times TP}{FP+FN+2\times TP}$	$[0,1]$
True Positive Rate	TPR	$\frac{TP}{TP+FN}$	$[0,1]$
False Positive Rate	FPR	$\frac{FP}{FP+TP}$	$[0,1]$
Volume Difference	VD	$\lvert 1-\frac{V_{seg}}{V_{GT}}\rvert$	$[0,1]$

Table 3.1: Metrics used to quantitatively assess the automatic segmentation. TP stands for true positives, FP stands for false positives, FN stands for false negatives, V_{seg} is the volume of the segmentation resulting mask and V_{GT} is the volume of the expert's manually delineated mask.

Finally, the Pearson's correlation coefficient (BLAND, 2000) was used to analyze the volumetric agreement between the automatic segmentation and the manual delineation. The value of this particular metric lies in the interval $[-1,+1]$ and it is defined as a linear correlation metric between two variables. Values of -1 and $+1$ indicate complete negative correlation and complete positive correlation, respectively, while a value of 0 indicates that there is no correlation between the two variables being analyzed. The coefficient formula, commonly called r, is given by

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{3.46}$$

where n represents the number of samples, x_i and y_i represent the variables and $\bar{x}$ and $\bar{y}$ are their means, respectively.

A metric that is commonly associated with the correlation coefficient if the *p-value*, that helps decide if the observed data have statistical significance. The *p-value* is associated with

the following question: if there is no correlation between variables x and y, what is the chance that a random sampling would result in a correlation coefficient as far from zero as observed in the experiment? If the *p-value* is small, we can reject the idea that the correlation was caused by random events (HUBBARD, 2004). As a comparison parameter, a significance level threshold was defined as $\alpha = 0.01$. This way, if *p-value* $< \alpha$, then it is possible to state that a random sampling does not cause the observed values and, consequently, the correlation stands true (NUZZO, 2014).

3.5.2 Feature enhancement metrics

Regarding the image enhancement process and white matter mask estimation mentioned in Chapter 1, we used other four metrics to quantitatively assess how well our approach was in enhancing MS lesions. The first one was the intensity profile distinction (IPD), calculated as

$$IPD = \left(\frac{\text{average(lesions)}}{\text{average(tissue)}} - 1 \right) \times 100, \tag{3.47}$$

in which we simply divide the average intensity of lesions by the average intensity of a tissue of interest (white matter or gray matter) in a particular image and then scale it in terms of percentage. Thus, we can determine how brighter, percent-wise, the lesion cluster is compared with other tissues.

We also used the well-known sensitivity and specificity metrics to measure the accuracy of our white matter mask estimation. Sensitivity (SS) is defined as

$$SS = \frac{TP}{(TP + FN)} \tag{3.48}$$

and in the context of this work measures the proportion of WM and lesion voxels that were correctly identified as such during the estimation.

Specificity (SP) is defined as

$$SP = \frac{TN}{(FP + TN)}, \tag{3.49}$$

where TN represents the number of true negatives. In this work, SP measures the proportion of non-WM and non-lesion voxels that were correctly identified as such. Both SP and SS fall in the interval $[0, 1]$ and similar to DSC, the closer they are to 1, the better.

Another metric we used to assess the white matter mask estimation was the lesion intersection (LI), defined as

$$LI = \frac{|\text{Lesion}_{GT} \cap \text{Mask}_{estim}|}{|\text{Lesion}_{GT}|} \times 100. \tag{3.50}$$

Similar to IPD, LI is also calculated in terms of percentage and provides a quantitative tool to analyze the lesion load that was kept during the white matter mask estimation.

3.5.3 Classification metrics

Finally, to assess the classification of active and nonactive voxel lesions we used the specificity, sensitivity and F1-score metrics defined in Table 3.2.

Metric	Evaluation
Specificity	$\frac{TP}{TP+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
F1-score	$2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

Table 3.2: Metrics used to assess classification of active and nonactive voxel lesion. TP, FP and FN are true positives, false positives and false negatives, respectively.

The ratio of active to nonactive lesions in our databases (which will be described in more detail in Chapter 5) was roughly 1:3, i.e., the active voxel class had a representativity of approximately 25%. Given this imbalance and the paramount importance of correctly detecting active lesions, the sensitivity and F1-score metrics bear more meaning to the assessment of classification performance. This is due to the fact that the former is related to the actual ground truth and the latter provides a better grasp on the overall performance - which can be distorted by the uneven distribution among both classes of interest.

We also conducted a lesion-level analysis. To do so, the active and nonactive voxels were grouped into connected components, thus indicating how many actual lesions were present in each patient. Given the classification probability of each lesion voxel, we calculated the average probability of any given lesion component being active or nonactive. Let C be a lesion component and p the probability of a given lesion voxel inside C. If the average probability of component C was equal to or greater than 50%, we ranked it as being active; otherwise, we ranked it as nonactive, as

$$\text{Class} = \begin{cases} \text{Nonactive}, & \frac{1}{\|C\|}\sum p_{p\in C} < 0.5 \\ \text{Active}, & \text{otherwise.} \end{cases}. \tag{3.51}$$

3.6 Final considerations

This chapter presented detailed information about the Student's t-mixture model, the Expectation-Maximization (EM) algorithm and how the latter was used to estimate the parameters of the former, along with how the *maximum a posteriori* (MAP) rationale is used for the actual clustering process. Also, a description of the XGBoost classifier, Gray Level Co-Occurrence Matrix and Run Length Matrix texture algorithms was provided. Moreover, the metrics used for the assessment of segmentation, feature enhancement and classification were also presented and detailed.

As mentioned in previous chapters, we decided to use the Student's t-mixture model due to the distribution's intrinsic heavy tail, which makes it less susceptible to the effect of samples that are too far off from the cluster centroids in the feature space. And since a Student's t-distribution tends to a Gaussian distribution when $\nu \to \infty$, it becomes easy to generalize it and compare the results from both distributions if necessary.

Textures are commonly used features for image classification and segmentation, and we opted for using GLCM and RLM because of their simplicity and extensibility to a three-dimensional space. We leveraged their spatial and neighborhood information to distinguish active and nonactive multiple sclerosis lesions, which will be further explored in the next chapters.

Chapter 4

LITERATURE REVIEW

This chapter presents a comprehensive review of works in the literature focused on the automatic segmentation and classification of multiple sclerosis lesions.

4.1 Image segmentation using supervised and unsupervised techniques

Many works based on probability theory for MR image segmentation have been proposed in the literature in the past decades (BISHOP, 1995, 2006; LI, 2009; MCLACHLAN; PEEL, 2000; STANFORD, 1999). There is a wide range of approaches regarding automatic MS lesion segmentation and classification (GARCÍA-LORENZO et al., 2013; X. et al., 2012), mainly because lesions do not have any specific shape, size or location. Moreover, there are many different protocols for image acquisition, which poses an extra challenge and indicates that MS lesion segmentation and classification in MRI are challenging problems and are open to new approaches in the medical image processing field.

One of the most commonly used unsupervised techniques for segmentation purposes is the Gaussian mixture model (GMM) (JAIN; DUIN; MAO, 2000; TITTERINGTON; SMITH, 1985), which has a strong mathematical foundation and is very similar to the model explained in Chapter 3, Section 3.1. To estimate the parameters of a GMM, the EM algorithm (Chapter 3, Section 3.2) is widely used, but since it is an iterative technique used to estimate the parameters of a model via maximum log-likelihood estimation, it is likely to fall in local minima, making it sensitive to initialization. This caveat must be taken into consideration when using this approach (MOORE; MCCABE, 1999). Some strategies have been proposed in the literature to mitigate the initialization sensitivity of the EM algorithm and therefore minimize errors in image segmentation. Some alternatives include a) initializing the model multiple times and selecting the set of pa-

rameters that minimize a cost function (DUDA; HART; STORK, 2000), b) using the output of the K-Means algorithm (HARTIGAN; WONG, 1979) as input for the parameter estimation (ZHANG; WU; NGUYEN, 2013) or c) probabilistic anatomical atlases to calculate statistical distributions of brain tissues (FREIRE; FERRARI, 2016), for the specific problem of MR brain image segmentation.

As mentioned in Chapter 3, one of the main advantages of the Student's t-distribution compared to GMM is the fact that the former has a heavy tail, making it less susceptible to noise (CHATZIS; KOSMOPOULOS; VARVARIGOU, 2009), which tends to dislocate the center of a Gaussian distribution and overestimate its dispersion.

In Nguyen and Wu (2012), the authors proposed a Student's t-mixture model to segment brain tissues in MRI and incorporated spatial information using the Dirichlet distribution (BLEI; NG; JORDAN, 2003). They called their approach *Student's t-mixture model spatially constrained algorithm* (SMM-SC). The authors used the gradient descent technique (BISHOP, 2006) to estimate the parameters of their model. According to them, their approach was quantitatively better than seven other segmentation techniques: GMM (MCLACHLAN; PEEL, 2000), *Student's t-mixture model* (SMM) (PEEL; MCLACHLAN, 2000), *Mean Field Algorithm* (MEANF) (FORBES; PEYRAR, 2003), *Spatially Variant Finite Mixture Model* (SVFMM) (BLEKAS et al., 2005), *Modified Fuzzy C-Means* (MFCM) (AHMED et al., 2002), *Fuzzy and Noise Tolerant Adaptive Segmentation Method* (FANTASM) (PHAM, 2001) and *Fuzz C-Means based Hidden Markov Random Field* (HMRF-FCM) (CHATZIS; VARVARIGOU, 2008).

The spatial information, modeled by the Dirichlet distribution, incorporated the relationship between neighboring pixels into the parameters of the distribution itself and was represented as a linear smoothing filter. The classification of each pixel in the authors' proposal followed three steps: first, they analyzed its neighborhood (in a $n \times n$ window), then averaged the probability of each neighboring pixel belonging to a given class and, finally, used the neighborhood information along with the gray level intensity of the central pixel of the window to classify it.

The estimation of the model parameters (both for the Student's t and Dirichlet distributions) was made using the iterative gradient descent algorithm. The initial values were provided by the user and used as a starting point. Then, each iteration maximized a log-likelihood function and the whole process halted when a specific convergence value was met.

The authors used synthetic, simulated and real images to test the accuracy of their approach. The metrics used to quantitatively assess the results were the misclassification ratio (MCR) (ZHANG; SMITH; BRADY, 2001) and the Dice similarity coefficient (DSC) (DICE, 1945). The simulated images came from the BrainWeb database (AUBERT-BROCHE et al., 2006), and the

real ones came from the Center for Morphometric Analysis at Massachusetts General Hospital (Massachusetss General Hospital, 2012). An important point to highlight is that even though simulated and real images represented a volume (the human brain), the segmentation proposal was to perform the segmentation in a slice-by-slice manner, i.e., 2D. For every set of images, the author's approach had the lowest MCR average compared to other techniques.

One main advantage of their work was that they made use of a probability distribution that had few parameters to be estimated. Using a relatively simple method, they were able to get good MCR results and outperform the other segmentation approaches. However, the author's technique needed to know the number of clusters in the images beforehand. Also, the neighboring window size used to bring context information to the algorithm can have a significant influence on the segmentation outcome. Another limitation of this technique lied in the fact that the segmentation itself was performed in a slice-by-slice manner, which can interfere with the estimation of pathology volumes such as MS lesions.

In Zhang, Wu and Nguyen (2013), the authors used a weighted Student's t-mixture model with spatial and clustering information to perform image segmentation and mitigate classification errors caused by noise. The authors compared their results with five other techniques: SMM (PEEL; MCLACHLAN, 2000), *Fuzzy Local Information C-Means* (FLICM) (KRINIDIS; CHATZIS, 2010), SVFMM (BLEKAS et al., 2005), HMRF-FCM (CHATZIS; VARVARIGOU, 2008) and SMM-SC (NGUYEN; WU, 2012). Comparatively, the authors' approach, called *Weighted Student's t-mixture model* (WSMM), yielded the best quantitative results for the MCR metric and the probabilistic rand (PR) index (UNNIKRISHNAN; PANTOFARU; HEBERT, 2005).

According to the authors, their proposal differed from other approaches, including from their previous work (NGUYEN; WU, 2012), in two aspects. The first one was that every pixel now had its own prior probability, which was dependent on its neighborhood. Their approach considered the spatial relationship between neighboring pixels and the link between spatial information and clustering information, which was incorporated into the technique using a Markov Random Field (MRF) (ZHANG; SMITH; BRADY, 2001). The second aspect regarded the weight each neighboring pixel had on the central one; the farther they were from the center, the smaller was their influence.

The estimation of the model parameters was achieved using the EM algorithm. The K-Means algorithm was used for initialization. As mentioned before, this is a commonly used strategy to mitigate local minimum problems. EM was then used to refine the parameters to accurately describe the clusters in the images.

The authors applied their technique to both synthetic and real images (MARTIN et al., 2008).

As mentioned before, they used the MCR metric and the PR index to assess the segmentation quality. For every image, the WSMM algorithm outperformed the other techniques.

However, a significant point to highlight is that the authors used only 2D images. For this reason, it was not possible to state if their technique would perform well in 3D medical images. Moreover, even though the authors' proposal achieved the best results compared to other techniques, one can argue that the gains in accuracy were not significant enough to justify the increase in the model's complexity.

In Simões et al. (2013), the authors used a GMM with context information to segment WM hyperintensities in FLAIR images. They used 40 images from a trial conducted at the University Hospital of Essen, in Germany. The images were divided into three groups according to their lesion load: small load ($\leq 10cm^3$), moderate load ($> 10cm^3$ and $\leq 30cm^3$) and severe load ($> 30cm^3$). The authors compared their results with manual annotations provided by an expert and with other automatic segmentation techniques (KHADEMI; VENETSANOPOULOS; MOODY, 2012; CHEN; ZHANG, 2004).

First, the images were preprocessed for skull stripping and correction of inhomogeneities caused by bias field, which is a phenomenon that causes similar regions to appear with different intensities in MRI. Then the segmentation was performed using GMM with gray level intensity and context information. The parameters were estimated by the EM algorithm. In this case, the authors defined three clusters: cerebral spinal fluid (CSF), white matter/gray matter (WM/GM) and hyperintensities in white matter (WMH). At the end of the whole pipeline, each voxel was clustered into one of these three clusters.

Using DSC, overlap fraction (OF) and extra fraction (EF) (ANBEEK et al., 2004), the authors observed that their approach presented a high correlation with manual annotations (DSC values above 0.70 for moderate and severe load cases) and outperformed the other automatic techniques. Comparing to manual annotations, the average DSC values for the small load (18 images), moderate load (13 images) and severe load (9 images) were, respectively, 0.51, 0.70 and 0.84. DSC values above 0.70 indicate a good spatial correlation between the manual annotation and automatic segmentation (BARTKO, 1991). A DSC value below 0.70 for the set of the small load was expected, since in this case segmentation errors have a greater impact in the similarity metric, as already pointed by other studies (ADMIRAAL-BEHLOUL et al., 2005; ANBEEK et al., 2004; DYRBY et al., 2008).

To further assess their approach, the authors also applied their technique to 23 images from the MICCAI MS Lesion Segmentation Challenge 2008 (MICCAI MS Lesion Segmentation Team, 2008). In this database, the accuracy of the segmentation is assessed by points, in a scale from

0 to 100, and a score greater than or equal to 90 is considered to be comparable with a human expert. Four different aspects weight the score: absolute volume difference between the ground truth and the automatic segmentation, the average distance between the volumes, false positive rate and true positive rate. The author's technique scored 82.005 points, less than 2 points below the method ranked first at the time the paper was published.

One advantage of the method by Simões et al. (2013) is that it did not require any alignment (registration) procedure since it used only FLAIR images. However, regarding the preprocessing stage, the authors stated that further investigations would be necessary to suppress any undesired effects caused by bias field correction and skull stripping in the final segmentation. Still, the technique proved to be interesting because of a) its simplicity, b) it required only FLAIR images, c) it was able to outperform other automatic techniques in the literature and d) it provided a good agreement with manual annotations.

Tomas-Fernandez and Warfield (2015) proposed a segmentation technique based on GMM and graph-cuts called MOPS. The authors used a population of 15 T1-w, T2-w and FLAIR MRI scans from healthy subjects as a reference model to detect MS lesions as outliers. The reference population was segmented into three clusters (i.e., CSF, GM and WM) using GMM. The EM algorithm was used to estimate the parameters of the model. The authors observed that the global gray level intensities of these three tissues in healthy subjects had a significant overlap with MS lesions. To overcome this problem, they decided to couple global intensity with local intensity by analyzing a window of radius $r = 2$ around each voxel. By doing so, lesions differed much more significantly from other brain tissues.

To perform the actual segmentation of MS lesions, the approach by Tomas-Fernandez and Warfield (2015) aligned the healthy reference population to scans from patients with MS and selected a set of voxels with a low probability of belonging to CSF, GM and WM (considering both global and local intensity information). From this initial set, they partitioned it into two using the Otsu´s thresholding method (OTSU, 1975): a collection of candidate MS lesion voxels and a set of hyperintense voxels. Then, for the set of candidate MS lesions, the authors applied a max-flow algorithm (BOYKOV; VEKSLER; ZABIH, 2001) to compute the optimal cut of the desired graph and thus achieve the final segmentation.

The authors used DSC, lesion true positive rate (LTPR) and lesion false positive rate (LFPR) to assess their proposal. They applied their algorithm to synthetic and real image datasets. The synthetic dataset of choice was BrainWeb, using images with mild, moderate and severe lesion loads. They compared their combined local/global intensity approach with a global intensity approach only and were able to verify a significant improvement in sensitivity and specificity

when using the former instead of the latter. The real image dataset of choice was the MS Grand Challenge Clinical Dataset held at MICCAI 2008 (MICCAI MS Lesion Segmentation Team, 2008). Their proposal outperformed the other 17 entries at the time, ranking it at the top with a score of 84.5 (as mentioned before, a score of 90 is considered to be similar to an average human rater).

The MOPS technique performed well on the two datasets the authors applied it to and indicated that global intensity approaches suffer from intensity overlaps between MS lesion profiles and healthy brain profiles, thus making it necessary to incorporate local information to achieve a better distinction between clusters. The drawbacks of MOPS lie in the fact that it requires a reference population of healthy subjects, which entails a thorough and robust intensity normalization and image registration procedure, and the voxel lesion probability (i.e., low probability of belonging to CSF, GM and WM) is defined by the user instead of being automatically estimated.

In Roura et al. (2015), the authors performed MS lesion segmentation using T1-w and FLAIR images. Their approach was based on two stages. First, they segmented the image into three clusters (CSF, GM and WM) on T1-w using the SPM8/12 algorithm, which is based on GMM (ASHBURNER; FRISTON, 2005). Then, they used the GM cluster (the most intense one among the three) to compute its distribution in FLAIR and selected voxels considered to be outliers as candidate MS lesions.

The outlier threshold itself was not sufficient to select only MS lesion voxels, so the authors proposed three post-processing steps to remove false positives and improve the accuracy of their method. The first step analyzed the percentage of voxels of a candidate lesion lying in WM and GM over CSF. If more voxels were lying in CSF than in WM and GM, then that lesion was considered to be a false positive. The second post-processing step analyzed the proportion of WM in the immediate neighborhood of a candidate lesion voxel; the neighborhood should have a more significant percentage of WM compared to GM in order to the central voxel be kept in the lesion cluster. Finally, components with a volume smaller than $3mm^3$ were discarded.

To evaluate their approach, the authors used DSC, true positive rate (TPR) and positive predictive value (PPV) and computed them between the ground truth and automatic segmentation. Like other proposals mentioned before, the authors used the MICCAI 2008 dataset to assess their approach and compare it with other techniques. They were able to score 82.34 points and, at the time of submission, were the only ones to use an unsupervised approach.

An important point to highlight is that the authors made their segmentation approach available through the SPM8/12 toolbox, allowing it to be used by other researchers and physicians. It was a rather simple approach, and even though it was not able to perform better than other

techniques (for instance, the procedure mentioned earlier by Tomas-Fernandez and Warfield (2015)), it can still be used as an initial segmentation that can be improved by another approach.

In Griffanti et al. (2016), the authors created a supervised tool called BIANCA based on the k-nearest neighbors algorithm for white matter hyperintensity segmentation. Their approach used the weighting of spatial information, local spatial intensity averaging and multiple image weights to gather enough features to provide a consistent segmentation.

The k-nearest neighbors (k-NN) algorithm is a method for classifying objects based on the closest training examples in its feature space. In BIANCA, this feature space was comprised of intensity and spatial features. Regarding MR image acquisitions, the authors used mainly T1-w, T2-w and FLAIR images during segmentation. They also used patches to introduce average local intensity information for each sample. The spatial features were incorporated into the model using spatial weighting, which was defined by the user. If the weighting was high, then the model would be trained with samples coming from very similar regions. On the other hand, a low weighting would make the model to use training samples that could be spatially far apart.

The authors applied their algorithm to two datasets with 85 and 474 participants. Both datasets were acquired at the University of Oxford. They assessed BIANCA with different metrics, being the most relevant ones DSC and intra-class correlation (ICC). The authors were able to observe that using T1-w and FLAIR images provided the best results in terms of high DSC and lowest cluster-level false positive ratio. Also, they observed that using subjects with a high load of WM hyperintensities (WMH) and a different number of training points for WMH (2000) and non-WMH (10,000) yielded the best results.

Direct and indirect comparisons with other works and techniques in the literature were made, and BIANCA achieved comparable results with all of them. The indirect comparison comprised works using supervised and unsupervised approaches to segment different pathologies (MS, neurodegenerative diseases, strokes and vascular problems). Direct comparison was made with three publicly available algorithms (SCHMIDT et al., 2012; DAMANGIR et al., 2012) applied to the two datasets the authors had access to. In this last scenario, BIANCA outperformed all the other three techniques.

Overall, BIANCA was able to achieve good results and is currently available as part of the FSL toolbox[1]. The drawback of this technique is that it requires labeled data to be trained, whereas other unsupervised approaches in the indirect comparison presented by the authors were able to achieve better DSC and ICC scores than supervised ones.

[1] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL`

In Egger et al. (2017), the authors applied two variations of a segmentation algorithm by (SCHMIDT et al., 2012) and compared their results with three independent, experienced raters to verify if the automated segmentation was comparable to manual annotations.

They applied the segmentation algorithm to a dataset of 50 T2-w images of patients with Relapsing-Remitting Multiple Sclerosis (RRMS) and compared the results with the other three raters. To do so, they used the DSC and ICC metrics to provide a quantitative way to compare the outcomes. They were able to verify a strong correlation between manual and automated lesion segmentation (ICC = 0.958 and DSC = 0.60), which was not statistically different from inter-rater correlations (ICC = 0.949 and DSC =0.66). They also observed a significant variability regarding the number of lesions found by the automated algorithm and human raters.

The study by Egger et al. (2017) had significant importance due to its comparison with three experienced raters. Most studies usually rely on one or two experts at most to provide a correlation between automatic and manual outcomes. Their approach was also essential to show that an agreement among experts regarding MS is usually only partial, indicated by a DSC smaller than 0.70 (BARTKO, 1991) and by the variability in the number of lesions found by each rater. These findings provide more evidence on the difficulty to segment MS lesions, since not only the problem itself is challenging, but even human raters have biases that interfere with the assessment of the segmentation accuracy.

Using an energy minimization approach, Zhao et al. (2017) proposed a technique to segment MS lesions using T1-w and FLAIR images. The foundation of the idea was based on the widely accepted MR image formation model $I(x) = I_0(x)b(x) + n(x)$, where $I(x)$ is the observed MR image, $I_0(x)$ is the ideal image, $b(x)$ is the bias field and $n(x)$ is additive noise. Building on the work of (LI; GORE; DAVATZIKOS, 2014), which aimed to segment healthy brain tissues (CSF, GM and WM) in MRI using energy minimization, the authors expanded this idea by incorporating more than one image weight in the segmentation and including MS lesions as a fourth tissue type to be segmented by the model.

The authors defined an energy function $F(u,c,w)$, where u was a membership function (i.e., if a voxel belonged to a given tissue), c was a constant vector related to physical properties of the tissues being imaged and w was a coefficient vector that could be tuned to achieve energy minimization. By and large, the authors' technique iteratively explored the relationship between the image signal and bias field to estimate the probability of a given voxel belonging to a particular tissue. They started with an initial estimation and refined it in terms of u, c and w until a convergence criterion was met.

After convergence, an initial mask of segmented lesions was generated, which was then

used as input for a post-processing step to remove false positives by thresholding areas of all connected components in the lesion mask. In other words, the authors analyzed each connected component and excluded those that did not have a minimum number of voxels within it.

The results presented in (ZHAO et al., 2017) were only qualitative. They applied their technique to a set of five images with apparent severe lesion load. They stated that their approach, at the current stage, was to be considered as an initial segmentation that should be refined. They also mentioned that some GM regions near the cortex were wrongly clustered as MS lesions since the former has an intensity overlap with the latter. According to the authors, this problem can be circumvented by using spatial information from probabilistic anatomical atlases. Overall, their technique presented visually promising results; however, due to the lack of quantitative data, it was not possible to assess how well this approach performed compared to other works.

As it can be noted, there are many different approaches used throughout the years to tackle the MS lesion segmentation problem in MRI. Though there have been improvements on techniques, both supervised and unsupervised, the results are not yet as good as the ones achieved by human experts. For this reason, there is room for new ideas to be explored in this scenario.

The segmentation techniques described in this section are summarized in Table 4.1.

Author	Year	Technique	Dataset	Results
Nguyen and Wu (2012)	2012	SMM	18 subjects	DSC ≈ 0.85 (WM)
Zhang, Wu and Nguyen (2013)	2013	SMM	12 images	PR index $\approx$ 0.790
Simões et al. (2013)	2013	GMM	40 subjects	DSC > 0.70
Tomas-Fernandez and Warfield (2015)	2015	GMM	51 subjects	LTPR ≈ 0.72
Roura et al. (2015)	2015	SPM8	70 subjects (clinical), 23 subjects (challenge)	DSC ≈ 0.30, Score: 82.34
Griffanti et al. (2016)	2016	k-NN	2 datasets	DSC ≈ 0.75
Egger et al. (2017)	2017	SPM8	50 subjects	DSC ≈ 0.60
Zhao et al. (2017)	2017	Energy min.	5 subjects	—

Table 4.1: Overview of segmentation papers mentioned in Section 4.1.

4.2 Texture as a feature for segmentation and classification

The concept of texture in images follows the same idea of texture in daily life. Familiar terms as rough, silky or bumpy can be used to describe a surface we are interested in. More importantly, textures are intrinsically related to neighborhood analysis around pixels or voxels. In other words, a single point is not sufficient to provide enough information to extract textural features from it.

Textures can be used to identify a number of characteristics in images such as gray level differences (contrast), size of areas where changes occur and directionality (or the lack of it) in a region of interest. The notion of textures in medical images is related to tissue manifestations defined by form, structure, composition and arrangement of its smallest components. They can be seen as mathematical derivations that characterize the distribution of gray level intensities that reflect the structural consistency of any given tissue. In this sense, textures are directly affected by disease progression. Since texture analysis is commonly used to detect subtle structural changes, it stands out as a suitable candidate to evaluate the activity and evolution of pathologies. For this reason, they can be incorporated into classification and segmentation models in order to provide more information and improve distinction among clusters or classes.

There are a myriad of techniques used to extract texture measurements from images. The most commonly used in the context of MS are Gray Level Co-occurrence Matrix (GLCM) (HARALICK; SHANMUGAM; DINSTEIN, 1973), Run-length Matrix (RLM) (GALLOWAY, 1975) and Local Binary Patterns (LBP) (OJALA; PIETIKAINEN; MAENPAA, 2002). GLCM tries to identify the spatial dependency between gray levels in an image, i.e., given a voxel located at p with intensity i, we count the occurrences of another intensity, j, in the neighborhood of p. RLM is used to identify the granularity of a texture by taking into consideration the repetition of certain intensity levels in a given direction. And finally, LBP computes a local representation of texture by comparing each voxel with its surrounding neighborhood. It is important to note that all these three techniques depend on direction; so in order to achieve rotation invariance, the outcome is commonly averaged over all directions.

In Zhang et al. (2008), the authors analyzed textural features extracted from GLCM, RLM, autoregressive model (AR) (SUKISSIAN; KOLLIAS; BOUTALIS, 1994) and wavelet analysis (RAJPOOT, 2002) with the goal to differentiate MS lesions from normal appearing white matter (NAWM) and normal white matter (NWM). Distinguishing them is of vital importance, both from clinical and computational standpoints, because it helps physicians grasp the disease progression more easily with quantitative data and it can also help improve segmentation accuracy

on automatic techniques.

A total of 16 MS subjects and 16 healthy subjects were selected for the authors' study. They used T2-w scans and chose 16 regions of interest (ROIs) of MS patients for MS lesions and NAWM, respectively, and 16 ROIs of healthy subjects for NWM. A radiologist and a neurologist manually performed the selection of ROIs.

After selecting the ROIs, the authors extracted more than 200 textural features using the Mazda software[2]. Feature selection was performed based on the highest difference among different texture groups to determine which of them were the most useful for classification. By doing so, the authors were able to reduce the number of significant features down to 27. It is important to note that these features came from GLCM, RLM, AR and wavelet analysis and were combined to provide more powerful discriminating characteristics.

The authors were able to observe that the combined set of features was able to perfectly distinguish MS from NWM and MS from NAWM. Regarding the distinction between NWM from NAWM, the classification accuracy was less successful (58.33% accuracy). They also compared the results of combining GLCM, RLM, AR and wavelets with GLCM features only and verified that the first outperformed the latter in every scenario, indicating that using textures extracted from more than one technique can improve the distinction power of the model.

The authors also mentioned a few drawbacks regarding their approach. The first one was that the ROI selection step was performed manually. Choosing ROIs in a semi-automated or automated fashion could reduce inherent biases from human experts. Another issue mentioned by the authors was the requirement of large sample sizes in the Mazda software, which could become an obstacle in studies with small cohorts. But overall, the authors' approach was a good indication of the discriminative power of textures and that physicians can use them and computer techniques alike to improve diagnostic and follow-up procedures in the MS context.

In Roy, Bhuiyan and Ramamohanarao (2013), a new textural feature set for MS lesion segmentation was proposed by combining local and global neighborhood information. The set was comprised of GLCM, RLM, gradient and histogram features, along with gray level intensity and tissue probabilities. These features were used to train a Support Vector Machine (SVM) (MURPHY, 2012) classifier to perform MS lesion segmentation.

The global neighborhood information was obtained by creating a lesion probability map with patches across the whole image. The authors defined a patch of size N and for each voxel v, analyzed the neighborhood of 10 random patches spread across the brain region. The more

[2] http://www.eletel.p.lodz.pl/cost/progr_mazda_eng.html

times v had a gray level intensity greater than the patches, the more likely it was to be a lesion. On the other hand, local neighborhood information was achieved by defining a 5×5 window around each candidate pixel and calculating gradient, histogram, GLCM and RLM information from it. Then, as mentioned before, the authors combined these features with gray level intensity and tissue probabilities to use as input for the SVM classifier.

The authors applied their technique to 8 subjects from the MS Lesion Segmentation Challenge 2008[3] and compared their results with three other state-of-the-art methods (ABDULLAH; YOUNIS; JOHN, 2012; GEREMIA et al., 2011; SOUPLET et al., 2008). To assess their performance, the authors used the F1-score (GONZALEZ; WOODS, 2008) metric. Their technique outperformed the other works in five out of eight cases. For the three remaining subjects, the difference in score between their proposed method and the best one was small according to the authors.

Overall, they presented a discriminative set of texture features, and the lesion probability map created from global neighborhood information was a significant contribution, though the definition of the number of patches and the use of absolute gray level differences seemed somewhat arbitrary. A larger dataset would provide more insight into the proposed method and allow more conclusions to be drawn from the results.

In Michoux et al. (2015), the authors used texture analysis to distinguish normal appearing white matter (NAWM) from active/nonactive MS lesions in T2-w images. They selected 21 patients and extracted features from both GLCM and RLM to help differentiate between these two classes. Each patient was imaged for T1-w, T2-w and FLAIR modalities, but the features themselves were extracted only from T2-w.

Two experts annotated a total of 44 active lesions, 37 nonactive lesions and 44 NAWM ROIs. MS lesions were categorized as active or nonactive from the analysis of T1-w images. Since the authors wanted to check if the perfusion of contrast would affect the outcome of T2-w images, they also selected 9 extra patients to be scanned twice (pre- and post-contrast). Of this second group, 14 active lesions were annotated.

Using linear discriminant analysis (LDA), partial least squares (PLS) and logistic regression (LR), they were able to get different sets of textural features and results with each technique. The authors achieved the highest accuracy when applying PLS to a 6-texture parameter model, which yielded a sensitivity of 0.88 and specificity of 0.81. These results were achieved by two different six-feature sets. The first one was a combination of four GLCM and two RLM features, whereas the second one was an equal combination of three GLCM and RLM features. Moreover, they did not see any statistically significant difference between pre- and post-contrast

[3] http://www.ia.unc.edu/MSseg/

textures in T2-w images, demonstrating that the contrast perfusion had no effect on them.

According to the authors, a limitation of their study concerned the limited number of patients (a total of 21) and active lesions (a total of 58), indicating that further investigations should be conducted in order to confirm their findings in a larger set of data. Another important point to be made is that the goal of the authors was to distinguish active MS lesions from NAWM, not active from nonactive lesions. Though the correct distinction between NAWM and other MS pathologies is crucial, it is not a sufficient condition to withdraw the usage of Gadolinium-based contrast.

In Leite et al. (2016), the authors proposed a 3D texture-based classification algorithm for brain white matter lesions (WML) in MRI. Texture analysis was based on 55 texture attributes extracted from gray level histogram, GLCM, RLM and gradient. The goal was to compare 3D texture-based with 2D texture-based approaches to check for accuracy improvements by using the former instead of the latter.

The authors extracted textural features from FLAIR images of 61 subjects diagnosed with carotid artery atherosclerosis. The ROIs containing WML were selected using Cerebra-WML (LU et al., 2014) and regions of NAWM were chosen using an automated approach developed by the authors that either selected the contralateral part of a WML ROI or the closest region to the WML using the same WML size and shape. A total of 2658 WML samples and 2637 NAWM samples were selected for the 2D analysis, while for 3D they chose 69 samples for WML and NAWM each.

After selecting samples, the authors used SVM with a radial basis function (RBF) kernel ($C = 1000; \gamma = 0.1$) for the classification task. They were expecting the 3D case to perform worse than 2D since the former had a much smaller number of samples. However, they observed the opposite. Using 3D samples yielded the best classification and sensitivity results. The authors stated that this result might have happened because the 3D implementation was more robust and comprised relevant volumetric structural information. Still, both 2D and 3D approaches presented high accuracy rates (97.41% and 99.28%, respectively) and, according to the authors, could easily be used in a real computer-aided diagnostics (CAD) system to distinguish NAWM and WML. The authors also analyzed the 30 most discriminant texture attributes for the 2D and 3D approaches using a method known as Gini importance (BREIMAN et al., 1984). They observed that the histogram and GLCM features were the most relevant ones.

In conclusion, their proposal was able to distinguish WML from NAWM correctly and also provided a comparison between 2D and 3D approaches. Though images used in this particular work had lesions caused by carotid artery atherosclerosis, one can speculate that the same set

of textural features could also be used to distinguish MS lesions. However, a different level of accuracy would undoubtedly be expected in this case.

In Tiwari et al. (2016), the authors used co-occurrence matrix homogeneity, neighboring gray level dependence matrix and multi-scale Gaussian derivatives textural features to distinguish radiation necrosis (RN) from brain tumor recurrence. RN is, unfortunately, a common side effect of the treatment of cancerous pathologies and is caused by radiation that is dispersed across the brain during radiotherapy. Its onset is typically a few weeks after the end of radiotherapy treatment and has a very similar appearance to brain tumor recurrence in MRI, making it hard to distinguish to the naked eye. For this reason, the authors wanted to explore computer-extracted textures to aid in the differentiation between RN and brain tumor recurrence.

The study population was comprised of datasets imaged at the University Hospital Case Medical Center and the University Texas Southwestern Medical Center. The data were divided into training cohort and metastatic subgroups. In total, 43 subjects were made available for the study, scanned in T1-w, T2-w and FLAIR weights. The ROIs containing tumors were manually selected on all three image weights by an experienced radiologist. Then 119 2D texture features were extracted from each tumor. To identify the most discriminative features, the authors did a minimum redundancy and maximum relevance (PENG; LONG; DING, 2005) feature selection analysis and used the top 5 most discriminative features as input for a SVM classifier. They observed that adding more than 5 features did not significantly improve the distinction capabilities of the classifier.

The performance metric used by the authors was the area under the receiver operating characteristic curve (AUC). They observed that FLAIR images provided the best results for both training and metastatic cohorts, with an $AUC = 0.79$. Moreover, using 15 studies from a hold-out cohort, the SVM classifier identified 12 of 15 studies correctly, whereas two experienced human raters diagnosed 7 out of 15 and 8 out of 15 studies correctly, respectively.

Despite being a feasibility study, the findings of Tiwari et al. (2016) indicate that textures can be used in the context of various brain pathologies and have the potential to not only help physicians distinguish them from side effects of treatments, but it can also be more reliable than human experts.

In Ardakani et al. (2017), the authors decided to use texture analysis (TA) to differentiate enhancing/active lesions (ELs) from non-enhancing lesions (NELs) and persistent black holes (PBHs). ELs are of particular interest for physicians because it is a sensitive indicator of active inflammation and serves as a proxy to monitor disease activity and impairment of subjects with MS. Most new lesions become enhanced and persist for 2-6 weeks (FILIPPI et al., 2001) and

are usually visible to the naked eye in Gadolinium-enhanced T1-w images, i.e., it requires the injection of Gadolinium-based contrast. In this sense, the authors wanted to check if they could distinguish ELs, NELs and PBHs in pre-contrast images using TA.

The dataset comprised 90 subjects scanned on T1-w and T2-w weights. Regarding T1-w, the subjects were scanned before and after being injected with a Gadolinium-based contrast agent. Overall, 116 ROIs consisting of 54 NELs, 27 ELs and 35 PBHs were manually selected by an expert for discrimination and classification. More than 300 texture features were extracted based on the histogram, absolute gradient, RLM, GLCM, AR and wavelet information. One-way analysis of variance (ANOVA) was used to assess differences between the three groups of interest (EL, NEL, PBH) and identified a total of 14 features as the most discriminative ones.

The authors used linear discriminant analysis (LDA) to transform the most discriminative features to lower-dimensional spaces to further increase their distinctive power. They were then used as input for a k-NN classifier. Using AUC as a metric for classification accuracy, the authors achieved perfect classification for NEL vs. EL and EL vs. PBH. For NEL vs. PBH, the classification yielded a result of $AUC = 0.975$. The authors also observed that GLCM textures presented the most discriminative power among the group of the top 14 features. Moreover, they verified that using all significant features yielded a higher level of performance than using each texture alone.

Overall, the work of Ardakani et al. (2017) indicated that TA is a viable alternative to discriminate between different groups of MS lesions. More importantly, it showed that textures have the potential to identify EL without injecting any contrast, which can have positive outcomes both for patient's health and imaging costs. Unfortunately, the selection of ROIs was made manually, and an automated approach to leverage the benefits of this approach is yet to be proposed.

In Verma et al. (2017), the authors applied dynamic texture parameters analysis (DTPA) to differentiate enhancing MS lesions from glioblastoma and lymphoma. All these three pathologies are mostly visible using Gadolinium contrast, and the initial diagnosis may be difficult to achieve based solely on conventional Gadolinium-enhanced MRI. In this sense, textural features provided by DTPA were examined in order to verify if they could provide enough information to help physicians better distinguish one pathology from another.

A total of 32 subjects, 12 with MS, 15 with glioblastoma and 5 with lymphoma, were selected and comprised only T1-w images after intravenous gadolinium application. ROIs were then selected by an experienced radiologist containing only areas enhanced by the gadolinium contrast agent.

The authors applied GLCM to the images in different time intervals (inflow of contrast, outflow of contrast and reperfusion). They used GLCM solely because this technique is regarded as the most common and easy-to-understand class of textural features and usually provides a good distinctive power to classifiers. Incorporating time intervals added a time dependency to the model, making it dynamic. It was observed that the most significant differences among textures appeared during the inflow phase, followed by outflow and reperfusion phases. These results indicate that there must be differences in micro-vascularization depending on the disease type, which affects the inflow phase of the contrast agent passage and, to a lesser extent, the outflow phase. This particular finding, according to the authors, could only be uncovered thanks to texture analysis.

Unfortunately, no classification stage was implemented to quantitatively verify how well the texture features would distinguish one pathology from another. Moreover, the authors stated that, as with any feasibility study, the number of patients was small. More subjects would be necessary to encompass a broader spectrum of variabilities in image scans and pathologies. Also, the manual delineation of the ROIs could introduce the expert's bias into the model; in this sense, an automatic approach would be more adequate. Overall, the work by Verma et al. (2017) was important to show that there are significant, yet simple, textural features that can be used to distinguish MS from other pathologies and also to show that the most distinguishable phase occurs during the inflow phase of the contrast agent.

In Bento et al. (2017), a probabilistic segmentation of brain white matter lesions using textures was proposed by the authors. They used a supervised approach and combined gray level intensities with LBP morphological gradient to perform the segmentation of white matter hyperintensities.

The experiments were conducted on a public dataset from the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention in 2015 (MICCAI-15). This particular data were put together for the Medical Image Segmentation Challenge on Ischemic Stroke Lesion Segmentation[4]. MRI exams, comprised of T1-w, T2-w and FLAIR images, of 28 subjects diagnosed with stroke were made available. However, in their study the authors used only FLAIR images. Every patient had his/her scan analyzed by an expert, who provided a manual segmentation of the stroke lesions.

The methodology comprised preprocessing, feature extraction, classification and post-processing steps. The preprocessing stage included intensity normalization and segmentation of the WM region using a semi-automated approach (LU et al., 2014). Regarding features, seven attributes

[4] www.isles-challenge.org

were extracted for each voxel, including the gray level intensity of the voxel, intensity value in the LBP image, structural gradient, morphological gradient and WM average intensity value across the whole WM region. In the classification stage, the authors applied three different supervised classifiers: k-NN, SVM and random forest (MURPHY, 2012), being the latter the classifier that achieved the best results. To train these algorithms, the authors used 12 of the 28 subjects, and the testing phase was conducted on the remaining 16 images. Instead of outputting a class label, they used a probabilistic implementation (PEDREGOSA et al., 2011). The post-processing step used a series of morphological operators (GONZALEZ; WOODS, 2008) to prune the probabilistic outcome and generate a binary mask with the actual lesion segmentation.

The authors achieved a high DSC coefficient (0.84) and outperformed five other automatic segmentation techniques (HAVEI et al., 2015; KAMNITSAS et al., 2015; CHEN; BENTLEY; RUECKERT, 2015; FENG; ZHAO; HUANG, 2015; HALME; KORVENOJA; SALLI, 2015) that used the same database. They noted the DSC results were highly variable and independent of the total number of white matter lesions, and stated their approach was limited to properly segment lesions greater than 10 pixels, since texture features are based on changes across a local neighborhood, thus making it difficult to achieve an accurate segmentation on lesions that were too small. Though this is a limitation, it is fundamentally related to the minimum size lesions may have to be correctly segmented by the current state-of-the-art. Moreover, despite being a supervised approach, it is important to note that this technique was able to achieve good results using a relatively small number of images with fairly simple classification algorithms; this is an indication that the right set of features can outweigh the need for complex classifiers.

Using some concepts of deep learning, the work of Fernandes and Cardoso (2017) proposed using LBP in a cascade fashion in order to explore higher dimensional features from textures. Since LBP is, by design, an approach that explores local relationships and patterns between neighboring pixels/voxels, cascading it in multiple layers would allow one to leverage higher order information and improve the accuracy of classifiers by providing a broader feature space.

As mentioned by the authors, while it is trivial to apply convolutions in a cascade fashion recursively, traditional LBP are not able to "process" their output. Such issue arises from the fact that they rely on receiving as input an image with the domain in an ordered set (e.g., gray level intensities). However, LBP codes are not an ordered set, thus dismissing the direct recursive application of standard LBP. To circumvent this issue, the authors suggested using, for instance, a dissimilarity function defined by the user to ensure ordering. In other words, given a dissimilarity function, a mapping of the LBP codes can then be made to an ordered set, making

it possible to use them in cascade.

To assess their approach, the authors used accuracy and class rank (position (%) of the ground truth label in the ranking of classes ordered by confidence) as metrics. The final decision model was a random forest with 1000 trees. They applied their proposal to seven publicly available datasets and compared results with a varying number of layers. The authors observed that despite the capability of the proposed deep architectures to achieve large gain margins, the deep LBP operator saturated rapidly, since most of the best results were found on architectures with up to three deep layers.

Overall, the technique proposed by the authors was able to aggregate information from local neighborhoods into higher abstraction levels, being able to surpass the performance obtained by simple non-cascade LBP. However, further research must be conducted in several areas to verify how this approach would behave in real-world problems and if the increase in complexity justifies the gain in accuracy.

The texture-based techniques described in this section are summarized in Table 4.2.

4.3 Symmetry and asymmetry analysis in MS

There are not many works in the literature that make use of symmetry and asymmetry information in the context of MS. Most authors have been using this particular feature to analyze structural and volumetric differences between healthy and pathological brains. While this application is relevant and helps physicians better understand the effects pathologies have on brain tissues, we believe it has the potential to be further explored from a segmentation pipeline point of view. In this sense, it is important to note that symmetry/asymmetry analysis remains an open field of research, especially when it comes to incorporating it into automatic segmentation and classification models.

In Prinster et al. (2006), the authors experimented to measure gray matter loss in relapsing-remitting MS subjects, which can alter the physiological asymmetry pattern of a healthy brain and add even more differences between healthy and MS patients.

The dataset used by the authors was comprised of 51 MS patients and 34 healthy subjects scanned on T1-w, T2-w and PD weights. First, the authors applied a non-parametric segmentation technique (B et al., 2000) to all images. Then, using a voxel-based morphometry process (ASHBURNER; FRISTON, 2000), they compared the GM clusters from both groups and looked for differences between them across the whole brain on a voxel-by-voxel basis. To get a better

Author	Year	Textures	Dataset	Results
Zhang et al. (2008)	2008	GLCM, RLM, AR, wavelets	32 subjects (MS)	100% accuracy for MS vs. NWM and NAWM
Roy, Bhuiyan and Ramamohanarao (2013)	2013	GLCM, RLM, gradient, histogram	8 subjects (MS)	F1 score = 0.463
Michoux et al. (2015)	2015	GLCM, RLM	30 subjects	Se = 0.88 Sp = 0.81
Leite et al. (2016)	2016	GLCM, RLM, gradient, histogram	61 subjects (carotid artery atherosclero-sis)	99.28% accuracy in distinction
Tiwari et al. (2016)	2016	Laplacian textures	43 subjects (cancer)	AUC = 0.79
Ardakani et al. (2017)	2017	GLCM, RLM, gradient, AR and wavelets	90 subjects (MS)	AUC = 1.0 for lesion distinction
Verma et al. (2017)	2017	GLCM	32 subjects (MS)	—
Bento et al. (2017)	2017	LBP	28 subjects (ischemic stroke)	DSC = 0.84
Fernandes and Cardoso (2017)	2017	LBP	7 different datasets	Average accuracy = 67.49%

Table 4.2: Overview of texture papers mentioned in Section 4.2.

grasp of these differences, the authors also searched for significant asymmetries between the MS and control groups using a technique based on the work of Luders et al. (2004). By doing so, they were able to globally and locally analyze the effect MS had on the GM tissue.

The results regarding global GM volumes confirmed previous findings of GM loss in MS patients when compared to healthy subjects (CHARD et al., 2002; QUARANTELLI et al., 2003), with an average tissue volume of 607.7 ml for the former and 660.1 ml for the latter. A critical

finding from the asymmetry analysis was that significant GM loss in MS patients was mainly located on the left dorsolateral frontal lobe, but was also bilaterally present in other areas such as the anterior cingulate gyrus and caudate heads. However, as the authors noted, there was no correlation between the observed GM loss with the spatial location of lesions, since they appeared to be symmetric across MS subjects. Given that, the authors stated that there must be other mechanisms involved in determining cortical GM loss, which is associated with the progressive phases of MS (BO et al., 2006) and tends to be overlooked because MS lesions usually appear in WM.

In Geremia et al. (2011), the authors used spatial decision forests for MS lesion segmentation using T1-w, T2-w and FLAIR images. They incorporated not only local information into their model (i.e., gray level intensities) but also spatial information using MNI probabilistic anatomical atlases (MAZZIOTTA; TOGA; EVANS, 2001) and context-rich features. They applied their technique to the MICCAI MS Lesion Segmentation Challenge 2008 (MICCAI MS Lesion Segmentation Team, 2008) dataset and compared their results with state-of-the-art algorithms.

The authors defined two context-rich features. The first one came from regions of user-defined size around the voxel of interest. The second one consisted of using the counterpart of the voxel of interest, thus making use of symmetry. Since getting the counterpart was tricky due to interpolation problems, the authors decided to extract information from a counterpart region (instead of a counterpart voxel).

An interesting byproduct of using spatial decision forests, according to the authors, is that it allowed them to identify the most discriminative visual features for MS segmentation. In this regard, they were able to observe that region information was the most selected feature in the trees, followed by local information and symmetry. The authors mentioned that symmetry features were under-represented in trees because a large number of peri-ventricular MS lesions tend to develop symmetrically; however, they also observed that this feature was most relevant to classify ambiguous asymmetrical regions.

Regarding the results, the authors stated their technique compared favorably to other works ranked in the dataset. As mentioned in Section 4.1, the 2008 MICCAI dataset ranked the works from 0 to 100, with a score of 90 indicating an expert-like segmentation. The authors' proposal scored 82.07, and although they did not rank at the top at the time of submission, they mentioned that the reliability of automatic methods is generally higher than that of human experts and thus their result was promising from an output consistency standpoint. Moreover, they also observed a drawback in their technique: the location of some lesions in the test dataset did not appear in the training dataset, highlighting a challenge to overcome when using supervised approaches.

In Govindarajan et al. (2015), the authors analyzed the effect of using an in-painting strategy on cortical thickness measurements in MS. Since this is an important metric to be used as a proxy for brain health, the smaller the effect lesions have on cortical thickness measurement, the better.

The authors used a dataset comprised of 918 subjects who participated in a clinical trial. Each subject was scanned in a protocol that included T1-w and FLAIR images. The pipeline adopted by the authors was the following. First, they segmented the images into WM, GM, CSF and lesions using an in-house developed pipeline (DATTA et al., 2006; SAJJA et al., 2006). Then, using the WM cluster mask, the authors randomly selected intensity values from this mask and used them to replace the intensity values in the lesion mask. They applied a Gaussian filter to the lesion area following the replacement of voxel intensity to minimize the effect of noise due to random assignment of intensity values. Finally, the authors used the FreeSurfer software (DALE; FISCHL; SERENO, 1999; FISCHL; SERENO; DALE, 1999) to calculate the cortical thickness in T1-w images, both original and in-painted.

An interesting result that arose from this study was that lesion in-painting presented a different behavior for each brain hemisphere. The authors created a lesion probability map using lesion masks from the segmentation step and verified there was no significant spatial asymmetry in the lesion distribution between the two hemispheres. Still, the left hemispheres showed more significant differences in the cortical thickness between in-painted and non-in-painted images. The reason for this asymmetry, according to the authors, remains unclear.

Overall, the authors proposed an in-painting algorithm to mitigate effects of cortical thickness measurement in MS images and used symmetry analysis to verify if the effects of their technique were the same for both hemispheres. Not only they found a significant difference in the measurement of cortical thickness for in-painted images, but also noted that their technique had a more prominent effect on the left hemisphere. Since they also observed that the MS lesion distribution across the images was symmetric - to a certain extent - the impact on the left hemisphere, as mentioned before, is still to be uncovered and, to the authors' best knowledge, has nothing to do with lesion asymmetry.

Finally, in Shen, Zhang and Zheng (2017), the authors proposed a technique for brain tumor segmentation in multi-modal MR images using a fully convolutional network (FCN). The novelty of their approach was based on using symmetry features to gather more information to their model and thus increase the accuracy of their method.

Their approach made use of one downsampling path with convolutional and max-pooling layers and three upsampling paths with upsampling and convolutional layers. It is important

to note that the downsampling module was responsible for encoding high level abstract and contextual information, while the upsampling part was responsible for reconstructing finer details, such as tumor boundaries. The authors stated they opted for a simplified version of FCN, with fewer layers, because they observed that adding more convolutional blocks did not present any improvement in brain tumor segmentation accuracy. The reason for this, according to the authors, is that unlike natural images, medical images are mostly based on low-level texture features, thus limiting the amount of high-level information that can be learned from them.

Regarding symmetry, the authors stated that it is an important cue for brain tumor segmentation since tumors usually break the symmetric appearance of healthy brains. To leverage this feature, they calculated the midsaggital plane, found corresponding matching pixel pairs and calculated their intensity differences on a 11×11 window neighborhood. They applied this rationale to each MR image weight and combined them with the original modalities to create the input for the FCN.

The authors applied their algorithm to the publicly available dataset BRATS 2013[5], comprised of 30 subjects scanned on T1-w, T1-wc (T1-w with contrast), T2-w and FLAIR. From these 30 subjects, 20 of them were used for training and 10 for testing. The authors observed that using an FCN without symmetry did not have the same accuracy as when using it, especially regarding the removal of false positives. On the results, their technique ranked second in the dataset scoreboard (with an average DSC of 0.87) and outperformed most supervised and unsupervised proposals alike. The only exception was for a convolutional neural network (CNN) by Pereira et al. (2016), which scored better in the dataset but had to use a healthy population as a reference for their approach.

An important point the authors made was that their algorithm took, on average, 2 minutes to process each 3D volume. Compared to other works applied to the same dataset, their running time was the shortest. This advantage in processing time is a significant achievement because it indicates the feasibility of use in real-world situations, such as clinical trials. Moreover, another relevant observation made by the authors was that increasing the number of layers did not improve segmentation accuracy, but incorporating symmetry information into the model did. This remark is an indication that a complex model is not always the best option for challenging scenarios, and that improving the feature space with relevant data can positively impact the final results.

An overview of the symmetry papers described in this section are shown in Table 4.3.

[5] `http://martinos.org/qtim/miccai2013/`

Author	Year	Dataset	Results
Prinster et al. (2006)	2006	51 subjects (MS)	Reduction in GM volume caused by MS
Geremia et al. (2011)	2011	MICCAI 2008 (MS)	Score = 82.07
Govindarajan et al. (2015)	2015	918 subjects (MS)	Small effect of lesion in-painting for cortical thickness
Shen, Zhang and Zheng (2017)	2017	30 subjects (brain tumor)	DSC = 0.87

Table 4.3: Overview of texture papers mentioned in Section 4.2.

4.4 Final considerations

This chapter presented a literature review of relevant and up-to-date works on segmentation, classification, texture and symmetry applied to MRI. It can be observed that the segmentation task itself is challenging and many different approaches, both supervised and unsupervised, can be used to tackle the problem. In this sense, textural features can be incorporated into the models to provide more information regarding the distinction between healthy tissues and pathologies. However, the same rationale cannot be extended to symmetry analysis, which has been shown to have a rather weak distinctive power in the MS lesions domain, which is why we opted not to use it in our pipeline.

To the best of our knowledge, we do not know of any work that combines local information, in terms of gray level intensities, spatial information and textures to perform MS lesion segmentation and active lesion classification. In this sense, we believe this combination was worth exploring and can ultimately expand the state-of-the-art on this matter.

Chapter 5

METHODOLOGY

The methodology used in this work is explained in this chapter, encompassing techniques, their characteristics and how they were applied to the MS lesions segmentation and classification problem. We also present a description of the databases used in our pipeline.

5.1 Databases

The clinical and probabilistic anatomical image databases used in this work are described as follows.

5.1.1 Clinical images

The clinical image database was comprised of 21 time-points (exams) from 5 patients in "The 2015 Longitudinal MS Lesion Segmentation Challenge", conducted at the 2015 International Symposium on Biomedical Imaging. This database contains images of five patients, one male and four females, with a total of 21 time-points. The average age of the patients was 43.5 years and the average time between follow-up scans was one year. Eligibility criteria and more details on the patients and how they were chosen are described in (CARASS et al., 2017).

Each scan was imaged and preprocessed in the same manner, with data acquired on a 3 Tesla MRI scanner (Philips Medical Systems, Best, The Netherlands). The imaging sequences were adjusted to produce T1-w, T2-w, PD and FLAIR images.

Each subject underwent the following preprocessing: the baseline (first time-point) magnetization prepared rapid gradient echo (MPRAGE) was inhomogeneity-corrected using N4 (TUSTISON; GEE, 2009), skull-stripped (CARASS et al., 2007, 2011) and dura stripped (SHIEE et

al., 2014), followed by a second N4 inhomogeneity correction and rigid registration to a 1 mm isotropic MNI template. Since the baseline MPRAGE was in MNI space, it was used as a target for the remaining images, which included the baseline T2-w, PD, and FLAIR, as well as the scans from each of the follow-up time-points. These images were then N4 corrected and rigidly registered to the 1 mm isotropic baseline MPRAGE in MNI space. In the end, image dimensions were $181 \times 217 \times 181$.

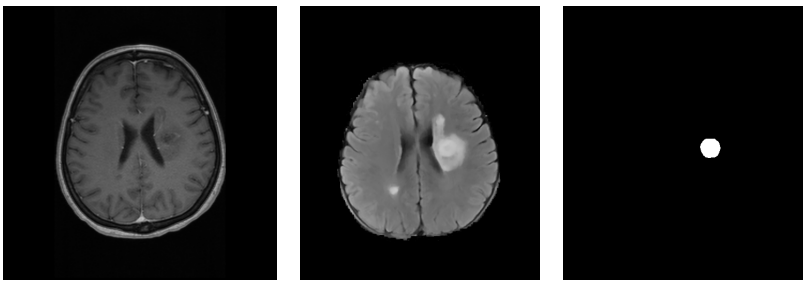
It is important to note that the training dataset also included manual MS lesion delineations by two experts for each time-point. Details about time-points and average MS lesion volume for each subject are summarized in Table 5.1.

	# of time points	Mean lesion volume (in ml) Expert 1	Mean lesion volume (in ml) Expert 2
Patient 1	4	16.67	19.07
Patient 2	4	30.52	31.80
Patient 3	5	5.40	7.81
Patient 4	4	2.17	3.23
Patient 5	4	4.55	3.96

Table 5.1: Number of time-points and average lesion volume for each patient.

5.1.2 Active lesions

Our active lesions database came from the Demyelinating Diseases Outpatient Clinics - Neurology & Neurosurgery Department - Universidade Federal de São Paulo - (UNIFESP), Brazil. It comprised scans of 33 patients with FLAIR and post-contrast T1-w images. An expert rater annotated active MS lesions in FLAIR using post-contrast T1-w as a reference for lesion location. Since this database existed prior to this work and the annotations were made on images as they were, 7 patients out of 33 had slightly different acquisition protocols regarding the number of slices per scan. While 26 patients were scanned with image dimensions $384 \times 512 \times 20$, the other 7 were scanned with image dimensions $384 \times 512 \times 25$. However, qualitatively speaking, the extra number of slices in this second group did not interfere with our region of interest (i.e., the brain itself). Voxel resolution was $0.44 \times 0.44 \times 6.5$ mm^3 across all images. An example of a patient from this database is shown in Figure 5.1.



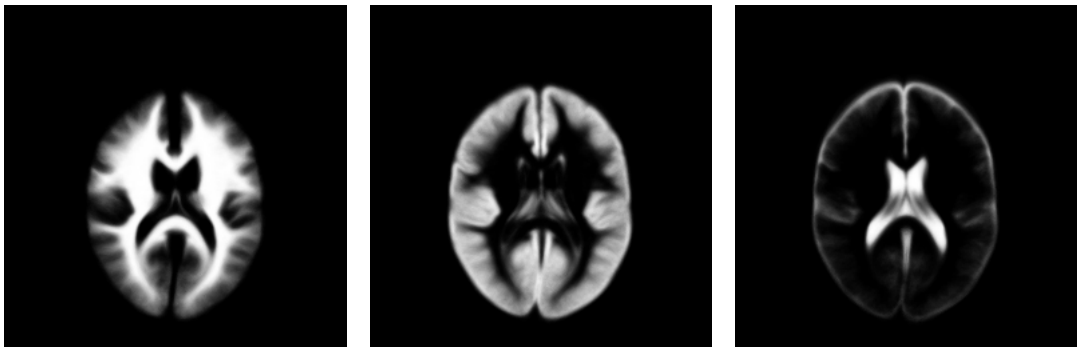
(a) (b) (c)

Figure 5.1: Patient with active lesions from the UNIFESP database. (a) T1-w, (b) FLAIR, (c) Active lesion mask annotated by an expert rater.

5.1.3 Probabilistic anatomical atlases

Probabilistic anatomical atlases are created from images of a normal population and are used as anatomical models to provide statistical information about tissues and cerebral structures (MAZZIOTTA et al., 1995).

In this work, three probabilistic anatomical atlases of CSF, GM and WM were used to provide spatial information to our segmentation technique. The images came from the ICBM project (FONOV et al., 2009) with dimensions $256 \times 256 \times 256$ and spatial resolution of 1 mm^3. Each voxel intensity in each atlas was in the $[0,1]$ interval, which indicated the probability of that particular voxel belonging to the corresponding tissue. The atlases are shown in Figure 5.2.



(a) (b) (c)

Figure 5.2: Probabilistic anatomical atlases. (a) WM, (b) GM, (c) CSF.

5.2 Preprocessing

A set of commonly used preprocessing techniques presented in the MR medical image processing literature were implemented and used in this work. The preprocessing pipeline

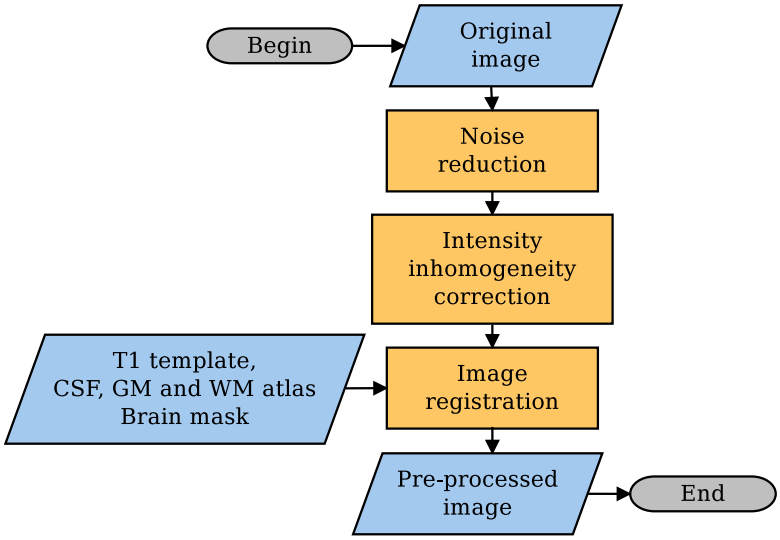
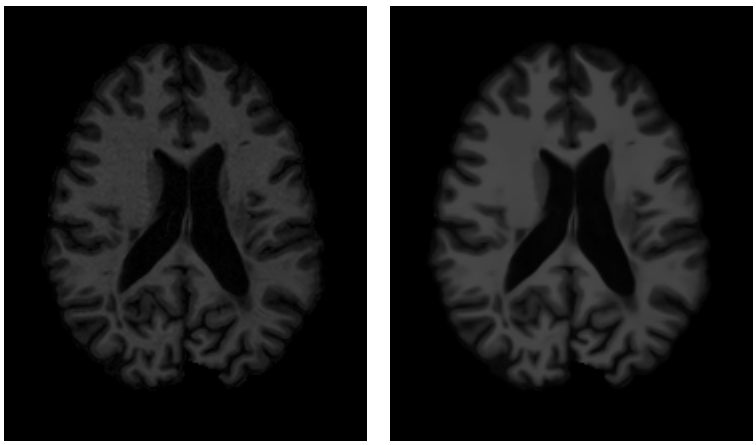


Figure 5.3: Preprocessing pipeline used in this work.

is shown in Figure 5.3. The description of each one of the stages is given in the following subsections.

5.2.1 Noise reduction

It is common for MR images to suffer from thermal noise, which interferes with the other stages of processing. To mitigate this issue, we applied the Non-Local Means (NLM) (BUADES; COLL; MOREL, 2005) filter to every time-point. This patch-based technique can smooth small variations of gray level intensities adaptively by decreasing the smoothness on edges (to preserve contrast) and increasing it otherwise. The algorithm is based on the fact that images have features that repeat not only locally, but globally. This way, it performs a weighted average of all voxels in an image region surrounding the analyzed voxel i in order to determine a noise-free estimation for it. In other words, it takes into account both the neighborhood information of a voxel and all regions similar to that neighborhood (patches) — hence the non-local characteristic. In Buades, Coll and Morel (2005) the authors conducted a qualitative and quantitative analysis of NLM compared to other well-known smoothing filters in the literature, such as Gaussian (LINDENBAUM; FISCHER; BRUCKSTEIN, 1994) and anisotropic diffusion (PERONA; MALIK, 1990; FERRARI, 2013), and were able to verify NLM's superiority. An example of the application of NLM is shown in Figure 5.4.



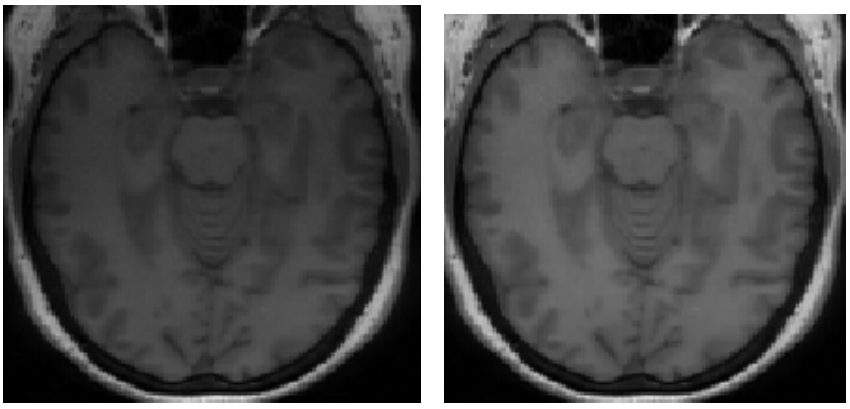
(a) (b)

Figure 5.4: Axial view of a T1-w image before (a) and after (b) applying the Non-Local Means noise reduction filter.

5.2.2 Intensity inhomogeneity correction

The presence of intensity inhomogeneity (or bias field) in MR images is a phenomenon caused by a myriad of factors, including poor radio frequency (RF) energy distribution, variations in the static magnetic field and sensitivity of RF coils. According to Arnold et al. (2001), these variations, if not corrected, can substantially degrade the result of automatic segmentation of brain tissues in MRI.

In this work, the N4 technique (Nick's Nonparametric Nonuniform intensity Normalization) proposed by Tustison and Gee (2009) was used to correct inhomogeneities. One advantage of this approach is that it does not depend on image pre-segmentation nor pulse sequence and it is robust to pathologies, which is the case of MR images with MS lesions. An example of the application of N4 is shown in Figure 5.5.



(a) (b)

Figure 5.5: Bias field effect on MR images before (a) and after (b) applying the N4 algorithm.

5.2.3 Image registration

An essential preprocessing stage is the spatial alignment between clinical images and probabilistic anatomical atlases. These atlases were used to constrain the segmentation area and reduce the number of ambiguities caused by voxels that have similar intensities but belong to different clusters. A study conducted by Freire and Ferrari (2014) analyzed the effect registration techniques had on MR images with MS and the authors verified that the lesions did not interfere with the registration procedure in a significant way, regardless of their load.

To preserve metadata of clinical images — such as origin, spacing and orientation — the clinical images were taken as fixed (or reference) images. This way, the transformations of the registration algorithm were applied to the atlases. In this work, we used the Nifty (Translational Imaging Group, 2014) method with B-Spline transformations and multi-resolution approach for non-rigid registration (MODAT et al., 2010).

5.3 Iterative segmentation

Clustering MS lesions using gray level intensities suffers from a false positive problem (SHIEE et al., 2010). Since lesions can present intensity levels that are close to those of other brain tissues, it becomes hard for the algorithm to distinguish and cluster them apart. For this reason, we applied an iterative segmentation approach (FREIRE; FERRARI, 2016). Doing so allowed a fine-tuning of the segmentation process to keep only the most hyperintense voxels clustered as MS lesions. The decision to preserve the most hyperintense voxels was made because of lesions commonly present e high-intensity profile in T2-w and FLAIR images, being the latter used in our segmentation pipeline. The iterative process, shown in Figure 5.6, was comprised of the following steps:

1. Let I_{seg} be a preprocessed image and BM a brain mask to constrain the segmentation region spatially.

2. Define a number of iterations (N_{iter}) and number of clusters $(N_{clusters})$ for each one of the iterations.

3. Segment I_{seg} in the region constrained by BM using the mixture model explained in Section 3.1 with the number of clusters defined for that iteration.

4. For each cluster, calculates its average gray level intensity.

5. Use regions from the $(N_{clusters} - 1)$ clusters with the highest intensity as the new binary mask BM to be used in the next iteration.

6. Decrease N_{iter}.

7. Repeat steps 3, 4, 5 and 6 until N_{iter} reaches zero.

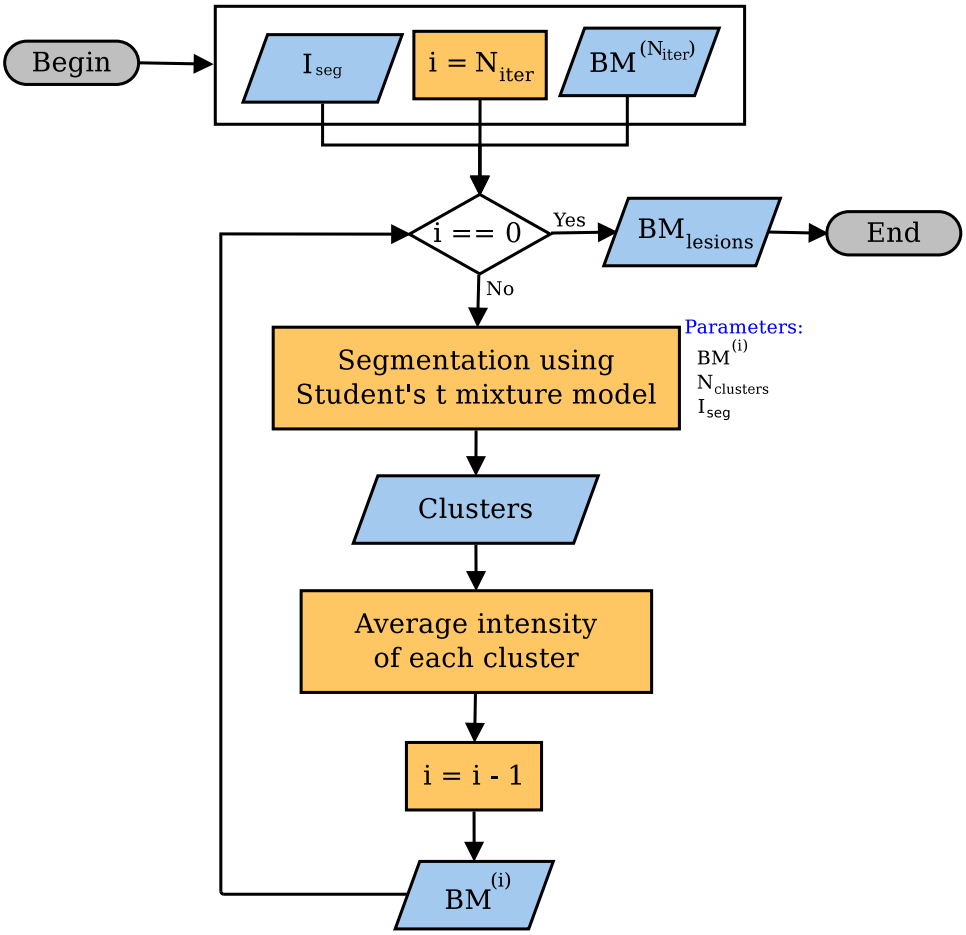


Figure 5.6: Flowchart of the iterative approach used in the MS lesion segmentation problem.

In this work, we set $N_{iter} = 5$, $N_{clusters} = 3$ for the first four iterations and $N_{clusters} = 2$ for the last one. These parameters were empirically defined using the images from the database described in Section 5.1.1 and provided the best compromise between execution time and segmentation accuracy.

The initial binary mask BM was created using the probabilistic anatomical atlases. The idea was to constrain the segmentation region to WM, which is where most MS lesions are located. However, the alignment between the atlases and clinical images is not perfect and the WM atlas may not encompass the whole WM region of the clinical image. In this sense, some lesion voxels located in the WM region of the clinical image may not present a high probability of belonging to the WM atlas due to alignment offsets. A threshold $t = 0.20$ was defined to

circumvent this problem and the locations of every voxel in the GM atlas with a probability smaller than t were used as input for the creation of the initial BM. The rationale was that voxels with a low probability of belonging to GM have, consequently, higher probabilities of belonging either to WM or CSF. Since voxels in the CSF region appear hypointense in FLAIR images, they are easily discarded from the lesion cluster during the iterative process. Moreover, hyperintense voxels that were located in the CSF atlas due to small misalignments, but in reality belonged to WM, were also part of the binary mask using this rationale.

At the end of the iterative process, a binary mask with lesions $BM_{lesions}$ was saved as the output. It is important to note that this iterative approach followed the hypothesis that MS lesions were representatives of the most hyperintense cluster in FLAIR images. The successive refinement of eliminating clusters with the lowest average intensity aimed at reducing the number of false positives. However, a limitation of this technique lies in the fact that it may also reduce the number of true positives in images where the variation of the MS lesions gray level intensities is significant enough to split them into two or more clusters.

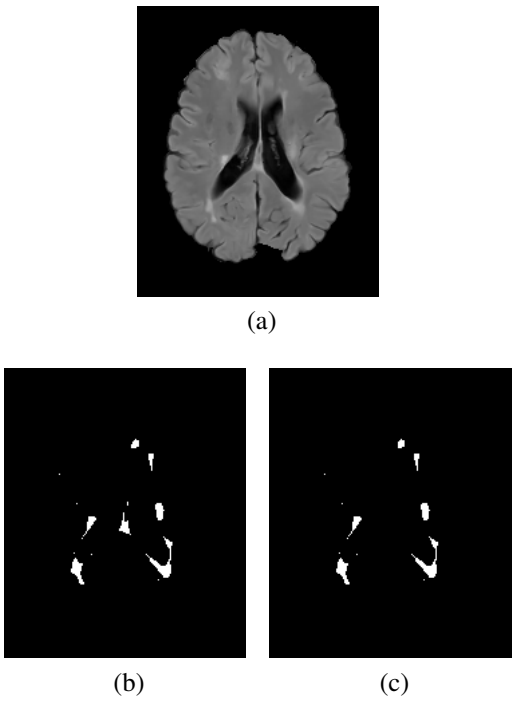
5.4 Segmentation post-processing

Post-processing techniques were used to remove false positives from the binary mask with MS lesions. Doing so involved detecting the midsaggital plane to eliminate voxels located in the third ventricle region and removing remaining voxels.

5.4.1 Midsaggital plane detection

Since the region between ventricles, known as the third ventricle (TV), in FLAIR images usually presents an intensity overlap with lesions (SHIEE et al., 2010), voxels in this particular area are mistakenly clustered as lesions and increase the number of false positives. Since actual lesions in this region are rare, we used a technique proposed by Ruppert et al. (2011) to detect the midsaggital plane (MSP). After that, a distance d_n was defined and every voxel with a distance to the plane smaller or equal to d_n were eliminated. In other words, given a voxel v and its distance to the midsaggital plane as $d(v) = d_{msp}$, v was removed from the binary mask if $d_{msp} \leq d_n$. In this work, we set $d_n = 8$.

This post-processing step proved to be effective for removing voxels in the TV region and keeping lesion voxels intact. A comparison between binary masks, before and after applying this approach, is shown in Figure 5.7.



(a) (b) (c)

Figure 5.7: Use of MSP to reduce false positives. (a) Original FLAIR image. (b) Binary mask before removing voxels located in the TV region. (c) Binary mask after removing voxels in the TV region using MSP detection.

5.4.2 Residue removal

The last step of the post-processing pipeline was removing residual voxels. In this work, a voxel was deemed as residue in the following way: let v be a voxel and N_v the neighborhood of v with radius r and size $s = (2 \times r + 1)^{dim}$, where dim is the dimension of the image. Since we focus on 3D images, $dim = 3$. We set $r = 1$ and consequently $s = 27$. We chose a neighborhood of radius 1 in order to analyze only the immediate surroundings of a voxel and thus try to remove only those that were likely isolated from the rest. In short, v was considered as a residual voxel if its neighborhood had less than $s/3$ active voxels; that is, voxels already present in the binary mask. A caveat of this approach is that it may also remove small lesions. However, as mentioned by Bento et al. (2017), very small lesions are commonly dismissed in automatic and manual approaches alike. And as shown in Table 5.2, the DSC metric increased with residue removal. In this sense, the net worth of applying residue removal is still positive.

5.5 Image enhancement

As mentioned before, FLAIR images are commonly used in the context of MS because lesions appear hyperintense in this particular image weight, making it easier for physicians to

Image source	DSC w/ residue	DSC w/o residue
Patient 1	0.675	0.698
Patient 2	0.850	0.861
Patient 3	0.674	0.681
Patient 4	0.685	0.722
Patient 5	0.657	0.680

Table 5.2: DSC results with and without residue removal. All comparisons were made between the automatic segmentation output from one time-point of each patient with the manual annotations from expert 1, which was randomly chosen.

identify them. Though lesions appear bright, their intensity profile overlaps with other brain tissues, mainly WM and GM (FREIRE; FERRARI, 2016), posing difficulties for accurate segmentation. In this sense, we proposed a lesion enhancement technique to dim down WM and GM regions and highlight hyperintensities. This was achieved by comparing each voxel's neighborhood average gray level intensity to some patches across a preprocessed image, creating a hyperintensity probability map. A byproduct of this proposal was that the enhancement could also be used to estimate a mask encompassing WM and MS lesions automatically.

Before generating the actual hyperintensity probability map, an intermediate image was created by applying the Sobel filter (DUDA; HART; STORK, 2000) to the input FLAIR image. Following the proposal in Roy, Bhuiyan and Ramamohanarao (2013), this intermediate image was generated as follows. Let $s = \{x, y, z\}$ be a particular spatial location, I_s the FLAIR intensity at s and g_s the gradient in the Sobel image at s. The edge and intensity information are combined as

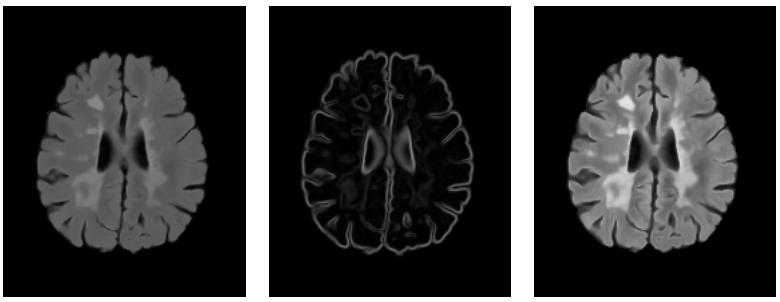
$$h(i) = \frac{1}{N} \sum_{s \in \{s | I_s = i\}} Prob\left(g \leq g_s\right), \tag{5.1}$$

where N is the total number of voxels with intensity i. In other words, Equation 5.1 goes over every FLAIR image intensity i and, given the probability density function (PDF) of the Sobel image, sums up the histogram bins that have a smaller frequency than g_s and then normalizes it.

After calculating $h(i)$, we compute the cumulative distribution function (CDF) of h as proposed in Roy, Bhuiyan and Ramamohanarao (2013):

$$q(i) = \sum_{k=1}^{i} h(k). \tag{5.2}$$

In the end, each $q(i)$ is used to replace each intensity i. An example of such intermediate image is shown in Figure 5.8.



(a) (b) (c)

Figure 5.8: Intermediate image generation. (a) FLAIR image, (b) Sobel image, (c) intermediate image.

5.5.1 Hyperintensity probability map

The hyperintensity probability map was created based on the intermediate image generated during the preprocessing stage detailed in Section 5.5. We devised an algorithm that automatically creates such map and does not depend on parameters that must be set by experimental observations, as opposed to (ROY; BHUIYAN; RAMAMOHANARAO, 2013; ROY et al., 2014).

The core principle behind this map is to compare each voxel neighborhood intensity with patches across different points in the image. The more times the voxel's neighborhood mean intensity is higher than the patches, then the more likely it is for that particular voxel to stand out and more likely for it to have a high hyperintense probability.

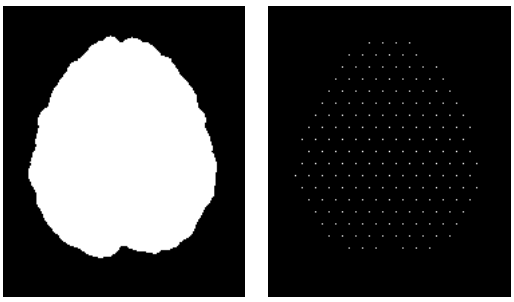
The first step to create the map was to define where each patch would be centered around. To do that, we defined a point net for each slice following the algorithm proposed in (ROY et al., 2014), which uses the combination of sines and cosines to distribute points across a slice evenly. Let $p = \{x, y, z\}$ be the coordinates of a candidate point. We then create new points $p' = \{x', y', z\}$ with

$$\begin{aligned} x' &= x + r\cos\theta \\ y' &= y + r\sin\theta \end{aligned}, \tag{5.3}$$

where θ is the angle and $r = 10$ is the radius. We set θ to zero and increase it by 60 degrees six times to complete a whole circumference. The six newly defined points become candidate points, and the process is repeated until no new point is found.

After defining such points, we cropped the net using a brain mask in order only to keep points that are inside our ROI. This procedure was done by simply purging points outside the brain mask. An example of a final point net set P for a particular slice is shown in Figure 5.9.

Now, let μ and σ be the mean and standard deviation of the whole intermediate image



(a) (b)

Figure 5.9: Point net generation. (a) Brain mask, (b) cropped point net. Each patch is centered around one point in (b).

within the brain mask. Then, for each voxel, we calculated its neighborhood mean intensity as

$$\mu_v = \frac{1}{N_v} \sum_{k=1}^{N_v} i_k, \tag{5.4}$$

where μ_v is the mean neighborhood intensity of voxel v, N_v is the number of neighbors of v and i_k is the intensity of neighbor k. The neighborhood size was defined as $3 \times 3 \times 3$ to maintain a good trade-off between sharpness and smoothness. The same rationale was used for the patches: the mean intensity was calculated as in Equation 5.4, thus creating μ_p for each patch.

Finally, we created a score S_v as

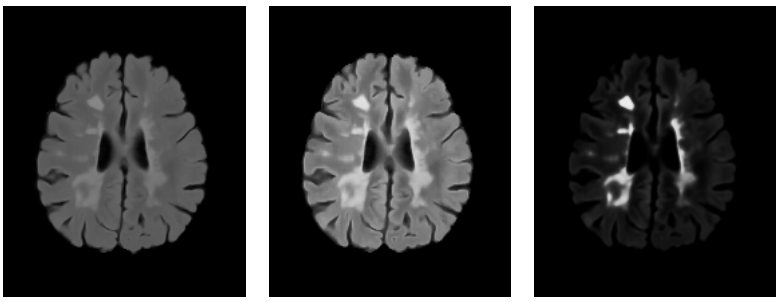
$$S_v = \frac{1}{|P|} \sum_{p \in P} \delta(\mu_v, \mu_p), \tag{5.5}$$

where $|P|$ is the cardinality of the patch set and

$$\delta(\mu_v, \mu_p) = \begin{cases} 1, & \text{if } \mu_v - \mu_p \geq \sigma \\ 0, & \text{otherwise} \end{cases}. \tag{5.6}$$

In other words, if the difference of intensity between a candidate voxel neighborhood and a patch was greater than or equal to the standard deviation of the whole image, then it was considered a hit. Otherwise, it was a miss. By doing so, voxels with bright neighborhoods were enhanced while other regions and tissues were dimmed out. Moreover, since we normalized the score S_v, each voxel remained in the range $[0,1]$, which also served as a hyperintensity probability indicator. An example of the map is shown in Figure 5.10.

It is important to note that our approach did not require a hard threshold for Equation 5.6 or a fixed number of patches, as in (ROY; BHUIYAN; RAMAMOHANARAO, 2013). Instead, the threshold was automatically calculated with respect to the standard deviation. Though simple,



(a) (b) (c)

Figure 5.10: Hyperintensity map generation. (a) FLAIR, (b) intermediate image, (c) hyperintensity map.

this was a significant improvement, since the main problem of using a hard threshold is that even normalized, intensities inherently vary from image to image. In this sense, a soft threshold such as the one we proposed in this work offers a better option for the enhancement of hyperintensities for it can adapt to each image intensity profile.

5.5.2 White matter mask estimation

The white matter region usually comprises most MS lesions (COMPSTON; COLES, 2008). An automatic brain segmentation into three clusters (WM, GM, and CSF) based on gray level intensities is most certainly going to mix lesions and cluster them as GM, WM or both (BATTAGLINI; JENKINSON; STEFANO, 2012; VALVERDE et al., 2015). In this context, being able to estimate a mask that encompasses both white matter tissue and MS lesions can help narrow down the ROI and increase the accuracy of lesion segmentation. To do so, we leveraged the fact that the map described in Section 5.5.1 could also be interpreted as a probability map and used it to get an estimate of such mask.

In this work, we made use of the Student's t-mixture model proposed in Freire and Ferrari (2016) and used T1-w and FLAIR images from each time-point to segment the brain into three different clusters and get an initial WM mask, herein referred to as $WM_{initial}$. To automatically identify the WM cluster from others, we used the WM probability map described in Section 5.1.3, averaged it over each cluster and selected the one with the highest WM probability.

Since the clinical database described in Section 5.1.1 did not provide WM ground truths, we created our own using a straightforward approach. Given any $WM_{initial}$, we simply merged it with the lesions ground truth to get the whole WM region in one single mask, herein referred to as WM_{whole}. Considering that each time-point had two different lesions ground truth, we created two WM masks for each time-point as well.

The actual WM estimation took place as follows. First, we calculated the mean (μ_{HI}) and standard deviation (σ_{HI}) of the region defined by $WM_{initial}$ on the hyperintensity map and the mean (μ_{prob}) of the region defined by $WM_{initial}$ on the WM probability atlas. The idea was to expand $WM_{initial}$ by considering voxels that were not part of the mask yet by analyzing $3 \times 3 \times 3$ neighborhoods centered around these voxels to verify their potential for being included. The expansion itself occurred by incorporating voxels that seemed as outliers; more precisely, voxels with mean neighborhood values greater than $\mu_{HI} + 1 \times \sigma_{HI}$ in the hyperintensity map and greater than μ_{prob} in the probability atlas. An example of the output of this estimation is shown in Figure 5.11.

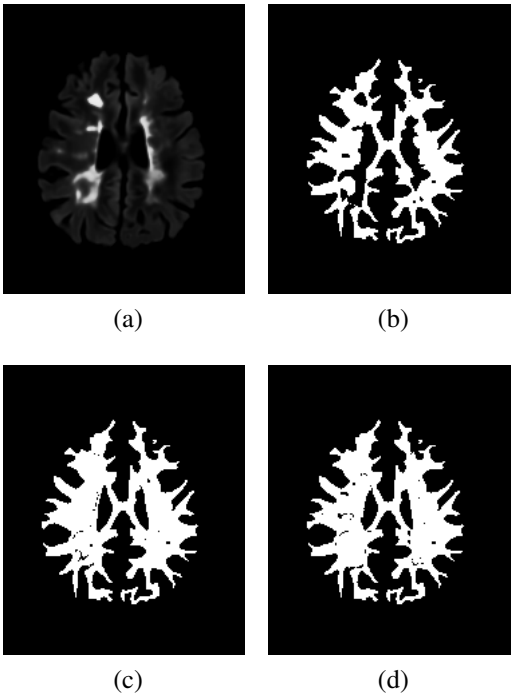


Figure 5.11: White matter mask generation. (a) Hyperintensity map, (b) $WM_{initial}$, (c) $WM_{groundtruth}$, (d) WM_{estim}.

5.5.3 Pure WM and GM clusters

To estimate the intensity profiles of white matter and gray matter clusters without lesions, we did the following. For each time-point, we segmented them into three clusters using the technique described in Section 3.1 and automatically identified the white and gray matter clusters by analyzing their mean intensities on white and gray matter probabilistic atlases. The cluster with highest white matter atlas mean intensity was taken as the white matter cluster ($WM_{initial}$); the same rationale was used for the gray matter cluster ($GM_{initial}$). Then, for each expert annotation, we simply excluded every voxel that had any intersection with the lesion ground truth

and these two clusters. Formally, $WM_{pure} = WM_{initial} - (WM_{initial} \cap GT_{E \in \{\text{Expert 1, Expert 2}\}})$ and $GM_{pure} = GM_{initial} - (GM_{initial} \cap GT_{E \in \{\text{Expert 1, Expert 2}\}})$. By doing so, we were able to obtain so-called "pure" WM and GM clusters, which were then used to calculate their intensity profiles and compare them to lesion profiles in Section 6.2.1.

5.6 Active lesions classification

As mentioned in Chapter 1, the distinction of active and nonactive lesions is currently achieved by the injection of a Gadolinium-based contrast in the patient's body a few minutes before the MRI procedure. However, the more injections a patient takes throughout their life, the more difficult it becomes to eliminate the contrast from their body. In this sense, we decided to extract a number of features and use the XGBoost classifier to tell these two classes apart with no aid whatsoever of Gadolinium injections.

To do so, we applied the same preprocessing pipeline shown in Figure 5.3 to FLAIR images from both datasets, except we also applied a histogram matching algorithm proposed by Nyul, Udupa and Zhang (2000) after the intensity inhomogeneity step to bring all images to a standard intensity domain. As a reference for this histogram matching, we chose the FLAIR scan in the clinical/nonactive dataset with the heaviest lesion load (32,3 ml). The rationale behind this choice was to make lesions as representative as possible when matching histograms in both active and nonactive datasets. Finally, all images were rescaled to the $[0, 255]$ interval.

After completing the preprocessing pipeline, we extracted features to aid in the distinction between active and nonactive lesions. Such features were be divided into two groups, as shown in Figure 5.12.

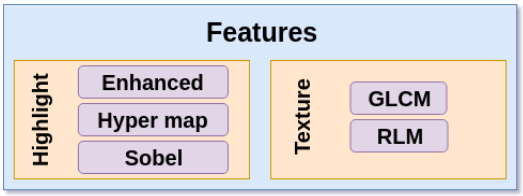
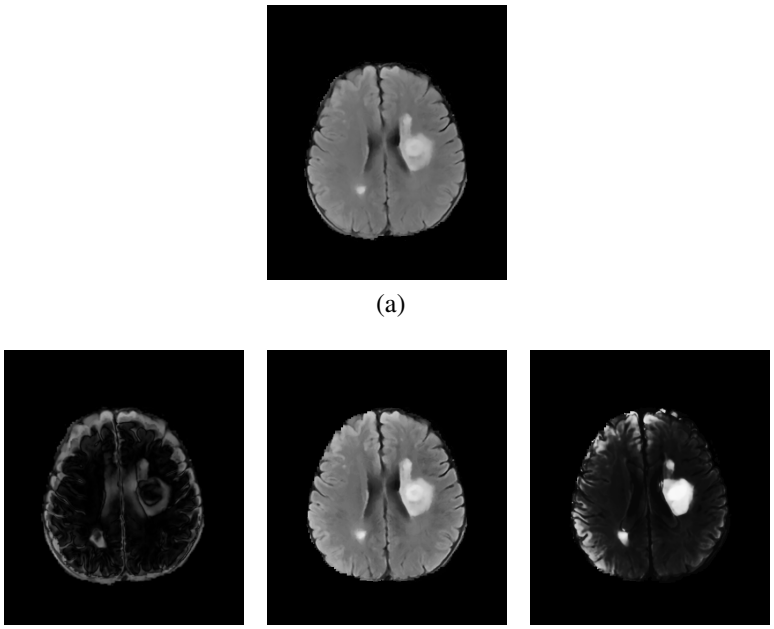


Figure 5.12: Classes of extracted features.

The highlight branch was comprised of Sobel and two other feature images, enhanced and hyperintensity map, generated by the algorithm proposed in Section 5.5. In short, these enhanced and hyperintensity map feature images were created by first scattering the histogram in order to attenuate mean intensities and enhance "tail" ones, using it to further highlight hyperintensities through patch comparisons. An example of the images generated from this branch

are shown in Figure 5.13.



(a)

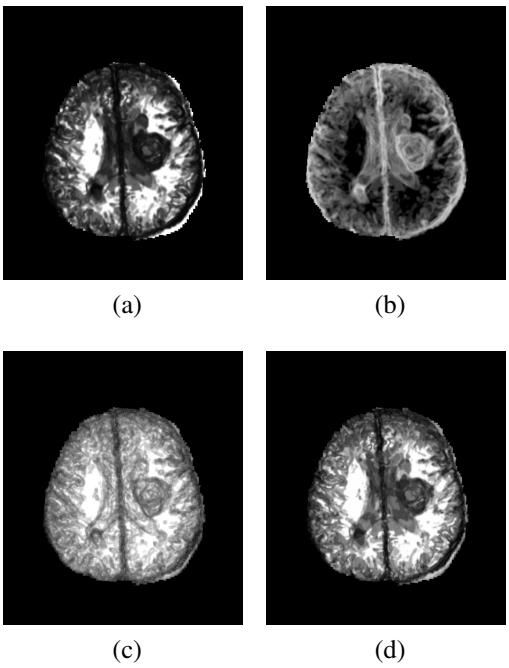
(b) (c) (d)

Figure 5.13: Highlight features extracted from an active (a) FLAIR image yielding (b) Sobel image, (c) Enhanced image and (d) Hyperintensity map.

The other branch of feature extraction concerned textures and was divided into two techniques: GLCM and RLM (both explained in Sections 3.4.1 and 3.4.2, respectively). In this work, we set the distance to a radius of 2 and 13 directions, scaled analogously from the default 4 directions - 0, 45, 90 and 135 degrees - used in 2D images, as shown in Figure 3.10. Each texture map was calculated as the average of all 13 directions. We decided to set the radius to 2 in order to compensate for the varying range of shapes and sizes each MS lesion can take. An example of features extracted from GLCM and RLM are shown in Figure 5.14.

After completing feature extraction, a total of 22 attributes were selected to be used in the classification step: 8 GLCM features, 10 RLM features, histogram matched FLAIR image, Sobel image, enhanced image and hyperintensity map.

One can argue that the differences in image resolution between the active and nonactive datasets could lead to differences in the texture outputs. While this is a well-grounded observation, we opted to preserve image resolutions as they were for several reasons. First, any up or downsampling procedure would involve alterations in the experts' annotations, which could lead to invalid lesion masks in the new resolution space. Second, we wanted our classification algorithm to work as close as possible to a real world situation regarding the identification of active lesions. And third, the main distinguishing and qualitative features used by the expert in the active dataset were present in the XY plane, which is already encompassed by the offsets



(a) (b)

(c) (d)

Figure 5.14: Textures extracted from the patient with active lesions depicted in Figure 5.13. (a) GLCM Energy, (b) GLCM Entropy, (c) RLM GLN, (d) RLM LRE.

shown in Figure 3.10.

5.7 Final considerations

This chapter presented the database, preprocessing, clustering rationale and post-processing steps used for MS lesion segmentation in MR images. The iterative approach was chosen as a way to partially overcome the intensity overlap between lesions and other brain tissues and thus improve segmentation accuracy. To further improve the distinction among clusters, we also presented a methodology for enhancing hyperintensities in FLAIR, which also had a byproduct of providing ways to estimate a WM mask that can be used to constrain the region of interest in an image. Additionally, we also discussed our classification approach to tell the active and nonactive lesion classes apart without the injection of Gadolinium-based contrasts in one's body prior to a MRI procedure.

The results of the segmentation, distinction provided by the enhancement and lesion classification are discussed in Chapter 6.

Chapter 6

RESULTS AND DISCUSSION

This chapter presents quantitative results on the iterative approach applied to the context of MS lesions in MRI, the distinction given by image enhancement on FLAIR images and the classification of active and nonactive lesions. We compare our results with other works and provide a discussion on strengths and weaknesses of our pipeline.

6.1 Segmentation results

The preprocessing, segmentation and post-processing steps explained in Chapter 5 were used to automatically segment MS lesions in the 2015 Longitudinal MS Lesion Segmentation Challenge (CARASS et al., 2017). As mentioned in Section 5.1.1, each time-point had two manual delineations performed by two experts. The results of our segmentation were compared with both of them. An example of the iterative process is shown in Figure 6.1.

Two segmentation outputs are shown in Figures 6.2 and 6.3. 3D rendered images were generated from the automatic segmentation and manual delineations and are also shown in Figures 6.4 and 6.5.

We extracted the mean and standard deviation of DSC, TPR, FPR and VD for each patient. The results are shown in Tables 6.1 and 6.2. It is important to note in Figure 6.1 that the initial binary mask was created using the probabilistic anatomical atlases to constrain the segmentation region as close to WM as possible.

The most significant differences in the values of the metrics between the ground truths from experts 1 and 2 in Tables 6.1 and 6.2 were observed for patient 4. As shown in Table 6.3, the agreement between the manual delineations of the experts in this case (DSC $= 0.612 \pm 0.019$) presented the lowest DSC among all patients, which indicates a high inter-experts variability.

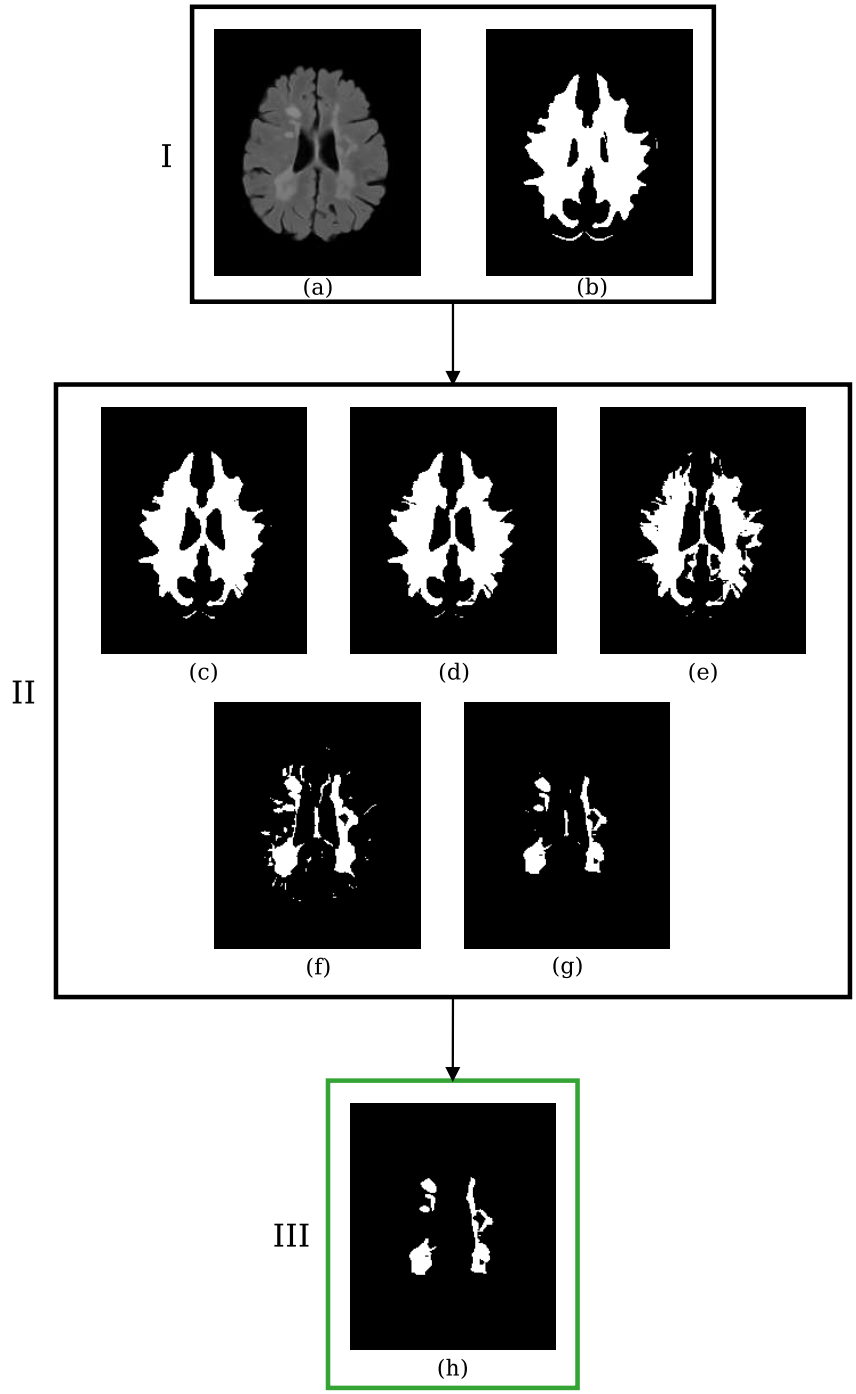


Figure 6.1: Iterative segmentation approach. In the first step (I), the image to be segmented (a) and the initial binary mask (b) are provided as input for the algorithm. In the second step (II), the iterative segmentation process is shown in binary masks (c)-(g). In the third step, the final lesion binary mask (h) is obtain after post-processing.

In this scenario, it is reasonable to expect the automatic segmentation to yield low agreement metrics when compared with manual annotations. An example of disagreement between experts in a time-point from patient 4 is shown in Figure 6.6.

The patients' lesion volume in the 2015 Longitudinal MS Lesion Segmentation Challenge

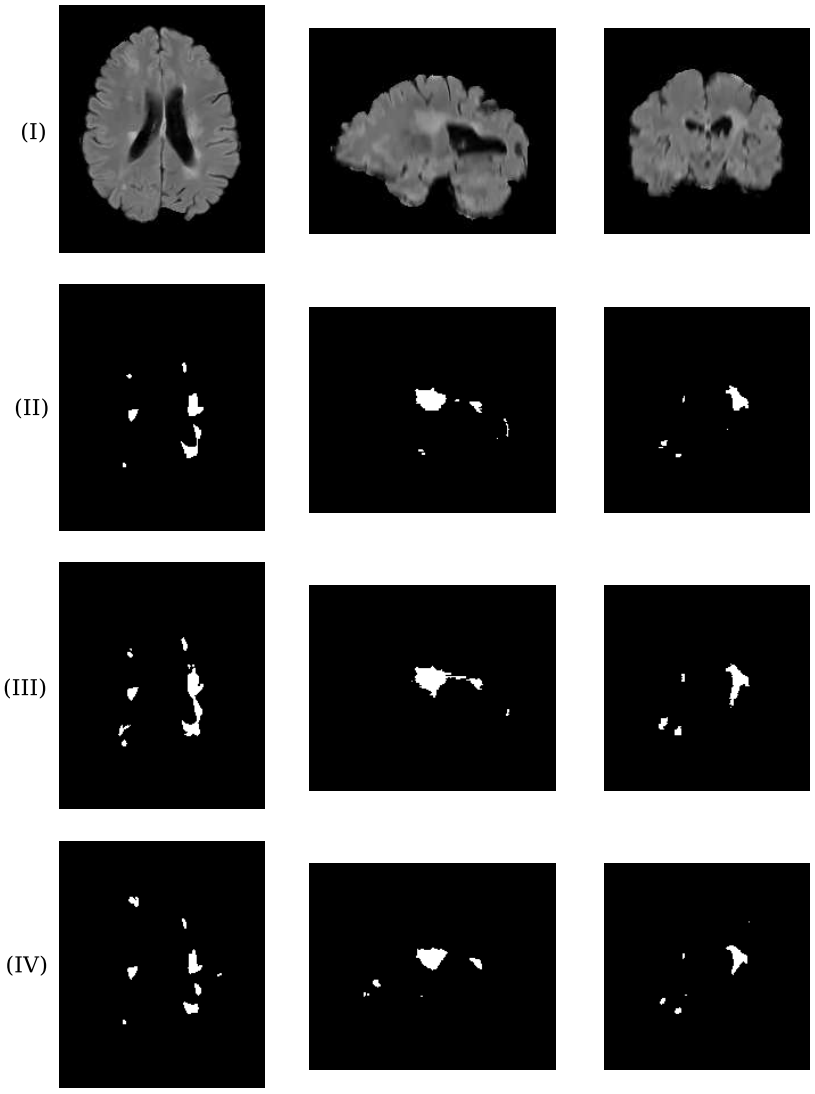


Figure 6.2: Final segmentation of a time-point from patient 1 with axial (left), saggital (middle) and coronal (right) views. In (I), FLAIR images. In (II), manual annotation from expert 1. In (III), manual annotation from expert 2. In (IV), automatic lesion segmentation.

ranged from 1.950 *ml* to 34.719 *ml*. Given the DSC, TPR and FPR metrics, the iterative approach presented consistent segmentation results in this range. However, as explained in Section 5.3, the iterative approach took the most hyperintense cluster as MS lesions. This may have excluded lesion voxels that did not fit in the hyperintense hypothesis, thus lowering the TPR and increasing the VD. A comparison made between the mean and standard deviation values of lesions annotated by experts 1 and 2 and segmented by our proposal in images from patient 1 are shown in Table 6.4.

In Table 6.4, it is possible to observe that the automatic segmentation had the lowest standard deviation, indicating our approach segmented a narrower portion of MS lesions. Moreover, except for time-point 2, our algorithm presented the highest mean values. Both high mean and low standard deviation indicate our technique indeed segmented MS lesions, but it was not able

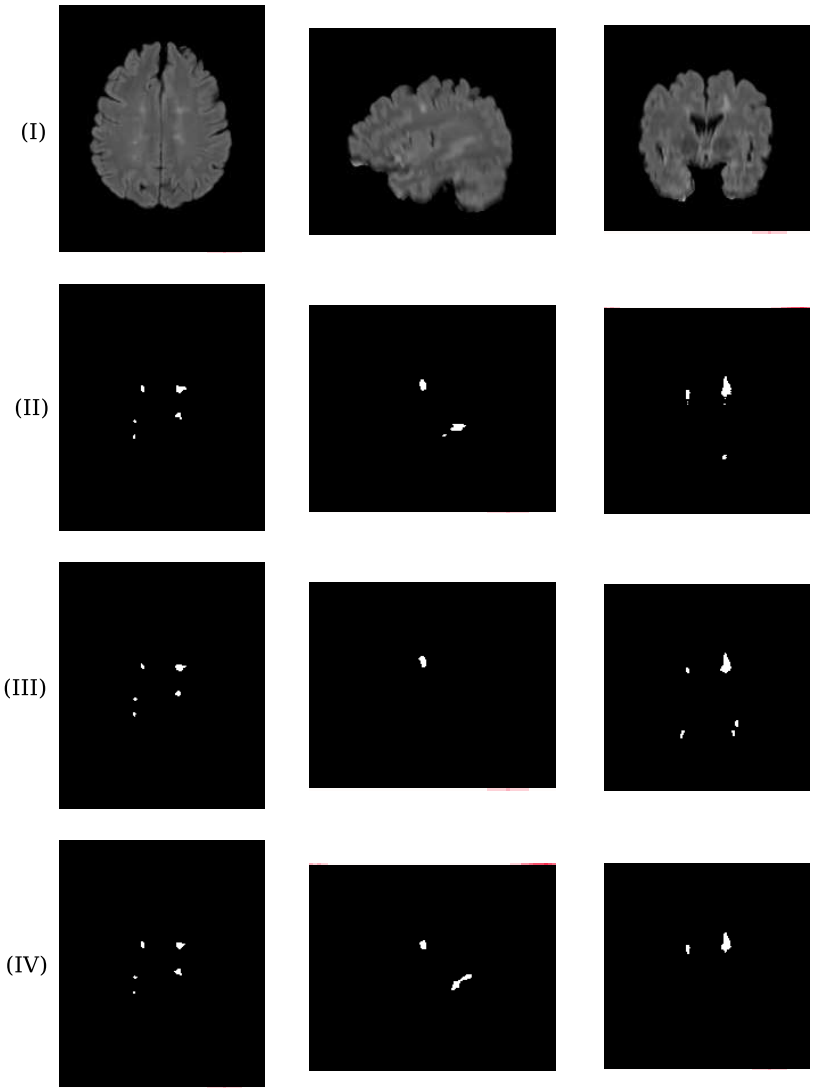


Figure 6.3: Final segmentation of a time-point from patient 5 with axial (left), saggital (middle) and coronal (right) views. In (I), FLAIR images. In (II), manual annotation from expert 1. In (III), manual annotation from expert 2. In (IV), automatic lesion segmentation.

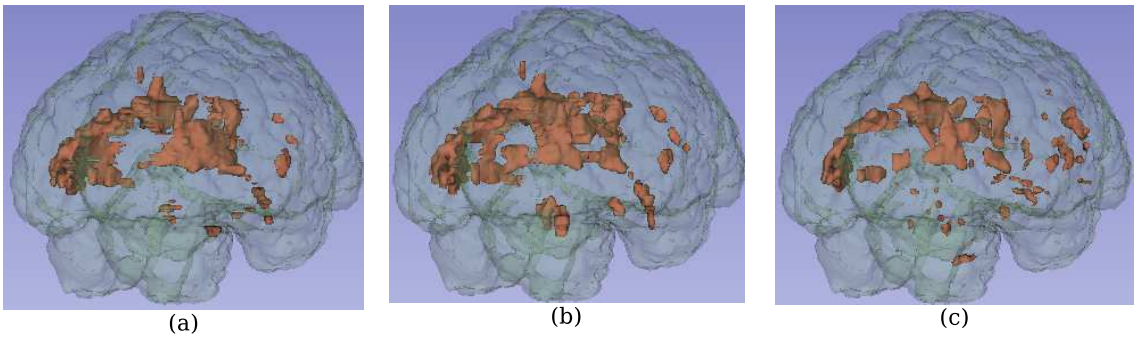
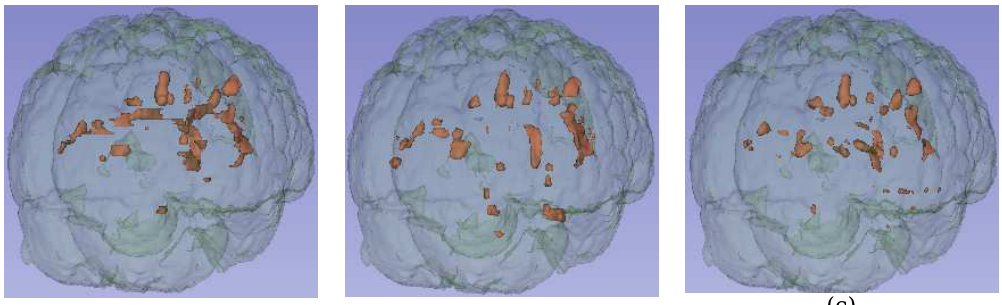


Figure 6.4: 3D rendered lesion images of a time-point from patient 1. In (a), manual annotations from expert 1. In (b), manual annotations from expert 2. In (c), automatic lesion segmentation.

to include lesion voxels that were located in a lower intensity range. The same behavior was observed for every other time-point from every other patient. An example of this problem is



(a) (b) (c)

Figure 6.5: 3D rendered lesion images of a time-point from patient 5. In (a), manual annotations from expert 1. In (b), manual annotations from expert 2. In (c), automatic lesion segmentation.

Source	# images	DSC	TPR	FPR	VD
Patient 1	4	0.644 ± 0.056	0.607 ± 0.164	0.259 ± 0.200	0.328 ± 0.09
Patient 2	4	0.716 ± 0.114	0.604 ± 0.181	0.05 ± 0.06	0.336 ± 0.230
Patient 3	5	0.642 ± 0.056	0.581 ± 0.074	0.245 ± 0.285	0.338 ± 0.115
Patient 4	4	0.622 ± 0.080	0.557 ± 0.159	0.220 ± 0.250	0.390 ± 0.120
Patient 5	4	0.634 ± 0.060	0.595 ± 0.097	0.272 ± 0.065	0.136 ± 0.146

Table 6.1: Segmentation results (mean $\pm$ std dev) using expert 1's delineations as *ground truth*.

shown in Figure 6.7.

We also calculated the Person's correlation coefficient (GAYEN, 1951), commonly called r, to assess the volumetric agreement between the automatic segmentation and the manual annotations performed by experts 1 and 2. For both of them, the *p-value* was smaller than 0.00001.

Considering the automatic segmentation output and the annotation from expert 1, $r = 0.8813$; similarly, for expert 2, $r = 0.8871$. The Pearson's correlation coefficient was also calculated for experts 1 and 2 with $r = 0.9841$ to expand the comparison. The scatter plots of this data are shown in Figures 6.8, 6.9 and 6.10.

The r and *p-values* related to the automatic segmentation indicated a strong positive correlation between the computed volumes from our approach and the manual annotations. By considering these results along with those shown in Tables 6.1 and 6.2, it is possible to observe that the automatic MS lesion segmentation proposed in this work indeed had a good spatial and volumetric agreement with the ground truths. Comparatively, the correlation between experts was higher compared to the automatic segmentation versus each annotation. This difference can be explained by the fact that our proposal segmented lesions with the highest gray level intensities, discarding those with a lower intensity.

Source	# images	DSC	TPR	FPR	VD
Patient 1	4	0.641 ± 0.083	0.573 ± 0.176	0.182 ± 0.121	0.345 ± 0.09
Patient 2	4	0.701 ± 0.095	0.579 ± 0.144	0.054 ± 0.061	0.364 ± 0.195
Patient 3	5	0.577 ± 0.086	0.464 ± 0.127	0.123 ± 0.153	0.410 ± 0.262
Patient 4	4	0.478 ± 0.043	0.367 ± 0.095	0.155 ± 0.186	0.474 ± 0.277
Patient 5	4	0.586 ± 0.055	0.592 ± 0.110	0.411 ± 0.098	0.153 ± 0.107

Table 6.2: Segmentation results (mean $\pm$ std dev) using expert 2's delineations as *ground truth*.

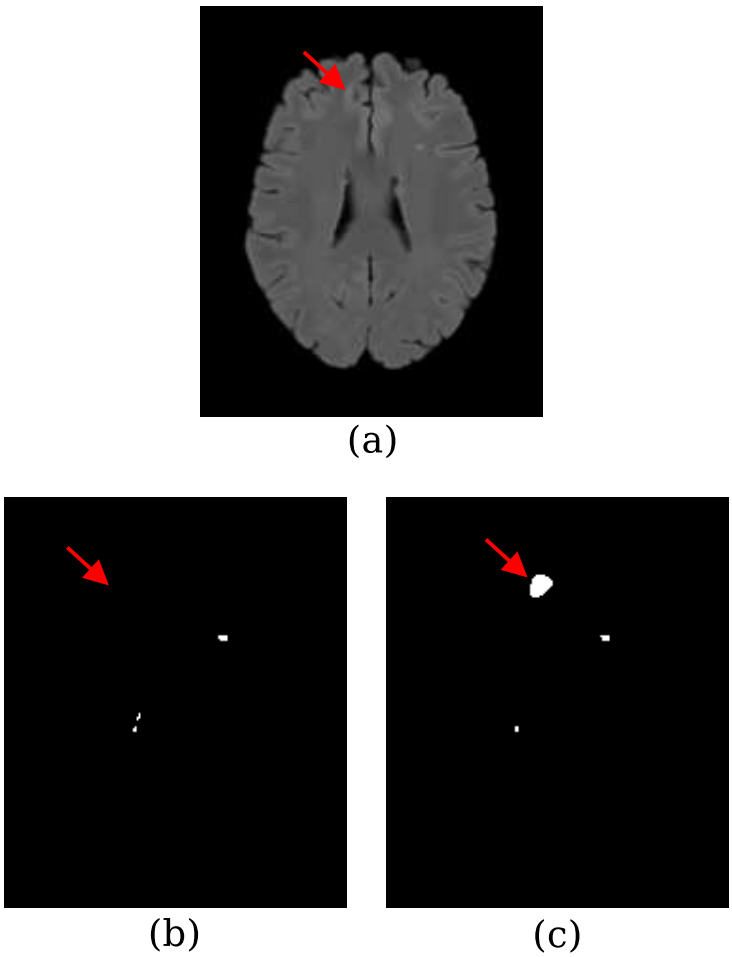


Figure 6.6: Example of disagreement between experts in a time-point from patient 4. In (a), axial view of FLAIR. In (b), manual annotation from expert 1 and in (c), manual annotation from expert 2. The red arrow indicates a region delineated only by expert 2 as being a lesion.

An advantage of our approach compared to some of the works described in Section 4.1 lies in the fact that it did not require a training step; each segmentation was performed in a stand-alone manner. Supervised techniques usually need a thorough training stage and additional data (such as a healthy population) to perform the segmentation. This restriction can be a problem when dealing with images from different hospitals or medical centers since they would have to

Source	# images	DSC (agreement between experts)
Patient 1	4	0.779 ± 0.017
Patient 2	4	0.846 ± 0.036
Patient 3	5	0.768 ± 0.076
Patient 4	4	0.612 ± 0.019
Patient 5	4	0.652 ± 0.048

Table 6.3: DSC metrics (mean $\pm$ std dev) for experts 1 and 2.

Time-point	$\mu_{GT1} \pm \sigma_{GT1}$	$\mu_{GT2} \pm \sigma_{GT2}$	$\mu_{seg} \pm \sigma_{seg}$
1	139.88 ± 12.49	137.58 ± 12.44	145.32 ± 9.27
2	112.20 ± 8.71	110.80 ± 9.05	110.52 ± 7.45
3	103.23 ± 8.31	102.978 ± 8.65	109.02 ± 6.26
4	118.20 ± 9.72	116.84 ± 10.23	123.37 ± 7.56

Table 6.4: Intensity comparison (mean $\pm$ std dev) of MS lesions among ground truth 1 (GT1), ground truth 2 (GT2) and the automatic segmentation output for images of patient 1.

go through a normalization stage. Such stage may introduce bias to the results since the training set is usually not large enough to encompass anatomical and physiological variabilities between different patients (DESPOTOVIC; GOOSSENS; PHILIPS, 2015).

The winner of the 2015 Longitudinal MS Lesion Segmentation Challenge was the work of Vaidya et al. (2015). The authors used neural networks and patches to perform the segmentation. Neural networks had to be trained before the actual segmentation could take place. Since the authors presented results for a few time-points of only one patient, it was not possible to thoroughly compare their work with ours. However, even with their DSC values being higher than those obtained by our technique, the TPR and FPR of our approach were better than the authors'. Such result is an indication that our proposal had smaller classification errors.

An overview of the results from the teams who participated in the challenge and our results are shown in Table 6.5.

It is important to note that our approach used only FLAIR images. Other image weights may bring more information to the model; however, preliminary tests using T1-w, T2-w and FLAIR and T1-w and FLAIR were conducted, and no significant accuracy gains were observed. Moreover, using only one image weight reduced the computational cost of the technique and

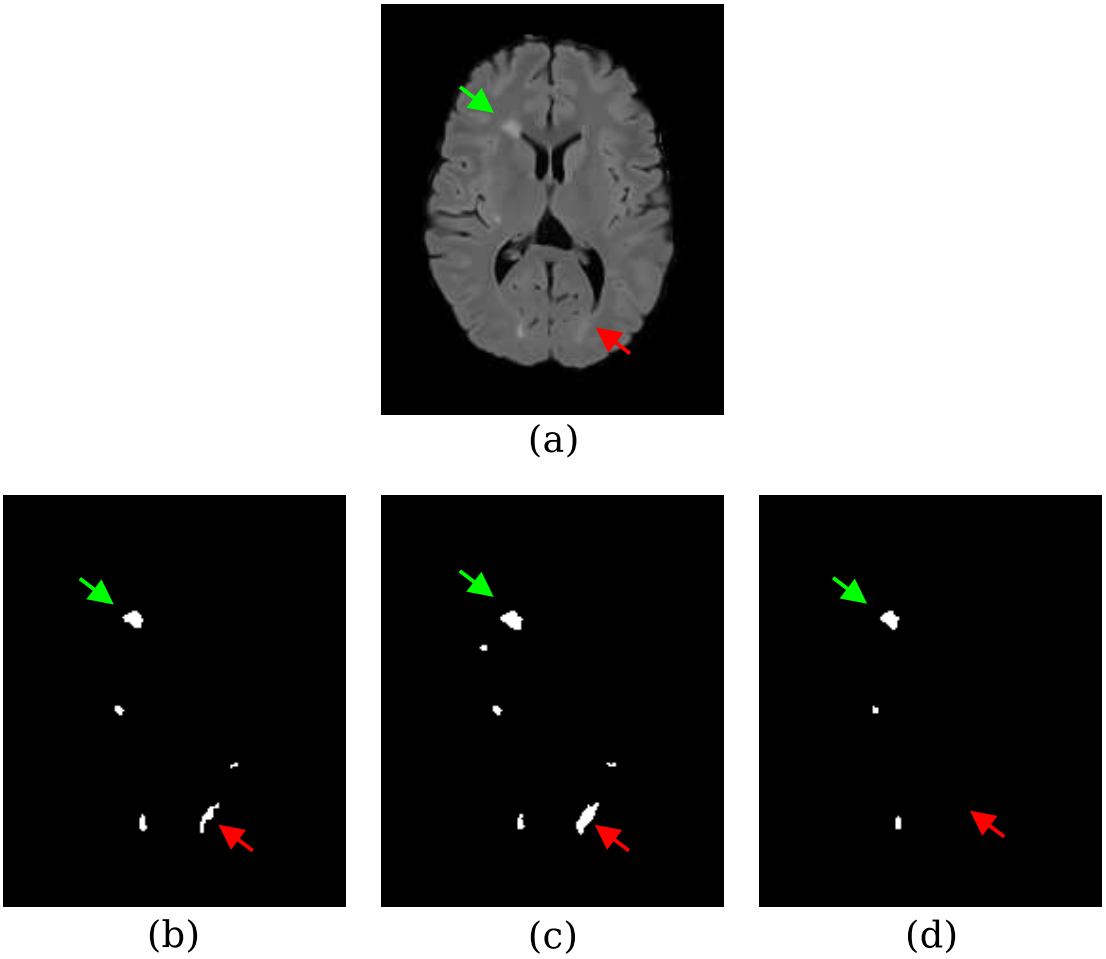


Figure 6.7: In (a), axial view of FLAIR. In (b)-(d), manual annotation from expert 1, 2 and automatic lesion segmentation, respectively. Green arrows indicate a hyperintense lesion identified in all three masks. Red arrows in the sub-figures indicate a less hyperintense lesion that was not identified by our proposed algorithm.

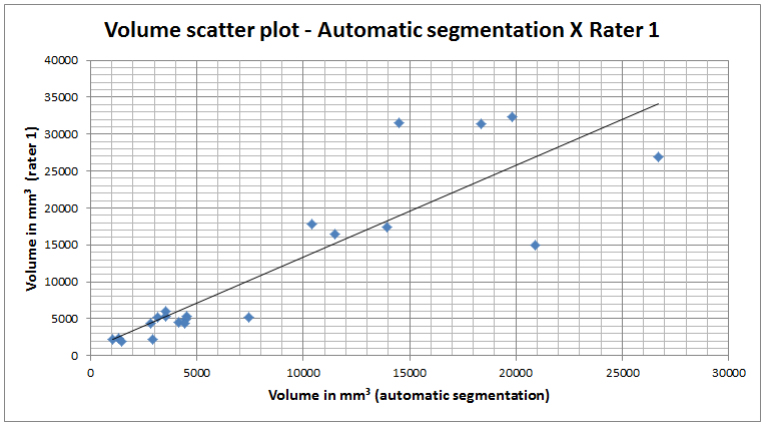


Figure 6.8: Scatter plot of volumes calculated for the automatic segmentation (horizontal axis) and annotation from expert 1 (vertical axis). For this data, $r = 0.8813$ (p-value < 0.00001).

decreased the execution time of the algorithm.

Overall, the results from our automatic segmentation presented a good agreement with man-

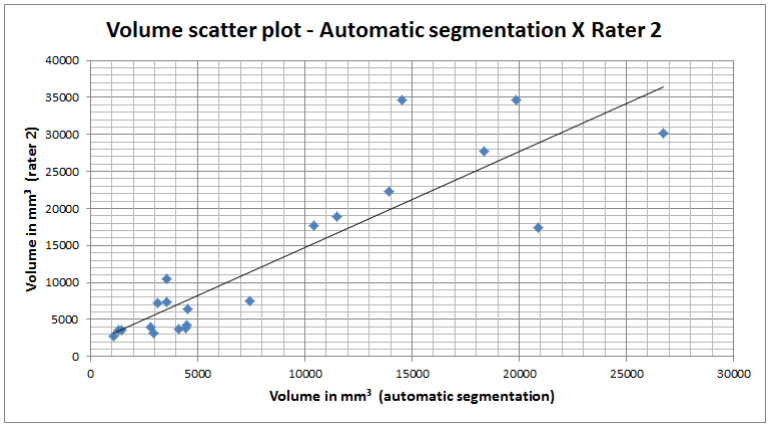


Figure 6.9: Scatter plot of volumes calculated for the automatic segmentation (horizontal axis) and annotation from expert 2 (vertical axis). For this data, $r = 0.8871$ **(p-value** < 0.00001**).**

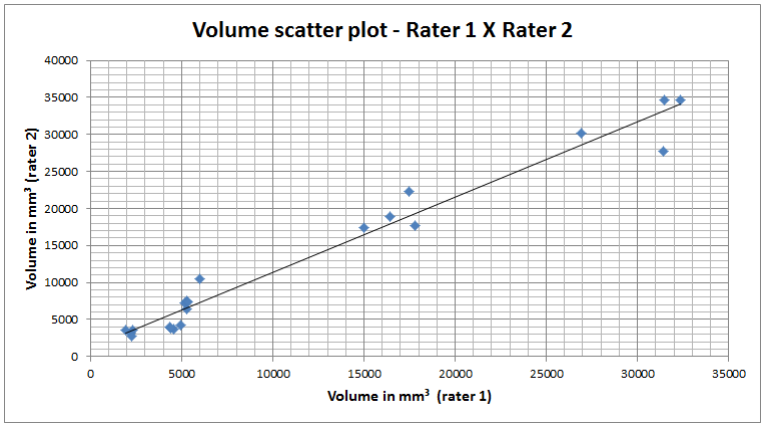


Figure 6.10: Scatter plot of volumes calculated for the annotations of expert 1 (horizontal axis) and expert 2 (vertical axis). For this data, $r = 0.9841$ **(p-value** < 0.00001**).**

ual annotations (especially from expert 1) and were able to keep consistency between images with different lesion loads. Besides, it did not require any training stage, so the segmentation of images from different acquisition protocols was not affected by any bias introduced by training images.

A difficulty faced during the development of the algorithm was the agreement variability between experts. This fact created problems because it failed to provide a solid foundation that could be used as a guide for adjusting the parameters of our technique. Ideally, a greater number of annotations made by different experts in the same database could help mitigate the variability problem. In this scenario, we could have a better grasp on the lesion annotations and improve the results of our algorithm based on the agreement of the majority of experts.

One limitation of our proposal lied in the image alignment stage. Since registration between images is hardly perfect, lesions located in regions that are not well aligned are likely to be left out during the region constraint step. The GM atlas was used as a way to include WM and

Authors	Technique	Results	Image weights	*Ground truth*
Jesson e Arbel	Random forest	DSC = 0.70/0.68 TPR = 0.61/0.50	T1-w, T2-w, FLAIR	GT1/GT2
Maier e Handels	Random forest	DSC = 0.70/0.65 TPR = 0.53/0.38	T1-w, T2-w, PD, FLAIR	GT1/GT2
Vaidya et al.	Neural network	DSC = 0.80 TPR = 0.40	T1-w, T2-w, PD, FLAIR	GT1 (only from patient 2)
Tomas-Fernandez e Warfield	GMM	DSC = 0.62 TPR = 0.53	T1-w, T2-w, FLAIR	Not clear
Prados et al.	Fast patch matching	DSC = 0.55 ± 0.14 TPR = 0.72 ± 0.10	T1-w, T2-w, PD, FLAIR	GT1 $\cap$ GT2
Freire e Ferrari	SMM-iter	DSC = 0.65/0.59 TPR = 0.58/0.51	FLAIR	GT1/GT2

Table 6.5: DSC and TPR results of works submitted to the 2015 Longitudinal MS Lesion Segmentation Challenged and ours. GT1 and GT2 are the ground truths from experts 1 and 2, respectively.

CSF voxels in the initial binary mask to overcome this issue, being CSF voxels easily excluded because of their hypointense characteristic in FLAIR images. Moreover, this constraint also reduces the number of false positives by eliminating voxels that have intensities close to MS lesions but are located outside the WM region.

Another limitation was the fact that MS voxels in a lower intensity range are likely to be left out of the MS cluster. This problem happens because our approach hypothesized that MS lesions represent the most hyperintense cluster in FLAIR images. If a significant variability in lesion intensity is observed, the less intense ones may not follow our hypothesis and, consequently, not be segmented appropriately. However, even with this caveat, our technique was comparable to algorithms applied to the same database, yielding similar or better results than most other works.

Finally, the number of iterations and clusters on each iteration were parameters that had a significant weight on the segmentation result. In this sense, an approach to automatically choose these parameters would improve the robustness of the whole technique and serve as an aid for the problem of excluding lesion voxels that are not as hyperintense as the lesion cluster.

6.2 Image enhancement results

In this section, we present results regarding the brightness intensity profile of MS lesions compared to gray matter and white matter tissues in FLAIR, intermediate and hyperintensity map images for each patient.

6.2.1 Brightness profile

Since the 2015 Longitudinal MS Lesion Segmentation Challenge provided two ground truths for each time-point, we extracted the intensity profiles for both annotations. We rescaled all images to the $[0, 1]$ interval, averaged the white matter, gray matter and lesion profiles and also calculated the standard deviation for each patient. The results are shown in Figures 6.11 and 6.12 for experts 1 and 2, respectively.

Each bar in Figures 6.11 and 6.12 represents the mean lesion intensity over the mean intensity of a given tissue (gray matter or white matter) in a particular image type (FLAIR, intermediate and hyperintensity map). For instance, the FLAIR WM bar in Figure 6.11(a) must be interpreted as "the average MS lesion profile in FLAIR images from Patient 1 is approximately 25% brighter than the average WM tissue intensity for the same image type and patient". This allows a direct comparison between tissues and images.

The results in Figures 6.11 and 6.12 indicated a significant difference in the MS lesion intensity profile in the hyperintensity map compared to FLAIR and the intermediate image described in Section 5.5. These outcomes are significant since they provide quantitative background to show the discriminative features of the HI map.

It is possible to note that the lesion intensity profile was more similar to the gray matter than to white matter. The difference in intensity between MS lesions and GM and WM in FLAIR and intermediate images was minimal compared with the HI map, which showed, in the worst case (Patient 1, Expert 2, GM), a 141% brightness gap. In contrast, this same case presented 18% and 32% IPD for FLAIR and intermediate images.

At the same time, the standard deviation in the HI map was far higher than in FLAIR and intermediate images, which is an indication that the map had a rather spread out MS lesion intensity profile. While this is a concern that must be addressed when using the map to segment lesions, whether manually or automatically, the overall difference in intensity between lesions and other brain tissues was still significant and potentially provides enough leverage to overcome, at least partially, the wide standard deviation variation.

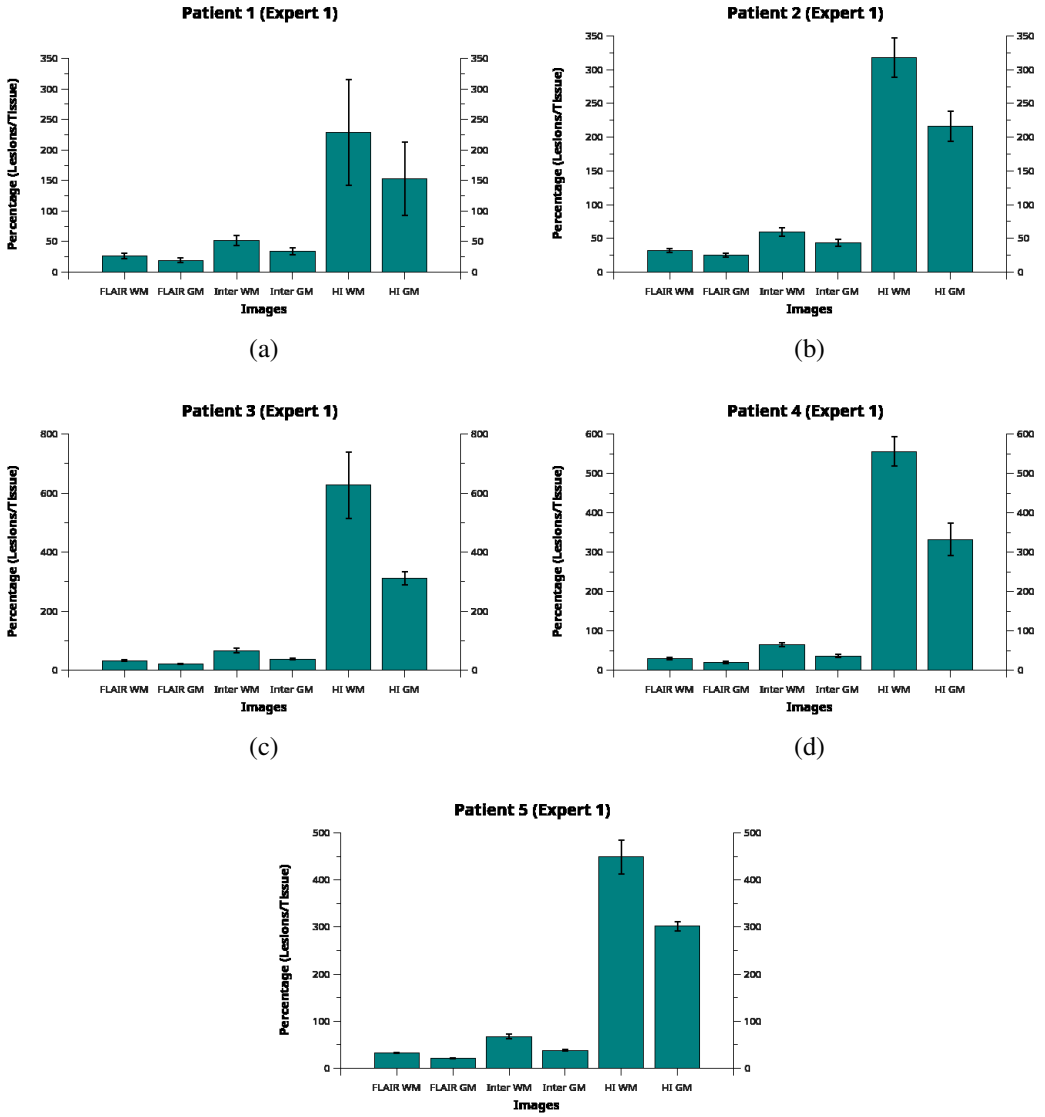


Figure 6.11: Lesion intensity profile compared with white matter and gray matter tissues using ground truths from expert 1. Here, "Inter" is the intermediate image and "HI" is the hyperintensity map. Each bar represents the mean lesion intensity over the mean intensity of a given tissue (white matter or gray matter) in a particular image type.

A drawback of the HI map is that it is highly dependent on the gray level intensity in FLAIR. This fact posed two problems that can be observed in Figure 6.13. The first one is presented in Figure 6.13(a)-(c) and concerns the natural intensity variation within the lesion profile. In this case, the lesion enclosed by the rectangle was not as enhanced as the one enclosed by the ellipse. As observed in Figure6.13(a), the lesion enclosed by the rectangle did not present a profile as hyperintense as the ellipse one in FLAIR, so this difference was propagated to the HI map. The other problem is shown in Figure 6.13(d)-(f). Areas enclosed by rectangles indicate regions that have no MS lesions and yet are enhanced in the map. Again, this occurred because

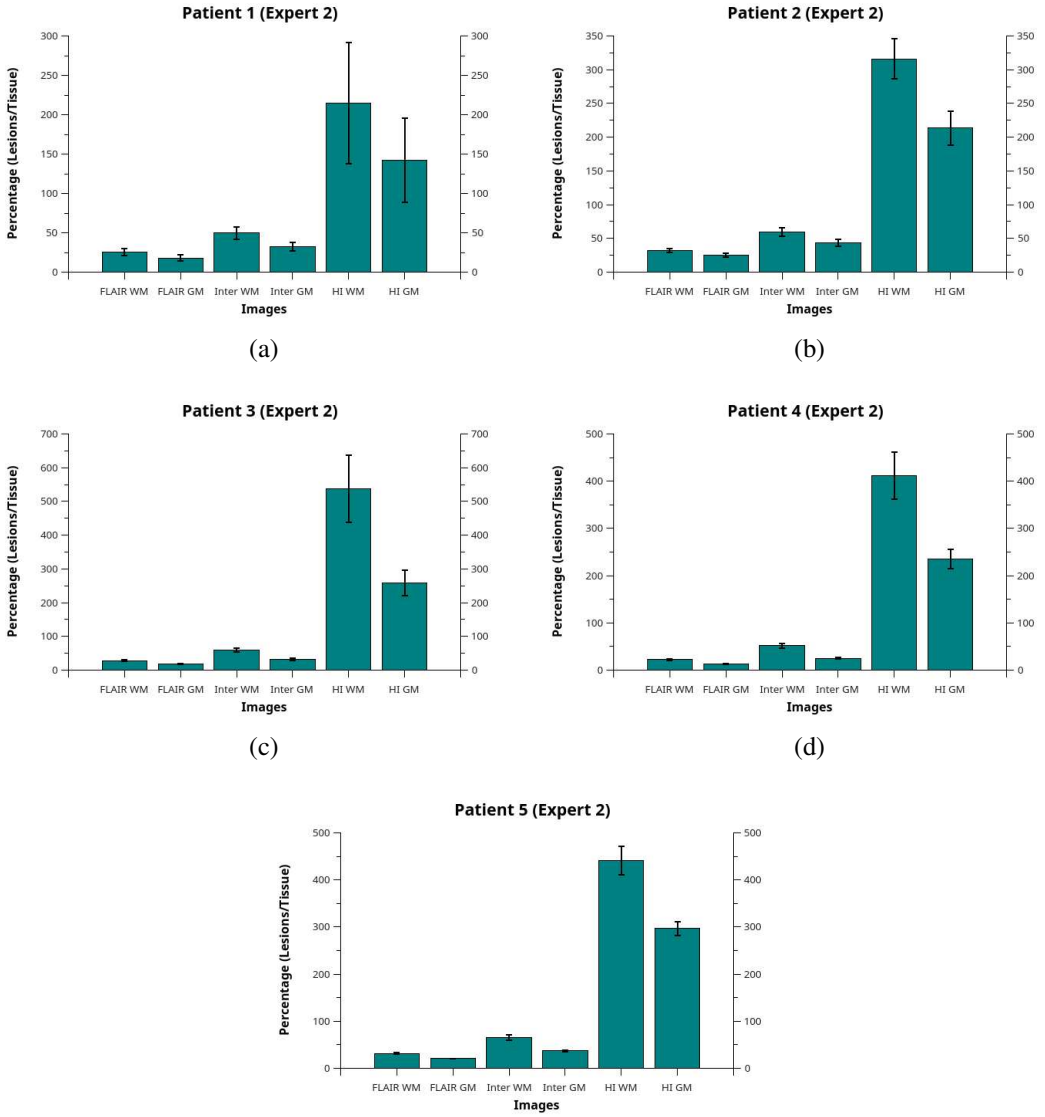


Figure 6.12: Lesion intensity profile compared with white matter and gray matter tissues using ground truths from expert 2. Here, "Inter" is the intermediate image and "HI" is the hyperintensity map. Each bar represents the mean lesion intensity over the mean intensity of a given tissue (white matter or gray matter) in a particular image type.

these regions presented rather high-intensity profiles in FLAIR and thus were enhanced in the map. Both these problems can interfere with lesion segmentation accuracy and indicate that the HI map should not be used as a stand-alone feature in manual and automatic segmentation techniques.

While there are some works in the literature (SHAH et al., 2011; TOMAS-FERNANDEZ; WARFIELD, 2015; BATTAGLINI; JENKINSON; STEFANO, 2012) that mention MS lesion intensity profiles and how they relate to other tissues, none of them provide a quantitative analysis regarding percentage, making it difficult to compare our results with theirs directly. The databases are also differ-

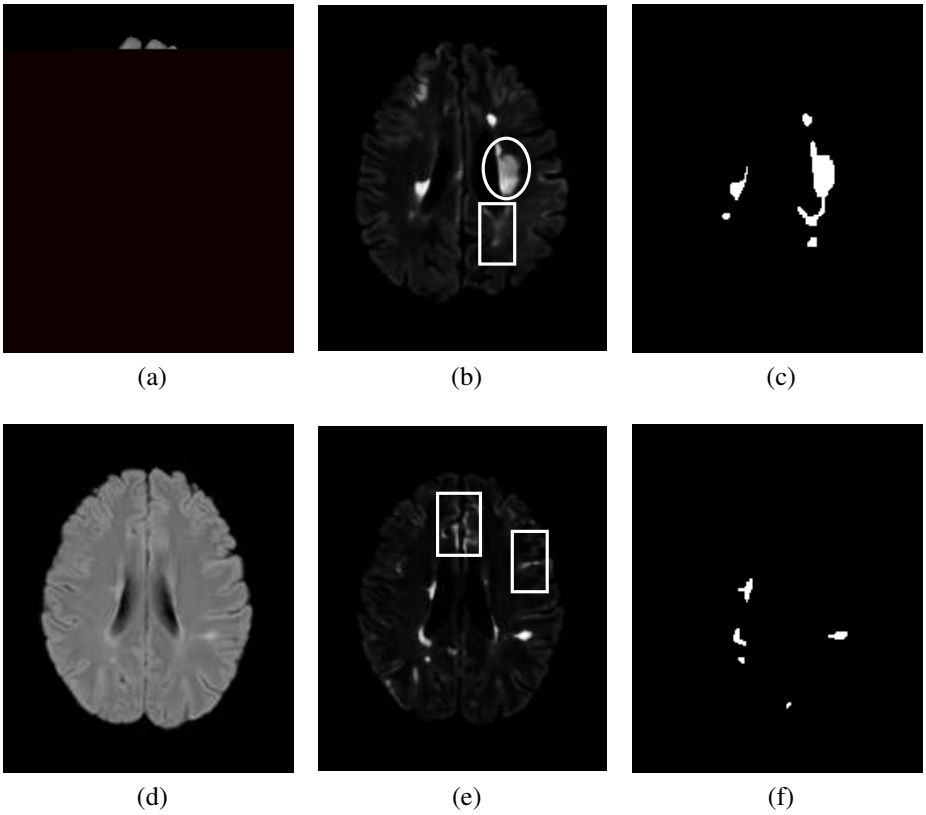


Figure 6.13: Intensity problems caused by wide lesion intensity range (first row) and hyperintensities in regions other than lesions (second row).

ent. However, by analyzing scatter plots in (SHAH et al., 2011; TOMAS-FERNANDEZ; WARFIELD, 2015), it is possible to observe that lesion intensity profiles present a significant overlap with other brain tissues. Hence, the HI map can undoubtedly help distinguish lesions more easily.

6.2.2 White matter mask comparison

We compared the white matter mask estimated in Section 5.5.2 with our WM ground truths using the DSC. We also extracted the percentage of lesions (intersection) present in each estimated mask to analyze the lesion load that was kept during the estimation. Again, since there were two lesion ground truths for each time-point, we extracted metrics for both experts. The results are shown in Table 6.6 and in Figures 6.14 and 6.15.

We also calculated the sensitivity and specificity of the estimated masks. The results for both experts are shown in Table 6.7.

The results shown in Tables 6.6 and 6.7 and in Figures 6.14 and 6.15 indicated high DSCs and significant intersection with lesions. It also pointed to very high sensitivity and specificity metrics. A relevant observation to be made is that the LI metric presented a consistent level

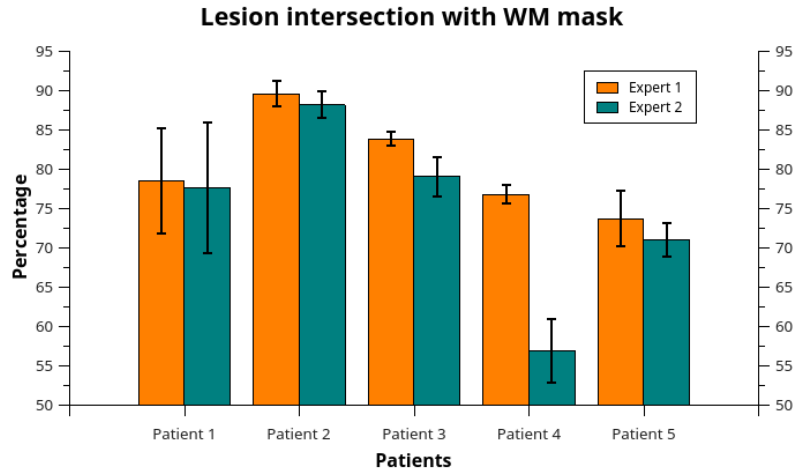


Figure 6.14: Lesion intersection with the estimated white matter mask using ground truths from both experts.

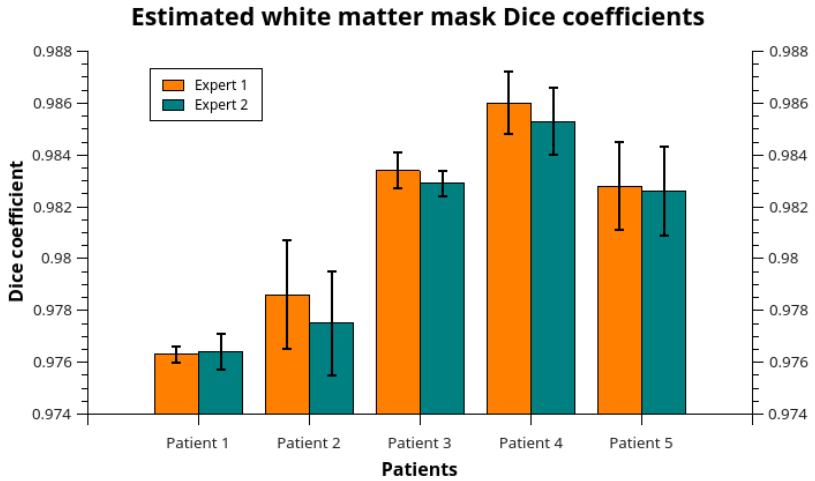


Figure 6.15: Dice coefficients of the estimated white matter mask compared to ground truths created using experts lesion annotations and automatic brain segmentation.

	LI (%) Expert 1 ($\mu \pm \sigma$)	LI (%) Expert 2 ($\mu \pm \sigma$)	Dice Expert 1 ($\mu \pm \sigma$)	Dice Expert 2 ($\mu \pm \sigma$)
Patient 1	78.56 ± 6.70	77.65 ± 8.30	0.9763 ± 0.0003	0.9764 ± 0.0007
Patient 2	89.60 ± 1.59	88.20 ± 1.71	0.9786 ± 0.0021	0.9775 ± 0.0020
Patient 3	83.88 ± 0.94	79.08 ± 2.51	0.9834 ± 0.0007	0.9829 ± 0.0005
Patient 4	76.83 ± 1.20	56.95 ± 4.03	0.9860 ± 0.0012	0.9853 ± 0.0013
Patient 5	73.73 ± 3.55	71.00 ± 2.16	0.9828 ± 0.0017	0.9826 ± 0.0017

Table 6.6: Lesion intersection (LI) and Dice coefficients for the white matter mask estimation for both expert ground truths.

of intersection regardless of lesion volumes, which is an indication of the robustness of our technique.

In Figure 6.14, it is possible to observe that patient 4 presented very different results on lesion intersection. The reason for this is that the expert annotations for this patient had the lowest DSC (0.612 ± 0.0019) among all patients, as mentioned in (FREIRE; FERRARI, 2016). In other words, experts did not have a high agreement coefficient on lesion segmentation for this particular case, which consequently made our technique present very different intersection values for each annotation.

Another point to be made about Figure 6.14 is that patient 1 presented the highest standard deviation of all, which comes from the fact that this patient's lesion intensities faded across time-points, making the enhancement less useful. This fading phenomenon can also be observed in Figures 6.11 and 6.12, since patient 1 had the highest standard deviation on the lesion intensity profile in the HI map compared with other patients.

6.3 Classification results

In this section we present the voxel-level classification results based on specificity, sensitivity and F1-score, as well as the accuracy of our method on a lesion-level basis. Before discussing the results per se, two observations regarding the number of folds (6 in total, as

	SS Expert 1 ($\mu \pm \sigma$)	SS Expert 2 ($\mu \pm \sigma$)	SP Expert 1 ($\mu \pm \sigma$)	SP Expert 2 ($\mu \pm \sigma$)
Patient 1	0.9931 ± 0.0024	0.9920 ± 0.0028	0.9786 ± 0.0012	0.9793 ± 0.0008
Patient 2	0.9928 ± 0.0012	0.9915 ± 0.0014	0.9823 ± 0.0023	0.9818 ± 0.0016
Patient 3	0.9983 ± 0.0001	0.9968 ± 0.0009	0.9816 ± 0.0010	0.9818 ± 0.0011
Patient 4	0.9990 ± 0.0008	0.9975 ± 0.0002	0.9818 ± 0.0015	0.9818 ± 0.0016
Patient 5	0.9974 ± 0.0002	0.9974 ± 0.0002	0.9822 ± 0.0020	0.9820 ± 0.0020
Overall	0.9962± 0.0028	0.9813 ± 0.0021	0.9951 ± 0.0030	0.9814 ± 0.0017

Table 6.7: Sensitivity (SS) and specificity (SP) values for the white matter mask estimation for both expert ground truths.

shown in Table 6.8) are deemed necessary.

First, we set the number of patients per fold to five as a trade-off between the size of the training and test sets in order to make the classifier general enough to avoid overfitting problems. However, since the number of patients with active lesions was 33, the last fold would comprise either 3 or 8 patients. We chose the latter for the sake of generalization. The results using only 3 patients in the training set indicated a strong overfitting, especially for the active lesion class. Hence, we decided to include these 3 patients in the last 5-patient fold. Since this experiment design choice implied in less samples for the training set, thus making the classification task "harder", we decided it was a valid approach. This is the reason why the sixth fold included eight instead of five patients from each class.

Second, as mentioned in Section 5.1, the nonactive database was comprised of 5 patients, four of them with four time points and one with five. The rationale to choose which patients with nonactive lesions would go into each fold was to randomly choose one time point per patient when the fold size was five. For the last fold, with eight patients, we simply chose three additional random time points from the first three patients in the nonactive dataset (i.e., the last fold had two time points from the first three patients and one time point from the rest).

The voxel-level classification results are shown in Table 6.8. The nonactive classification

yielded higher results when compared to the active class, especially regarding the specificity metric. However, as mentioned in Section 3.5.3, sensitivity and F1-score are more significant when it comes to analyze imbalanced datasets as the one we used in this work. In this sense, we can observe that these two metrics were similar between the active and nonactive classes. Moreover, we verified that most misclassifications occurred when a nonactive voxel was mistaken for an active one. Even though this is not ideal, this result is still better for a physician - who can discard these misclassifications when reviewing the outcome of our approach - than the other way around.

Patients	Class	Results		
		Specificity	Sensitivity	F1-score
1-5	Active	0.945	1	0.972
	Nonactive	1	0.993	0.997
6-10	Active	0.7	0.996	0.823
	Nonactive	1	0.966	0.983
11-15	Active	0.765	1	0.867
	Nonactive	1	0.977	0.988
16-20	Active	0.99	1	0.995
	Nonactive	1	0.998	0.999
21-25	Active	0.887	0.863	0.875
	Nonactive	0.978	0.982	0.98
26-33	Active	0.999	0.958	0.978
	Nonactive	0.991	1	0.995
Weighted avg.	Active	0.892	0.968	0.924
	Nonactive	0.994	0.987	0.991

Table 6.8: Specificity, sensitivity and F1-score metrics for the 6-fold cross-validation.

A lesion-level overview is shown in Table 6.9. The results show that 154 out of 157 lesions were correctly classified as active and the classification was perfect for the nonactive class. As mentioned in Section 3.5.3, we averaged the probability of each lesion as a whole and labeled them according to the majority class. We can see that even though the active class did not have specificity, sensitivity and F1-score as high as the nonactive class in a voxel-level analysis, the mistakes were compensated by the surrounding voxels comprising the lesion itself. In other words, the voxel-level classification provided a margin of safety for errors, making

them less significant when viewed from a coarser granularity. Since the outcome of interest for neurologists and physicians is on the lesion-level, this margin indicates there is room for mistakes on a voxel-level that, to a certain extent, do not interfere on the lesion-level results.

Table 6.9: Mean (std) for lesion-level classification.

Fold	Active ID	Active prob. (mean±std)	Active hits	Nonactive ID	Nonactive prob. (mean±std)	Nonactive hits
1	1	1 (1.04e-13)	1/1	1	0.982 (0.009)	26/26
	2	1 (1.07e-11)	5/5	6	0.998 (0.001)	31/31
	3	1 (3.36e-08)	12/12	9	0.983 (0.009)	26/26
	4	1 (2.07e-09)	3/3	14	0.993 (0.002)	20/20
	5	1 (2.04e-12)	1/1	18	0.987 (0.007)	19/19
2	6	1	1/1	2	0.930 (0.049)	33/33
	7	1 (1.21e-10)	4/4	5	0.989 (0.005)	41/41
	8	0.989 (0.00681)	6/6	11	0.912 (0.061)	28/28
	9	1 (2.48e-15)	3/3	16	0.966 (0.019)	21/21
	10	1 (2.7e-14)	2/2	20	0.961 (0.024)	22/22
3	11	1 (3.29e-12)	2/2	3	0.944 (0.034)	20/20
	12	1 (5.03e-12)	1/1	8	0.992 (0.001)	33/33

Table 6.9: Mean (std) for lesion-level classification.

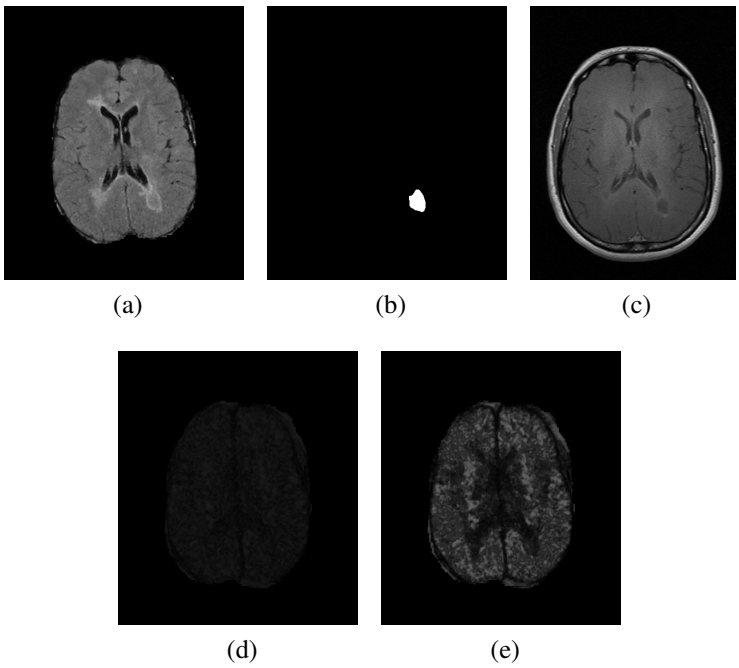
Fold	Active ID	Active prob. (mean±std)	Active hits	Nonactive ID	Nonactive prob. (mean±std)	Nonactive hits
4	13	1 (2.66e-15)	6/6	12	0.947 (0.034)	26/26
	14	1 (5.81e-12)	3/3	17	0.985 (0.006)	17/17
	15	1 (1.3e-12)	1/1	21	0.963 (0.022)	20/20
	16	1 (3.08e-08)	6/6	1	0.994 (0.002)	26/26
	17	1 (7.63e-07)	2/2	6	1 (6.89e-05)	31/31
	18	1 (8.26e-06)	13/13	11	0.994 (0.003)	28/28
	19	1 (1.06e-06)	2/2	16	0.997 (0.001)	21/21
	20	0.986 (0.00418)	4/4	18	0.995 (0.002)	19/19
5	21	0.956 (0.039)	4/4	1	0.95 (0.027)	26/26
	22	0.687 (0.213)	3/3	5	0.994 (0.001)	41/41
	23	0.974 (0.0245)	4/4	9	0.965 (0.018)	26/26
	24	0.585 (0.241)	1/1	15	0.988 (0.004)	25/25
	25	0.883 (0.0451)	10/10	19	0.958 (0.022)	23/23

Table 6.9: Mean (std) for lesion-level classification.

Fold	Active ID	Active prob. (mean±std)	Active hits	Nonactive ID	Nonactive prob. (mean±std)	Nonactive hits
6	26	0.996 (0.001)	9/9	1	0.999 (0.004)	26/26
	27	1 (2.08e-05)	26/26	7	1 (1.84e-05)	33/33
	28	0.988 (0.003)	13/13	10	0.997 (0.001)	28/28
	29	1 (2.15e-09)	1/1	14	0.999 (0.005)	20/20
	30	0.995 (0.001)	2/2	18	0.998 (0.001)	19/19
	31	1 (2.92e-07)	2/2	2	1 (7e-05)	33/33
	32	0.259 (0.076)	0/3	6	0.999 (0.001)	31/31
	33	1 (3.59e-09)	1/1	11	0.995 (0.001)	28/28

Only one patient, ID 32, did not have their lesions identified as active. Some features from this particular patient are shown in Figure 6.16. Visual analysis indicate that not only were the lesions small, they also had an intensity profile too similar to that of nonactive lesions. Moreover, the quality of the FLAIR image itself was rather poor. Even though textural and enhanced features aid in the distinction between both classes, there are cases where the classification step fails to yield the expected output due to conditions that are out of our reach but yet have influence on the final outcome.

We can also observe that the standard deviation across all lesions was quite small and the average lesion probability for most patients, both active and nonactive, was close or equal to



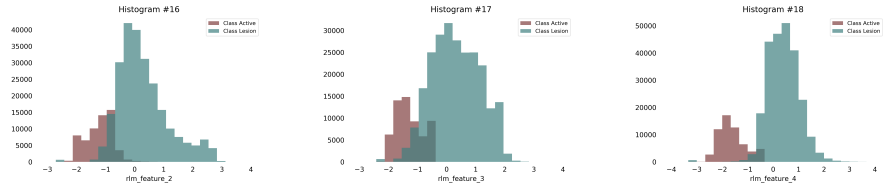
(a) (b) (c)

(d) (e)

Figure 6.16: Example of slices from patient ID 32 who had zero lesion hits. (a) FLAIR, (b) active lesion annotation, (c) T1-w with Gadolinium enhancement, (d) RLM LRE, (e) RLM GLN.

one. It indicates our classifier was able to accurately distinguish both classes with a high degree of certainty, which is important when dealing with a sensitive health care scenario as this one.

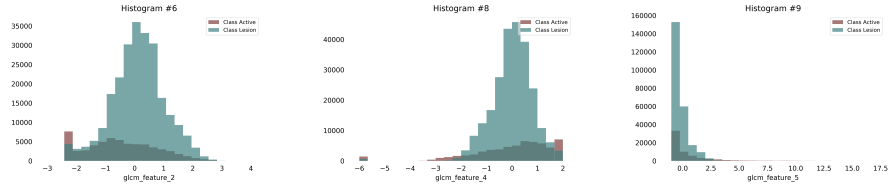
To analyze the effect of the features in the classification step, we plotted histograms for all 22 attributes extracted from all images of both active and nonactive datasets. Doing so we observed the RLM features RLE, RLN, GLN showed a significant distinction between both classes, whereas GLCM features, in general, had an overlap between one class and the other. This is an indication that active and nonactive lesions are better separated when accounting for their smoothness profile. Histograms from RLM and GLCM features are shown in Figures 6.17 and 6.18.



(a) (b) (c)

Figure 6.17: RLM feature histograms: (a) RLE, (b) RLN and (c) GLN. Class "Lesion" represents the nonactive lesions.

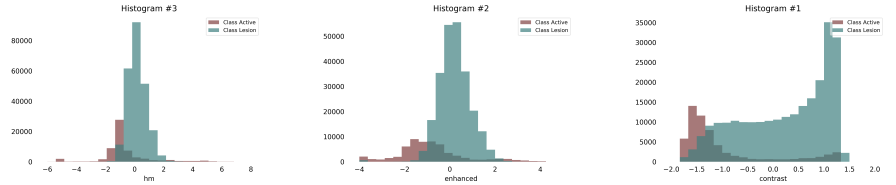
The enhancement effect mentioned in Section 5.5 can be seen in histograms shown in Figure 6.19. We can see that the enhanced and hyperintensity map features offset and spread the class



(a) (b) (c)

Figure 6.18: GLCM feature histograms: (a) Entropy, (b) Inverse Difference Moment and (c) Inertia. Class "Lesion" represents the nonactive lesions.

histograms apart, making them more distinguishable. Extracting good features is of paramount importance for the supervised learning pipeline, since they have a direct effect on the classification step. Therefore, combining dissimilar and relevant textural and enhanced features creates a domain space that is adequate for a proper class separation. However, it is also relevant to note that features with significant overlays cannot be promptly discarded because their combination with other more key/distinct features can be useful for the classification step as a whole, especially when the overlay areas between these features are complementary (i.e., the overlay area between classes in one histogram is the most distinctive area in another).



(a) (b) (c)

Figure 6.19: Enhanced feature histograms: (a) Histogram matched (no enhancement), (b) Enhanced and (c) Hyperintensity Map. Class "Lesion" represents the nonactive lesions.

Overall, the classification accuracy on a lesion-level for the active class was over 98%, indicating that it is possible to identify this particular kind of MS lesion using only FLAIR images and without any kind of contrast agent. In addition, RLM features proved to be the most distinctive ones in this scenario. Performing the classification step on a voxel-level provided us with enough samples to properly train our classifier and also allowed for a margin of safety to take place, "smoothing out" the effects of misclassifications on a lesion-level .

6.4 Final considerations

This chapter presented the MS lesion segmentation results using an iterative approach combined with the Student's t-mixture model applied to FLAIR images. The algorithm made use of

probabilistic anatomical atlases to constrain the segmentation region to WM to reduce the number of false positives introduced by voxels of intensities similar to MS lesions but that belong to different clusters. The iterations proved to be a successful idea because the refinement in the segmentation was able to gather most lesions as the most hyperintense cluster.

We compared the segmentation results with manual annotations and with other techniques applied to the same database. To quantitatively analyze the results, we extracted the DSC, TPR, FPR, VD and Pearson's correlation coefficient. These metrics provided an overview of the spatial and volumetric agreement between our algorithm and the expert annotations and indicated our approach presented results that compared favorably to the state-of-the-art and yielded consistent outputs regardless of lesion load.

We also presented in this chapter an intensity enhancement technique that highlights hyperintensities in FLAIR images while dimming down non-hyperintense regions. The quality of the data provided to the model is just as important as the model itself, so we believe the hyperintensity map can provide valuable information about the location of lesions and thus improve the quality of the segmentation. Our results show the hyperintensity enhancement significantly improved the distinction of the lesion intensity profile compared with other brain tissues (WM and GM). However, there were some caveats in the technique that need to be addressed before actually incorporating it into the model. The main drawback currently lies in the fact that some hyperintense regions that do not correspond to MS lesions are highlighted in a way they end up being similar to lesions themselves. To circumvent this problem, spatial information provided by the atlases combined with both global and local intensity information of these regions can be investigated to eliminate the issue.

Moreover, we showed the hyperintensity map can provide a white matter mask estimation as its byproduct. As mentioned before, constraining the segmentation region to WM can help reduce the number of false positives, since voxels located mainly in the GM tissue have a significant intensity overlap with MS lesions. In this sense, our WM estimation had an excellent spatial agreement with the actual WM region and can undoubtedly be used as an alternative to the constraint step we used in the segmentation pipeline.

Finally, we presented a supervised classification pipeline to distinguish active and nonactive MS lesions in FLAIR images without using any contrast agent. We relied on textural and enhanced features to create a feature space where both classes could be easily distinguished using the XGBoost classifier. Our dataset comprised 54 patients, 33 with active lesions and 21 with nonactive ones, with varying lesion loads, shapes, sizes and locations.

Though active lesions represented a small portion of samples when compared to nonactive

ones, the results indicate that is possible to tell these two classes apart using information from FLAIR images and without the aid of contrasts of any kind. Out of the 157 active lesions, 154 were correctly identified as so. This is relevant because it unburdens patients from the accumulation of heavy metal-based contrasts in their systems, makes it possible for patients with kidney problems to get an assessment of their active lesion progress (because they could not take the contrast in the first place) and also lowers the cost of the MRI procedure, since the contrast itself is no longer be necessary.

Chapter 7

CONCLUSIONS

In this chapter we present the conclusions of this work and further investigations that can be derived from our results.

7.1 Overview and future investigations

The main goal of this research was to develop an automatic computational scheme for segmentation, volumetric assessment and analysis of multiple sclerosis lesions in magnetic resonance images of the human brain. We divided this goal into two main steps: tissue clustering and active lesion classification. Besides being challenging, these tasks pose difficulties that range from the image acquisition process to the intrinsic characteristics of MS lesions — which have no well-defined shape, size or location. A myriad of works in the literature regarding segmentation and classification are divided into two main groups: supervised and unsupervised approaches. To this day there is no consensus as to which technique or approach works best for all scenarios.

We opted for using an unsupervised approach in the segmentation of brain tissues and lesions because of the intrinsic complexity of the human brain. The available public datasets are usually not large enough to encompass all the physiological and pathological variations that subjects may present. Given the unsupervised approach, the decision to use a parametric technique based on the Student's t-mixture model was made based on promising results presented by some previous works, mentioned in Chapter 4, that used this technique in the segmentation context. Our approach, combined with iterative refinement of the segmentation, yielded very good results in the 2015 Longitudinal MS Lesion Segmentation Challenge and compared favorably both to manual annotations and other automatic techniques applied to the same database.

Moreover, we developed a technique to increase the distinction of MS lesions in MRI. This process had some caveats, the most glaring being the highlighting of partial volume areas that did not encompass any MS lesions (i.e., false positives). That was the reason why we did not use this input in the segmentation step. However, taking only lesions themselves into account, the hyperintensity map we created provided distinguishable information that was used for our classification purposes.

Regarding the distinction between active and nonactive lesions, our results were auspicious and indicated that textures were a powerful tool for telling these two classes apart. We chose an ensemble approach for this particular step because the fine and subtle differences between the two types of lesions would be better leveraged in a supervised domain. Moreover, the XGBoost classifier proved to be effective in reducing variance and bias, and even though the voxel-level results for the active lesion class were not as good as the nonactive one, a lesion-level analysis showed that misclassifications in the former were not severe, thus indicating a margin of safety regarding errors in a voxel-based classification in this scenario.

The benefits of this finding are many. First, patients will no longer have to go through contrast injections, which is commonly described as a rather uncomfortable procedure, and would not have to worry about its accumulation in their system. Second, MS patients with kidney problems will be able to have a more accurate follow-up regarding the progression of the inflammatory state of their lesions, since they could not undergo contrast injections in the first place. And third, the cost of MRI procedures for MS patients can get potentially lower due to the elimination of contrast injections altogether.

There were limitations in the classification step. A larger number of patients would be required to assess the classification performance on a broader set of lesions. Also, lesions were not rated by an independent neuroradiologist, and that could have introduced bias in our study. And finally, we lacked an external validation set, which would make our data more robust. Nonetheless, we were able to show that is very possible to correctly identify active MS lesions without using Gadolinium-based contrasts.

As future investigations that can be derived from this work, feature selection and analysis could be conducted in order to verify which inputs are the most relevant for the classification step and which can be discarded without affecting accuracy. Furthermore, a longitudinal study exploring MS lesions and their surroundings could provide important information about disease progression and indicate potential markers for new lesions and, consequentially, earlier diagnosis and treatment.

Appendix A

EXPECTATION MAXIMIZATION

A.1 The Expectation Maximization algorithm

The Expectation Maximization (EM) algorithm is an iterative approach widely used to compute maximum likelihood estimates (MLE), being particularly useful in problems with incomplete data (MCLACHLAN; KRISHNAN, 1997). The basis of EM is to associate an incomplete data problem to a complete one, where the maximum likelihood estimate is computationally feasible. In this scenario, the MLE of the complete data problem can have a closed solution or at least be calculated by well-established numeric methods in the literature.

Each iteration of EM is comprised of two stages. In the first one, commonly called E-step, we calculate the conditional expectation of the complete data (observable data along with the current parameters of the statistical model). In the second stage, known as M-step, we maximize the likelihood of the complete data to get an updated estimate of the model parameters. This process is repeated until the algorithm converges or a number of iterations is reached.

More formally, let $\mathbf{x}_i$ be the observed data (in this work, they are represented by gray level intensities in different image weights) for case i and let z_i be the missing data (corresponding to the labels of each brain tissue). Let $X_o = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and $X_{m=}\{z_1, \cdots, z_N\}$ be the sets of observed and missing data, respectively. The goal is to maximize the log-likelihood function of the observed data (MURPHY, 2012) given by

$$\ell(\theta) = \sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \theta) = \sum_{i=1}^{N} \log \left[\sum_{N} p(\mathbf{x}_i, z_i \mid \theta) \right], \tag{A.1}$$

where θ represents the model parameters.

However, Equation A.1 is hard to maximize since we do not know the missing variables z_i. The EM algorithm solves this problem iteratively by maximizing the expectation of the complete data as if we knew the missing data. To do so, we take an initial guess of the unknown values.

Let the likelihood of the complete problem be defined by

$$\ell_c(\theta) \triangleq \sum_{i=1}^{N} \log p(\mathbf{x}_i, z_i \mid \theta). \tag{A.2}$$

However, since Equation A.2 cannot be computed because z_i is unknown, we define the expectation of the log-likelihood function as

$$Q(\theta;\theta^t) = E\left[\ell_c(\theta) \mid X_p, \theta^t\right], \tag{A.3}$$

where t is the current iteration and Q is an auxiliary function. The computation of the expectation is done related to the old parameters, θ^t and the observed data X_p. The goal of the E-step is to compute $Q(\theta;\theta^t)$; more specifically, to compute the terms the MLE depends on.

In the M-step, function Q is optimized with respect to θ:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta;\theta^t). \tag{A.4}$$

To perform the maximum a posteriori (MAP) estimation, the M-step is modified to

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta;\theta^t) + \log p(\theta). \tag{A.5}$$

A.2 Expectation Maximization and the Student's t distribution

As explained in Section 3.1, the Student's t-distribution is a more robust alternative regarding outliers when compared to the Gaussian distribution. However, different from the Gaussian, the Student's t distribution does not have a closed form for the MLE. In this case, it is necessary to use iterative optimization methods, being the EM algorithm one of the most used approaches for that matter (MURPHY, 2012).

As shown in (PEEL; MCLACHLAN, 2000), a multivariate Student's t-distribution $t(\boldsymbol{x}_i \mid \mu, \Sigma, \nu)$ can be written as a weighted mixture of Gaussian distributions

$$t(\mathbf{x}_i \mid \mu, \Sigma, \nu) = \int N\left(\mathbf{x}_i \mid \mu, \frac{\Sigma}{u}\right) Ga\left(u \mid \frac{\nu}{2}, \frac{\nu}{2}\right) du, \tag{A.6}$$

where N is the Gaussian distribution given by

$$N(\mathbf{x}_i \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \right\}, \tag{A.7}$$

and u is a weighting factor that follows a Gamma distribution parameterized by

$$u \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \tag{A.8}$$

Equation A.6 can be seen as an infinite mixture of Gaussians, each with slightly different covariance matrices (MURPHY, 2012). Being z_i the missing data, the likelihood of the complete problem can be written as

$$\begin{aligned} \ell_c(\theta) &= \sum_{i=1}^{N} \left[\log N\left(\mathbf{x}_i \mid \mu, \frac{\Sigma}{z_i}\right) + \log Ga\left(z_i \mid \frac{\nu}{2}, \frac{\nu}{2}\right) \right] \\ &= \sum_{i=1}^{N} \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{z_i}{2}\delta_i + \frac{\nu}{2}\log\frac{\nu}{2} - \log\Gamma\left(\frac{\nu}{2}\right) \right. \\ &\quad \left. + \frac{\nu}{2}(\log z_i - z_i) + \left(\frac{\nu}{2} - 1\right)\log z_i \right], \end{aligned} \tag{A.9}$$

where the Mahalanobis distance is defined by

$$\delta_i = (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu). \tag{A.10}$$

It is possible to partition $\ell_c(\theta)$ into two terms, one with μ and Σ and the other with ν. This way, we get (MURPHY, 2012)

$$\ell_c(\theta) = L_N(\mu, \Sigma) + L_G(\nu), \tag{A.11}$$

$$L_N(\mu, \Sigma) \triangleq -\frac{1}{2} N \log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N} z_i \delta_i, \tag{A.12}$$

$$L_G(\nu) \triangleq -N\log\Gamma\left(\frac{\nu}{2}\right) + \frac{1}{2}N\nu\log\left(\frac{\nu}{2}\right) + \frac{1}{2}\nu\sum_{i=1}^{N}(\log z_i - z_i). \tag{A.13}$$

A.2.1 EM with known ν

For simplicity, first we will derive the EM algorithm assuming the value of the degrees of freedom, ν, is known. In this case, we can ignore L_G from Equation A.13. According to (ZHANG; WU; NGUYEN, 2013) and (MURPHY, 2012), we have that

$$p(z_i \mid \mathbf{x}_i, \theta) = Ga\left(z_i \mid \frac{\nu + D}{2}, \frac{\nu + \delta_i}{2}\right), \tag{A.14}$$

that is, the missing data z_i follow a Gamma distribution. If $z_i \sim Ga(a,b)$, then $E[z_i] = a/b$. Therefore, the E-step at iteration t is

$$\bar{z}_i^{(t)} \triangleq E\left[z_i \mid \mathbf{x}_i, \theta^{(t)}\right] = \frac{\nu^{(t)} + D}{\nu^{(t)} + \delta_i^{(t)}}. \tag{A.15}$$

This way, the M-step is obtained by maximizing $E[L_N(\mu, \Sigma)]$:

$$\hat{\mu}^{(t+1)} = \frac{\sum_i \bar{z}_i^{(t)} \mathbf{x}_i}{\sum_i \bar{z}_i^{(t)}}, \tag{A.16}$$

$$\begin{aligned} \hat{\Sigma}^{(t+1)} &= \frac{1}{N} \sum_i \bar{z}_i^{(t)} \left(\mathbf{x}_i - \hat{\mu}^{(t+1)}\right) \left(\mathbf{x}_i - \hat{\mu}^{(t+1)}\right)^T \\ &= \frac{1}{N} \left[\sum_i \bar{z}_i^{(t)} \mathbf{x}_i \mathbf{x}_i^T - \left(\sum_{i=1}^{N} \bar{z}_i^{(t)}\right) \hat{\mu}^{(t+1)} \left(\hat{\mu}^{(t+1)}\right)^T\right]. \end{aligned} \tag{A.17}$$

It is possible to note that $\bar{z}_i$ works as a precision metric of i; that is, if the value is small, the point corresponding to i will have a small weight in the computation of the mean μ and covariance Σ. This is one of the reasons why the Student's t distribution can be robust to outliers (MURPHY, 2012).

A.2.2 EM with unknown ν

To compute the MLE of the degrees of freedom, it is necessary to calculate $L_G(\nu)$, which concerns z_i and $\log z_i$. As mentioned before, $z_i \sim Ga(a,b)$; so it is possible to show that (MURPHY, 2012)

$$\bar{\ell}_i^{(t)} \triangleq E\left[\log z_i \mid \theta^{(t)}\right] = \psi(a) - \log b, \tag{A.18}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. From Equation A.14 we have

$$\begin{aligned} \bar{\ell}_i^{(t)} &= \psi\left(\frac{\nu^{(t)} + D}{2}\right) - \log\left(\frac{\nu^{(t)} + \delta_i^{(t)}}{2}\right) \\ &= \log\left(\bar{z}_i^{(t)}\right) + \psi\left(\frac{\nu^{(t)} + D}{2}\right) - \log\left(\frac{\nu^{(t)} + D}{2}\right). \end{aligned} \tag{A.19}$$

Substituting in Equation A.13, we have

$$E[L_G(\nu)] = -N \log\Gamma\left(\frac{\nu}{2}\right) + \frac{N\nu}{2} \log\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \sum_i \left(\bar{\ell}_i^{(t)} - \bar{z}_i^{(t)}\right), \tag{A.20}$$

and the gradient is given by

$$\frac{d}{d\nu}E\left[L_G(\nu)\right] = \frac{-N}{2}\psi\left(\frac{\nu}{2}\right) + \frac{N}{2}\log\left(\frac{\nu}{2}\right) + \frac{N}{2} + \frac{1}{2}\sum_i\left(\bar{\ell}_i^{(t)} - \bar{z}_i^{(t)}\right). \tag{A.21}$$

Equation A.21 has a unique solution in the interval $[0,\infty)$ that can be found iteratively using optimization techniques.

Appendix B

PUBLISHED PAPERS

Computers in Biology and Medicine 73 (2016) 10–23

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Automatic iterative segmentation of multiple sclerosis lesions using Student's *t* mixture models and probabilistic anatomical atlases in FLAIR images

Paulo G.L. Freire *, Ricardo J. Ferrari

Departamento de Computação, Universidade Federal de São Carlos, Rod. Washington Luís, Km 235, 13565-905 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 4 January 2016
Received in revised form
16 March 2016
Accepted 29 March 2016

Keywords:
Image segmentation
Multiple sclerosis
Mixture models
Student's *t* distribution
Probabilistic anatomical atlases

ABSTRACT

Multiple sclerosis (MS) is a demyelinating autoimmune disease that attacks the central nervous system (CNS) and affects more than 2 million people worldwide. The segmentation of MS lesions in magnetic resonance imaging (MRI) is a very important task to assess how a patient is responding to treatment and how the disease is progressing. Computational approaches have been proposed over the years to segment MS lesions and reduce the amount of time spent on manual delineation and inter- and intra-rater variability and bias. However, fully-automatic segmentation of MS lesions still remains an open problem. In this work, we propose an iterative approach using Student's *t* mixture models and probabilistic anatomical atlases to automatically segment MS lesions in Fluid Attenuated Inversion Recovery (FLAIR) images. Our technique resembles a refinement approach by iteratively segmenting brain tissues into smaller classes until MS lesions are grouped as the most hyperintense one. To validate our technique we used 21 clinical images from the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge dataset. Evaluation using Dice Similarity Coefficient (DSC), True Positive Ratio (TPR), False Positive Ratio (FPR), Volume Difference (VD) and Pearson's *r* coefficient shows that our technique has a good spatial and volumetric agreement with raters' manual delineations. Also, a comparison between our proposal and the state-of-the-art shows that our technique is comparable and, in some cases, better than some approaches, thus being a viable alternative for automatic MS lesion segmentation in MRI.

1. Introduction

Multiple sclerosis (MS) is a demyelinating autoimmune disease that attacks the central nervous system (CNS) and affects more than 2 million people worldwide [1]. It is mainly characterized by the presence of white matter (WM) lesions [2], which are visible on magnetic resonance imaging (MRI) scans and appear hyperintense on T2-weighted and Fluid Attenuated Inversion Recovery (FLAIR) images. Segmentation of MS lesions is usually done by a radiologist, who has to visually assess and manually delineate them when measuring total lesion volume. Since MRI brain scans are usually volumetric, the manual delineation procedure is done in a slice-by-slice manner, which is time consuming and suffers from large intra- and inter-rater variability and bias [3].

Clinical trials have shown that lesion volumes are meaningful outcomes for assessing disease burden in multiple sclerosis (MS) [4], and thus accurately measuring such volumes is of considerable interest in clinical practice [5]. In this scenario, an automated technique to segment MS lesions and measure their volumes would reduce the time needed from the rater and decrease the observer dependency as well. However, automatic segmentation of MS lesions is challenging, since a number of variables such as partial volume effect (PVE), bias field, acquisition parameters and different scanner magnetic field strengths may directly influence the segmentation outcome. There have been many proposals over the years regarding automatic MS lesion segmentation, but no single one appears to be widely used at the present time. The proposed techniques range from a myriad of approaches such as dictionary learning [6], logistic regression [7,8], patch-based [9], random decision forest [10] and mixture models [11,12].

In this paper, we propose a new fully-automatic technique for MS lesions' segmentation in MR FLAIR images using an iterative approach based on Student's *t* mixture models and probabilistic anatomical atlases. Our technique resembles a refinement approach by iteratively segmenting brain tissue classes into

* Corresponding author.
E-mail address: paulo.freire@dc.ufscar.br (P.G.L. Freire).
URL: http://www.bipgroup.dc.ufscar.br/ (R.J. Ferrari).

http://dx.doi.org/10.1016/j.compbiomed.2016.03.025
0010-4825/

subclasses until MS lesions are grouped as the most hyperintense one. Since our segmentation technique is intensity-based, we use probabilistic anatomical atlases to constrain the iterative process to the WM region, thus avoiding misclassification of voxels that have similar intensities to lesion voxels but are outside the WM tissue class.

To validate our technique we used 21 images from the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge (also known as 2015 ISBI Longitudinal Challenge) and assessed its performance using the Dice Similarity Coefficient (DSC), True Positive Ratio (TPR), False Positive Ratio (FPR), Volume Difference (VD) and Pearson's r coefficient on our results and rater manual delineations. It is important to note that each case in the dataset had 2 lesion delineations, each one done by a different rater, and we compared our results to both of them. Also, in some cases the manual delineations did not have a good agreement rate between themselves, which confirms the inter-variability and bias between observers.

We also compared our findings with proposals that took part in the 2015 ISBI Longitudinal Challenge. Since they used the same database as we did, a direct comparison was possible. We briefly describe the works that presented quantitative results in the following.

In [13], an automatic hierarchical framework for the segmentation of healthy tissues and lesions in brain MRI was proposed. The authors used a Markov Random Field segmentation framework that leveraged spatial prior probabilities for 9 healthy tissues through multi-atlas fusion and then used a random forest classifier to provide region level lesion refinement.

In [14], lesions were segmented using a fast patch matching approach, which was extended to multimodal data. To do so, the authors registered all available modalities to a common space and stacked them to form a 4D volume of multimodal intensities. Patches were defined and used to segment MS lesions.

An approach using random forest and local context intensity features was proposed in [15] to segment MS lesions. The authors extracted features from the images such as voxel intensity values (before and after image smoothing) and local histogram features and trained a random forest with supervised learning to segment MS lesions.

In [16], the authors proposed a 3D convolutional neural network (CNN) using a voxel-wise classifier with multi-channel 3D patches of MRI volumes as input. For each ground truth, a CNN was trained and the final segmentation was obtained by combining the probability outputs of these CNNs.

Finally, in [17] the authors used an estimation of spatially global within-the-subject intensity distribution and a spatially local intensity distribution derived from a healthy reference population to segment MS lesions. Using this approach, the authors aimed to distinguish locations in the brain with abnormal intensity levels when compared to the expected value at the same location in a healthy population.

All the works previously described are intensity-based methods and tried to use spatial information or patches to improve the segmentation results. The DSC, TPR and FPR values for these works are summarized in Table 10. These three metrics were the ones used by all authors, allowing a direct comparison of our approach to theirs. Along with DSC, TPR and FPR, we also used VD and Pearson's r coefficient metrics when comparing our proposal results to the raters' ground truths.

This paper is divided in the following manner. In Section 2 we explain the methodology of our work, including details about the dataset, metrics used to evaluate our model, preprocessing, our segmentation technique and post-processing stages. Results and discussions are presented in Section 3, where we compare results from our technique to each rater delineation and other proposals and also analyze how well the manual delineations for each image agree with each other. Finally, Section 4 concludes our paper.

2. Methodology

This section provides information about the dataset used in this work, along with the description of the metrics used to evaluate our results. It also presents information regarding our segmentation technique and pre- and post-processing steps.

2.1. Dataset

2.1.1. Clinical images

The dataset used to validate our technique consisted of longitudinal images from 5 patients obtained from the 2015 Longitudinal MS Lesion Segmentation Challenge[1] conducted at the 2015 International Symposium on Biomedical Imaging in New York, NY, April 16–19. Each longitudinal dataset included T1-, T2-, PD-weighted and FLAIR MR images with 4–5 time points acquired on a 3T MR scanner. Every longitudinal dataset had two manual lesion delineations provided by rater 1 and rater 2. Considering all longitudinal dataset cases, 21 time points were provided in total. T1-weighted images had approximately 1 mm^3 voxel resolution, while the other weighted images had a resolution of 1 mm^2 in plane with 3 mm thickness. To minimize the dependency of the results on registration performance and brain extraction, all images were already rigidly registered to the baseline T1-weighted image with automatically computed skull stripping masks. After registration, image dimensions were $181 \times 217 \times 181$ for every image.

2.1.2. Probabilistic anatomical atlases

Three probabilistic anatomical atlases, corresponding to gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), were used to provide spatial information to our algorithm. They were obtained from the ICBM[2] project [18]. Their spatial resolution was $1 \times 1 \times 1$ mm and their dimensions were $256 \times 256 \times 256$.

2.2. Metrics

To evaluate our technique, we used the Dice Similarity Coefficient (DSC) [19], True Positive Rate (TPR), False Positive Rate (FPR) Volume Difference (VD) and Pearson's r coefficient.

The DSC is defined as the ratio between the number of voxels where both the automatic and rater reference segmentation (ground truth) agree (true positives) and the sum of the total number of voxels labeled as lesion by both methods (manual and automatic). Also, according to Bartko [20], DSC values of 0.7 or higher suggest good agreement between two delineations. The TPR, FPR and VD metrics were calculated taking into account only the lesion voxels.

The Pearson's r coefficient [21] was used to assess the volumetric correlation between our segmentation and the ground truths from the raters. It is given by

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}, \tag{1}$$

where n is the number of time points, x_i and y_i are the absolute volumes of the ground truth and the automatic segmentation for a particular time point and $\bar{x}$ and $\bar{y}$ are their respective means. Pearson's r coefficient values lie inside the interval $[+1, -1]$.

[1] http://iacl.ece.jhu.edu/MSChallenge
[2] http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009

Table 1
Metrics used to evaluate the automatic segmentation. *TP* stands for True Positives, *FP* stands for False Positives, *FN* stands for False Negative, V_{seg} is the lesion automatic segmentation volume and V_{GT} is the lesion ground truth volume.

Metric	Abbreviation	Equation	Interval
Dice Similarity Coefficient	DSC	$\frac{2\times TP}{FP+FN+2\times TP}$	[0, 1]
True Positive Rate	TPR	$\frac{TP}{TP+FN}$	[0, 1]
False Positive Rate	FPR	$\frac{FP}{FP+TP}$	[0, 1]
Volume Difference	VD	$\left\|1-\frac{V_{seg}}{V_{GT}}\right\|$	[0, 1]

The DSC, TPR, FPR and VD metrics are summarized in Table 1. For DSC and TPR, the closer their values are to 1, the better. On the other hand, for FPR and VD, the closer their values are to 0, the better.

2.3. Preprocessing

A number of well known preprocessing techniques were applied to the dataset images before segmentation. These techniques included noise reduction, bias field correction and registration of probabilistic anatomical atlases to the clinical data (MRI scans), as shown in Fig. 1.

2.3.1. Noise reduction

Although the term noise in MRI can have different meanings depending on the context, the main source of noise in MR images is the thermal noise from the patients' body. Thermal noise may negatively affect the segmentation by increasing the number of misclassified voxels. To minimize noise effect, the Non-Local Means (NLM) [22] technique was applied to all images in the dataset. This filtering technique works by smoothing small grayscale variations in an adaptive manner, reducing the amount of blur on edge regions and increasing it otherwise. The NLM also takes into account the fact that some features in an image may occur not only locally, but globally, and therefore takes the weighted mean from all image voxels to estimate the value of a voxel v. In other words, the filter does not take into account just the neighborhood of a voxel, but rather all regions that are similar to that neighborhood (hence the non-local name). In [22], the authors compared their proposal, both quantitatively and qualitatively, to other well-known noise reduction algorithms such as the Gaussian filter [23] and the anisotropic diffusion filter [24,25] and showed the superiority of the NLM approach.

2.3.2. Bias field correction

The bias field effect is a phenomenon that causes similar regions to appear with different intensities in MR images. MRI bias field is caused by a range of factors such as poor radio frequency (RF) energy distribution, static magnetic field variations and RF coil sensitivity. According to Arnold et al. [26], if these issues are not corrected, they may substantially degrade results of automatic brain tissue segmentation techniques in MR images. In this work, the N4 (Nick's Nonparametric Nonuniform Intensity Normalization) proposed by Tustison and Gee [27] is used to correct the MRI bias field effect. Advantages of this technique are the non-requirement for pre-segmentation image, independence of pulse frequency and robustness to pathologies (which are desired in the case of images with MS lesions).

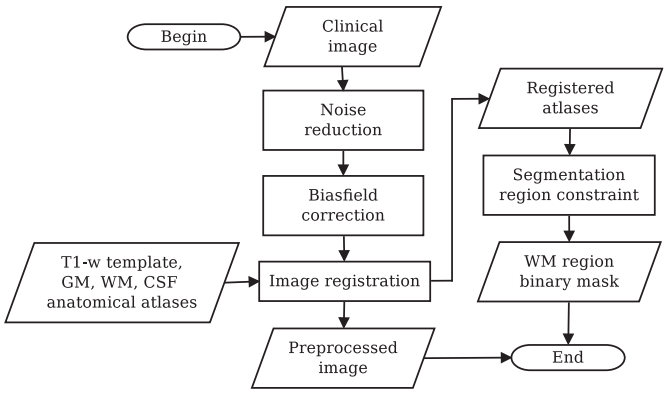


Fig. 1. Flow chart of preprocessing stages.

2.3.3. Image registration

An important step in the preprocessing pipeline is the spatial alignment between the clinical data and the probabilistic anatomical atlases, which are used to spatially constrain the segmentation area and eliminate classification ambiguities. In order to preserve metadata information (such as origin, spacing and orientation) of the clinical images and to avoid sub-pixel image registration artifacts, the atlases were registered to the clinical image instead of the other way around. A T1-weighted image originally registered to the atlases was used as the moving image ($T1_{mov}$), and for each patient time point its corresponding T1-weighted image was used as the reference image ($T1_{ref}$). $T1_{mov}$ was registered to $T1_{ref}$ and the registration information was then applied to the atlases, aligning them to each patient scan. In this work, the NiftyReg[3] tool was used with free-form B-Spline deformation model and multi-resolution approach for non-rigid registration [28].

2.3.4. Segmentation region constraint

Our technique focuses on segmenting MS lesions in WM. Since they can occur anywhere in the WM region [29], we decided to use the registered probabilistic anatomical atlases to constrain our segmentation region to WM. Given a voxel v, its corresponding probabilities in WM, v_{WM}, GM, v_{GM} and CSF, v_{CSF} and a set $S = \{v: v_{WM} > v_{GM}$ and $v_{CSF} > v_{GM}\}$, a binary mask was created using the voxels in S. By doing so, we constrained the segmentation region to WM and also reduced the number of voxels to be clustered by our algorithm.

2.4. Image segmentation

In this work, an iterative Student's t mixture model intensity-based segmentation algorithm is proposed to continuously refine the resulting segmented tissue sub-classes in FLAIR MR images until all MS lesions are grouped as one simple class. We call this approach *Iterative Student's t Mixture Model* (SMM-iter).

A mixture model is a stochastic model commonly used to represent subpopulations [30]. Different brain tissues – such as MS lesions, white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) – can be thought of as subpopulations (or classes) in a brain MR image. One way to distinguish one class from the other is to look at their voxels' intensities. The use of Student's t distribution to cluster brain tissues using intensity information has been proposed in some works [31,32]. The main advantage of the Student's t distribution over other distributions, such as Gaussian Mixture Models (GMM), is that it has a "heavy" tail, which makes it more robust to outliers[4] [33] and, therefore, less susceptible to noise.

Formally, let $\mathbf{X}$ be a random variable, with Student's t distribution probability density function (pdf) $p(\mathbf{X})$. Let $\boldsymbol{\mu}$ be a mean vector, $\boldsymbol{\Sigma}$ be a covariance matrix (positive, definite and real) with

[3] http://sourceforge.net/projects/niftyreg/

[4] We define an outlier as a data sample that is far too distant from all image clusters.

dimension $d \times d$, where d is the data dimension, and ν is the degrees of freedom, $\nu \in [0, \infty)$. Then, $p(\mathbf{X})$ can be written as [34]

$$p(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)\left[1+\nu^{-1}\delta\left(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right]^{\frac{\nu+d}{2}}}, \tag{2}$$

where $\delta(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ represents the Mahalanobis distance and Γ is the Gamma function.

The Student's t distribution is equivalent to a Gaussian distribution with a stochastic covariance matrix [35]. Consider a weight u following a Gamma distribution parameterized by ν:

$$u \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \tag{3}$$

Given that, a variable $\mathbf{X}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\frac{\boldsymbol{\Sigma}}{u}$ as:

$$\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu, u \sim N\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right). \tag{4}$$

We can write a mixture of K Student's t distribution components as

$$\phi\left(\mathbf{X}, \boldsymbol{\Psi}\right) = \sum_{i=1}^{K} \pi_i p\left(\mathbf{X}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i\right), \tag{5}$$

where $\pi_1 \ldots \pi_K$ represent the weights for each component and

$$\boldsymbol{\Psi} = \left(\pi_1, \ldots, \pi_K; \mu_1, \ldots, \mu_K; \Sigma_1, \ldots, \Sigma_K; \ldots; \nu_1, \ldots, \nu_K\right)^T \tag{6}$$

are the mixture model components' parameters.

The estimation of our model parameters was performed using the Expectation Maximization (EM) algorithm [31]. In this sense, consider the complete data vector

$$\mathbf{X}_c = \left(\mathbf{X}_1, \ldots, \mathbf{X}_N, \ z_1, \ldots, z_N, \ u_1, \ldots, u_N\right)^T, \tag{7}$$

where $z_1, \ldots, z_N$ are the component-label vectors and $z_{ij} = (z_j)_i$ is either one or zero, depending on whether the observation $\mathbf{X}_j$ is generated by the ith component or not. In this case, $\mathbf{X}_j$ is an intensity sample from the region to be segmented. Note that, following the definition of the Student's t distribution in Eqs. (2)–(4), the observed data augmented by z_j, $j = 1, \ldots, N$ are still incomplete since the component covariance matrix depends on the degrees of freedom ν. Due to that, the complete data vector also includes the additional missing data $u_1, \ldots, u_N$.

The EM algorithm is divided into two main stages: the E-step (expectation) and M-step (maximization). The E-step on the $(t+1)$th iteration requires the calculation of the posterior probability of $\mathbf{X}_j$ belonging to the ith component of the mixture [35], calculated as

$$z_{ij}^{t+1} = \frac{w_i^t p\left(\mathbf{X}_j; \boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t, \nu_i^t\right)}{\sum_{m=1}^{K} w_m^t p\left(\mathbf{X}_j; \boldsymbol{\mu}_m^t, \boldsymbol{\Sigma}_m^t, \nu_m^t\right)}, \tag{8}$$

as well as the expectation of the weights for each observation [36]

$$u_{ij}^{t+1} = \frac{\nu_i^t + d}{\nu_i^t + \delta\left(\boldsymbol{x}; \boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t\right)}. \tag{9}$$

We also incorporated neighborhood information in our model as proposed by [11,37,12] in the following way. Let v be a voxel and let N_V be the neighborhood of v with radius r and size $s = (2 \times r + 1)^{dim}$, where dim is the number of dimensions in the image (in this context, using 3D images, $dim = 3$). Then, in Eq. (8), w_i^t is calculated as

$$w_i^t = \frac{\pi_i^t + n_{iq}^t}{2}, \tag{10}$$

where

$$n_{iq}^t = \frac{\pi_i^t p\left(\mathbf{X}_q; \boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t, \nu_i^t\right)}{\sum_{m=1}^{K} \pi_m^t p\left(\mathbf{X}_q; \boldsymbol{\mu}_m^t, \boldsymbol{\Sigma}_m^t, \nu_m^t\right)} \tag{11}$$

and $\mathbf{X}_q \in N_v$. This is the mean field approximation of a Markov Random Field [38]. In other words, for each sample $\mathbf{X}_j$, w_i^t was calculated as the average between the ith component weight and the posterior probability of the neighborhood of $\mathbf{X}_j$ belonging to the ith component as well.

Maximizing the log-likelihood of the complete data provides the update equations of the mixture model parameters:

$$\pi_i^{t+1} = \frac{1}{N}\sum_{j=1}^{N} z_{ij}^t, \tag{12}$$

$$\boldsymbol{\mu}_i^{t+1} = \frac{\sum_{j=1}^{N} z_{ij}^t u_{ij}^t \mathbf{X}_j}{\sum_{j=1}^{N} z_{ij}^t u_{ij}^t}, \tag{13}$$

$$\boldsymbol{\Sigma}_i^{t+1} = \frac{\sum_{j=1}^{N} z_{ij}^t u_{ij}^t \left(\mathbf{X}_j - \boldsymbol{\mu}_i^{t+1}\right)\left(\mathbf{X}_j - \boldsymbol{\mu}_i^{t+1}\right)^T}{\sum_{j=1}^{N} z_{ij}^t}. \tag{14}$$

The degrees of freedom ν_i^{t+1} for the ith component are computed as the solution to the equation (adapted from [12])

$$\log\left(\frac{\nu_i^{t+1}}{2}\right) - \psi\left(\frac{\nu_i^{t+1}}{2}\right) + 1 - \log\left(\frac{\nu_i^t + d}{2}\right) + \frac{\sum_{j=1}^{N} z_{ij}^t\left(\log u_{ij}^t - u_{ij}^t\right)}{\sum_{j=1}^{N} z_{ij}^t} + \psi\left(\frac{\nu_i^t + d}{2}\right) = 0, \tag{15}$$

where $\psi(x) = \frac{\partial(\ln \Gamma(x))}{\partial x}$ is the digamma function. In this work, the degrees of freedom were calculated using the NLopt non-linear optimization library.[5]

After the parameters' estimation, we use the *maximum a posteriori* (MAP) classification rule. The classification of each voxel in its respective z label is done by associating it with the class that maximizes the posterior probability $P(z_k|\mathbf{X})$. Formally, the classification stage follows the Bayesian MAP decision given by

$$P\left(z_i|\mathbf{X}\right) > P\left(z_j|\mathbf{X}\right), \tag{16}$$

where $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, k$ and $i \neq j$.

Both sides of the decision rule in Eq. (16) follow the Bayes theorem as [39]:

$$P(z_i \Big| \mathbf{X}) = \frac{p\left(\mathbf{X}|z_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right)p\left(z_i\right)}{\sum_{j=1}^{K} p(\mathbf{X}\Big|z_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)p(z_j)}, \tag{17}$$

where $p(z_i)$ represents the *a priori probability of class i*, $p(\mathbf{X}|z_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ *is the pdf of class i and the parameter K represents the number of mixture components.*

In the context of this work, the random variable $\mathbf{X}$ represents the gray level intensities in FLAIR images. Since we did not use other image weights (such as T2, for instance), $d = 1$ and the mean vector $\boldsymbol{\mu}$ becomes simply the univariate mean and the covariance matrix $\boldsymbol{\Sigma}$ becomes the standard deviation.

[5] http://ab-initio.mit.edu/wiki/index.php/NLopt

2.4.1. Iterative approach

Classifying MS lesions using an intensity-based clustering technique often suffers from a large amount of false positives [29]. This is due to the fact that MS lesions may have pdf of the intensity levels close to those of other normal brain tissues. Given that, we decided to use an iterative approach to successively approximate the lesion segmentation (hence the *iter* in SMM-iter). This approach allows a segmentation refinement in a way that on each iteration the differences between lesions and other tissues increase, hence reducing the number of false positives.

The iterative approach, shown in Fig. 2, is composed of the following steps:

1. Define a preprocessed image I_{seg} to be segmented and a binary mask BM to spatially constrain the image segmentation.
2. Define the number of iterations (N_{iter}) and the number of clusters $(N_{clusters})$ for each iteration.
3. Segment I_{seg} using Student's t mixture model explained in Section 2.4 with BM and $N_{clusters}$ defined for that iteration.
4. For each cluster, calculate its mean intensity.
5. Use the regions of the $(N_{clusters} - 1)$ highest mean intensity clusters of that iteration as the new BM for the next iteration.
6. Loop to step 3 until the number of iterations reaches zero.

In this work, we used $N_{iter} = 5$, $N_{clusters} = 3$ for the first four iterations and $N_{clusters} = 2$ for the last iteration. These parameters were empirically determined using a training image dataset where they have shown to provide the best trade-off between processing time and lesion segmentation accuracy.

The initial binary mask $(BM^{(N_{iter})})$ was created using the anatomical probabilistic atlases. The idea was to constrain the segmentation region to the WM area. It is a fact that the alignment between the atlases and the clinical images is not perfect. The registered WM atlas may not adequately represent the real WM area in the clinical image. In this scenario, due to misalignments, some lesion voxels located in the clinical WM area may have low probabilities in the WM atlas. To circumvent this problem, a threshold $T=0.20$ was defined and all voxels in the GM atlas with a probability smaller than T were given as input to create $BM^{(N_{iter})}$. Voxels with low GM probability have higher probabilities of belonging either to CSF of WM. Therefore, the initial binary mask covered voxels from both these classes. CSF voxels appear hypointense in FLAIR images, being discarded from the lesion cluster during the iterative segmentation. On the other hand, hyperintense voxels that would be left out in the initial binary mask if we used only the WM atlas are taken into account using this approach.

Note that on each iteration we discard the lowest mean intensity cluster and update the binary mask with the remaining clusters of that iteration. The binary mask update is done in this manner because MS lesions appear hyperintense in FLAIR images. Therefore we wanted to preserve the high mean intensity clusters while refining our segmentation.

This iterative approach makes the assumption that MS lesions are the most hyperintense class in FLAIR images. The successive refinement eliminates classes that do not follow such assumption, thus reducing the number of false positives. However, a limitation of this approach lies in the fact that it may also reduce the number of true positives in images where there is a significant intensity variability among MS lesions.

The output of this iterative approach is a binary mask $BM_{lesions}$ representing all MS lesions segmented by our technique in image I_{seg}.

2.5. Post-processing

Post-processing techniques, including mid-sagittal plane detection to help eliminate in-between ventricles (third ventricle – TV) voxels and a residue removal method, were used to remove false positives from the resulting lesion binary mask.

2.5.1. Mid-sagittal plane detection

Since the TV region in FLAIR images commonly has an intensity overlap with MS lesions [29], voxels located in this region are mistakenly clustered together with lesions and increase the number of false positives. Since MS lesions located close to the TV region are quite rare, we used the technique proposed in [40] to detect the mid-sagittal plane (MSP), defined a distance threshold $d_{threshold}$ and removed all voxels in which distance to the plane was smaller or equal to $d_{threshold}$. In other words, given a voxel v and its distance to the mid-sagittal plane $d(v) = d_{msp}$, we removed v from the image if $d_{msp} \leq d_{plane}$. This approach was effective in removing misclassified voxels in the TV region while keeping the lesion voxels intact. A comparison between images before and after using this technique is shown in Fig. 3.

2.5.2. Residue removal

The final post-processing step was to remove residual voxels. We defined residual voxels as follows: given a voxel v and a neighborhood N_v with radius r around v and size $s = (2 \times r + 1)^{dim}$,

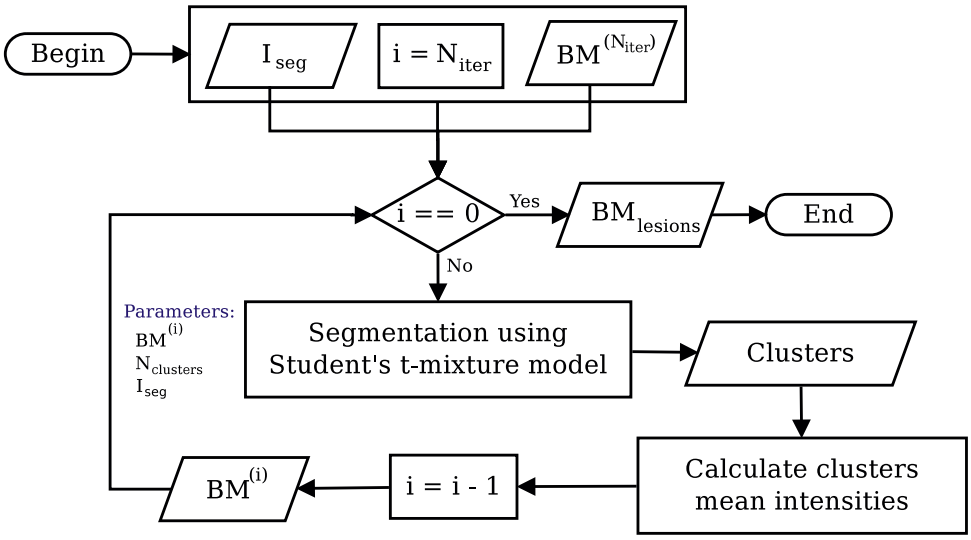


Fig. 2. Flow chart of the proposed iterative approach.

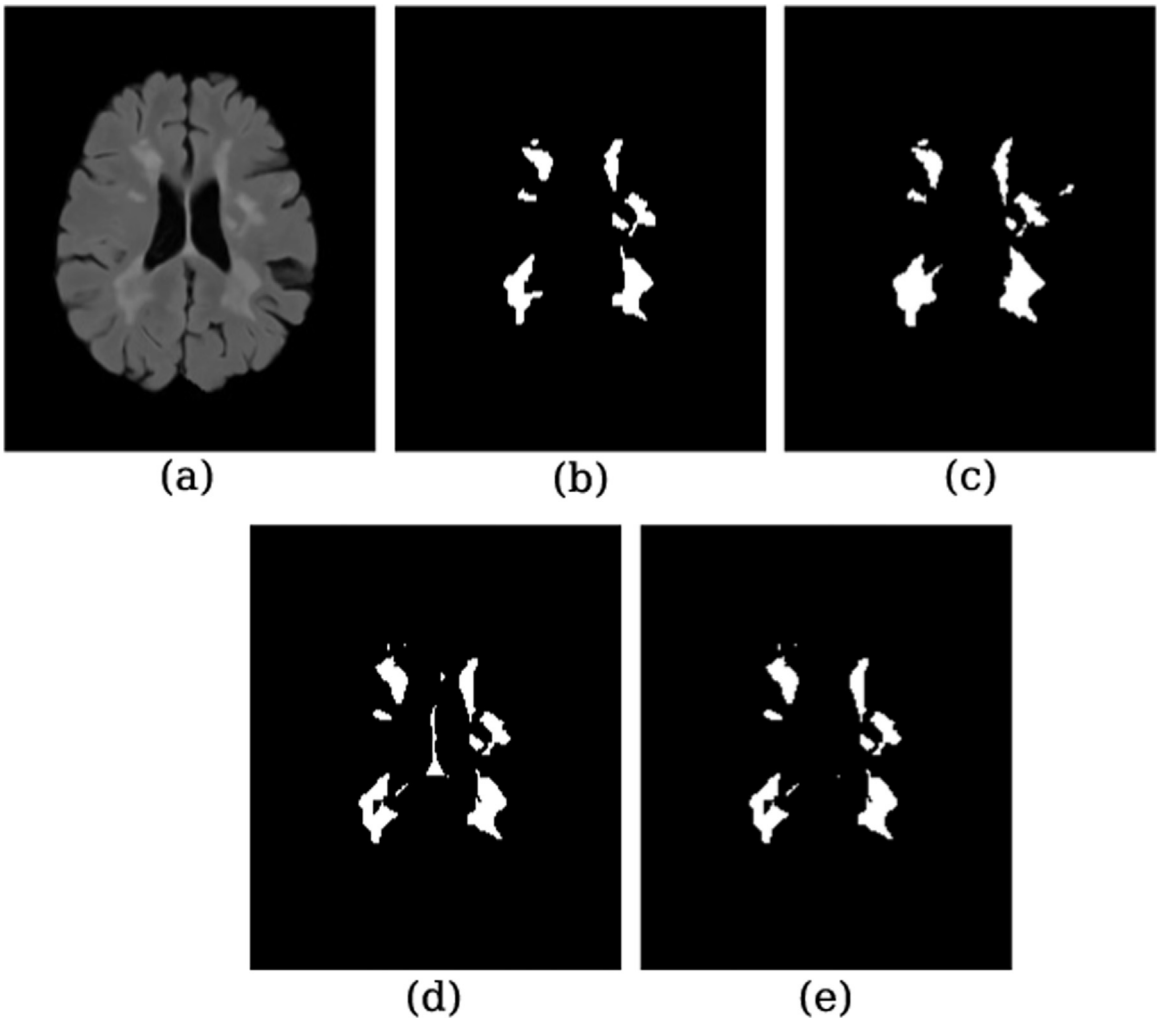


Fig. 3. Comparison between lesion segmentation masks. In (a), we have the original FLAIR image. The ground truths from raters 1 and 2 are shown in (b) and (c), respectively. In (d), we have the automatic segmentation without MSP detection and, finally, in (e) the output after applying the MSP detection and third ventricle removal is shown.

v is a residual voxel if its neighborhood has less than $\frac{s}{3}$ active voxels (by active voxels we mean voxels with values equal to 1 in the binary mask). In other words, we aimed to remove isolated voxels that were unlikely to belong to the MS lesion cluster.

Granted, this approach is prone to remove small lesions from the segmentation, and that is a limitation of our technique. However, as shown in Table 2, the DSC improved with the use of residue removal.

3. Results and discussion

In this section, the results of our segmentation algorithm are presented and discussed. An analysis of inter-specialist agreement is also provided.

Table 2
Results of our automatic image segmentation performed with and without the post-processing stage of residue removal. All comparisons were made between our proposed automatic segmentation method and the ground truth provided by rater 1.

Image source	DSC with residue	DSC without residue
Patient 1	0.675	0.698
Patient 2	0.850	0.861
Patient 3	0.674	0.681
Patient 4	0.685	0.722
Patient 5	0.657	0.680

3.1. Segmentation results

We used the preprocessing, segmentation and post-processing stages explained in Sections 2.3–2.5, respectively, to automatically segment MS lesions in our image dataset. As previously mentioned, each time point had two manual delineations performed by two different raters. We compared our segmentation to both of them. An example of the iterative segmentation followed by the post processing stage is shown in Fig. 4.

For qualitative analysis purposes, two automatic segmentation results along with the ground truths from both raters are shown in Figs. 5 and 6.

For each patient, we took the mean and standard deviation of DSC, TPR, FPR and VD metrics for all patient's scans. Results are shown in Table 3 and 4. We also took one time point from each patient and extracted their corresponding DSC value on each iteration (using rater 1 as ground truth) to offer an overview regarding the iterative approach progress; the results are shown in Table 5.

The greatest differences on the metrics values between rater 1 and rater 2 ground truths were observed in patient 4 time points. As shown in Table 6, the agreement between the raters' delineations for this case ($\mathrm{DSC} = 0.612 \pm 0.019$) presented the smallest DSC value among all patients, indicating a high rater inter-variability. Also for comparison, all lesion absolute volumes of all time points from patient 4 are presented in Table 7. Patient

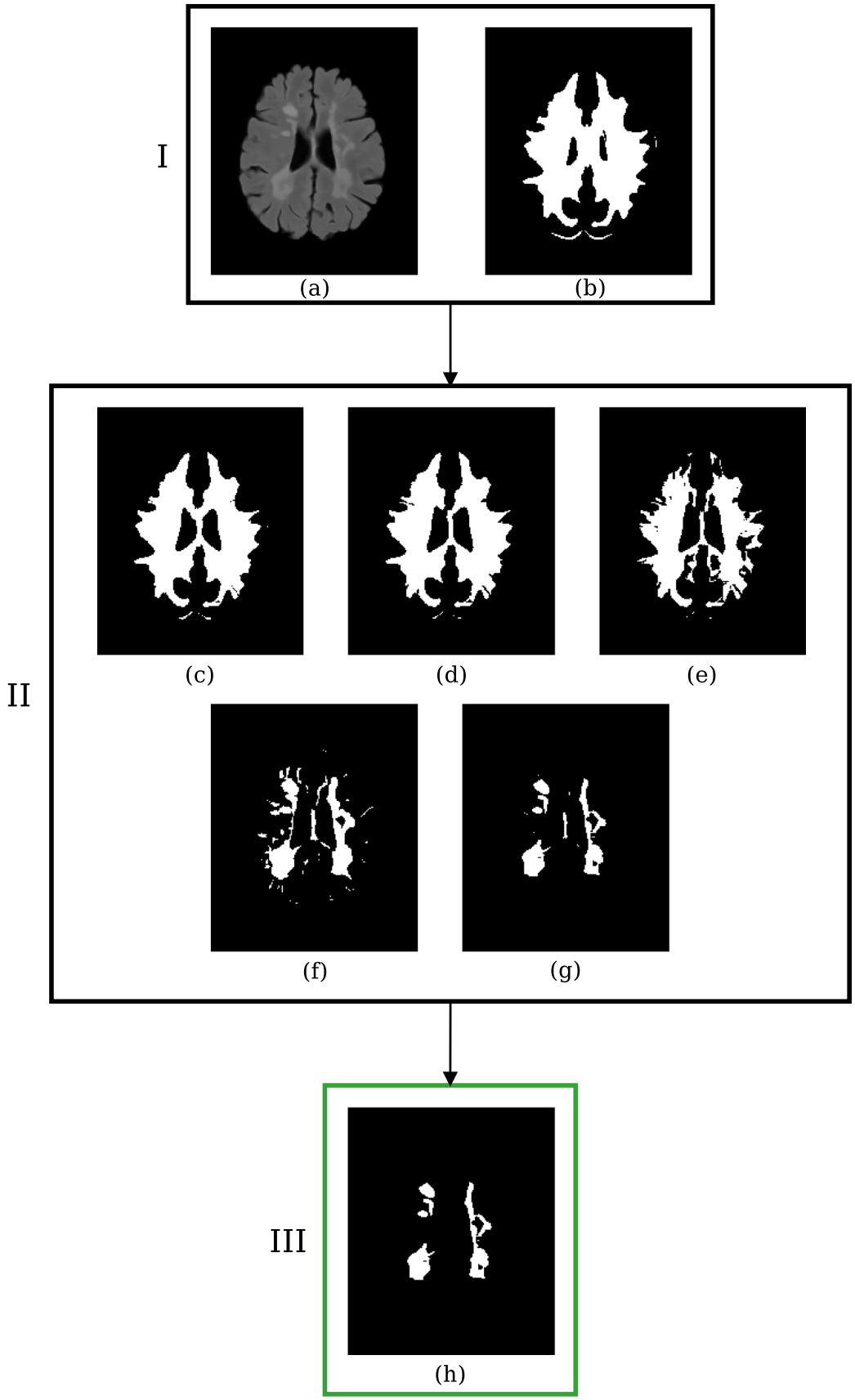


Fig. 4. Iterative segmentation. In (I), (a) is the image to be segmented and (b) is the binary mask used to constrain the initial segmentation region. In (II), the segmentation is performed in an iterative manner. Note that the binary mask shrinks from (c) to (g). Finally, in (III) we have the output of the post processing stage.

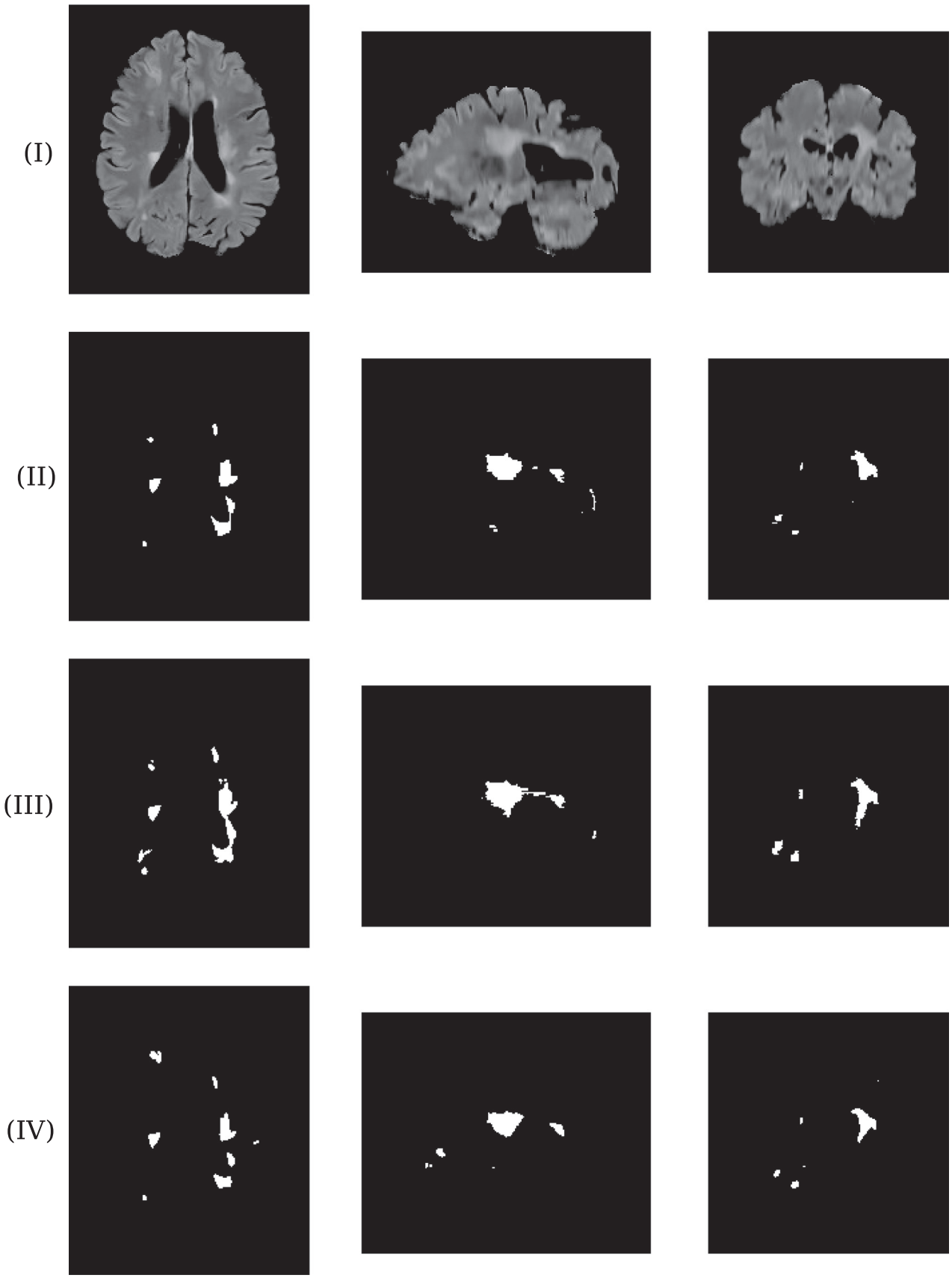


Fig. 5. Axial (left column), sagittal (center column) and coronal (right column) views of a time point from patient 1 (row I) and its respective binary masks from raters 1 (row II), 2 (row III) and our proposed technique (row IV).

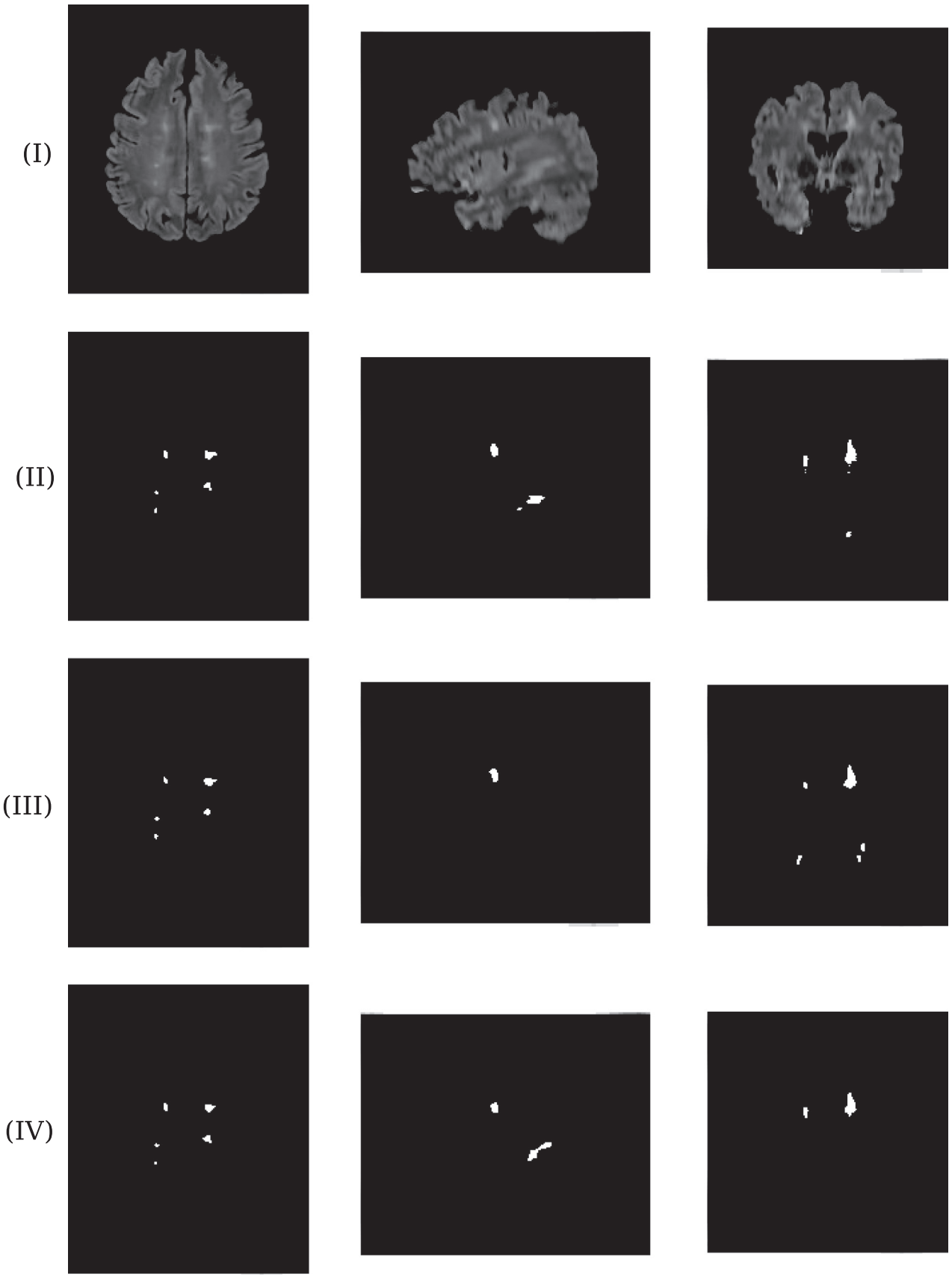


Fig. 6. Axial (left column), sagittal (center column) and coronal (right column) views of a time point from patient 5 (row I) and its respective binary masks from raters 1 (row II), 2 (row III) and our proposed technique (row IV).

Table 3
Segmentation results (mean ± standard deviation) using manual delineations from rater 1 as ground truth.

Image source	Scans	DSC	TPR	FPR	VD
Patient 1	4	0.644 ± 0.056	0.607 ± 0.164	0.259 ± 0.200	0.328 ± 0.09
Patient 2	4	0.716 ± 0.114	0.604 ± 0.181	0.05 ± 0.06	0.336 ± 0.230
Patient 3	5	0.642 ± 0.056	0.581 ± 0.074	0.245 ± 0.285	0.338 ± 0.115
Patient 4	4	0.622 ± 0.080	0.557 ± 0.159	0.220 ± 0.250	0.390 ± 0.120
Patient 5	4	0.634 ± 0.060	0.595 ± 0.097	0.272 ± 0.065	0.136 ± 0.146

Table 4
Segmentation results (mean ± standard deviation) using manual delineations from rater 2 as ground truth.

Image source	Scans	DSC	TPR	FPR	VD
Patient 1	4	0.641 ± 0.083	0.573 ± 0.176	0.182 ± 0.121	0.345 ± 0.09
Patient 2	4	0.701 ± 0.095	0.579 ± 0.144	0.054 ± 0.061	0.364 ± 0.195
Patient 3	5	0.577 ± 0.086	0.464 ± 0.127	0.123 ± 0.153	0.410 ± 0.262
Patient 4	4	0.478 ± 0.043	0.367 ± 0.095	0.155 ± 0.186	0.474 ± 0.277
Patient 5	4	0.586 ± 0.055	0.592 ± 0.110	0.411 ± 0.098	0.153 ± 0.107

Table 5
DSC values on each iteration for one time point from each patient using rater 1 as ground truth.

Iteration number	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Iteration 1	0.065	0.111	0.026	0.025	0.021
Iteration 2	0.074	0.117	0.064	0.040	0.035
Iteration 3	0.181	0.151	0.162	0.091	0.087
Iteration 4	0.398	0.454	0.383	0.226	0.244
Iteration 5	0.622	0.822	0.601	0.398	0.553
After post-processing	0.644	0.861	0.628	0.531	0.634

4 had the smallest lesion load among all other patients. In this case, it is reasonable to expect the automatic segmentation to perform rather poorly, hence the variability in volume and low spatial agreement with raters.

The results in Table 5 show that the iterative approach is able to refine the segmentation in a successive manner, improving the DSC on each step. The post-processing stage is then used to remove false positives, improving the spatial agreement between the rater and the automatic segmentation.

Lesion volumes in this study ranged from 1.95 ml to 34.71 ml. Given the DSC, TPR and FPR values, our technique was able to maintain a good segmentation consistency in this range.

Table 6
DSC values (mean ± standard deviation) for raters 1 and 2 manual delineations.

Image source	Scans	DSC (agreement between raters)
Patient 1	4	0.779 ± 0.017
Patient 2	4	0.846 ± 0.036
Patient 3	5	0.768 ± 0.076
Patient 4	4	0.612 ± 0.019
Patient 5	4	0.652 ± 0.048

Table 7
Absolute lesion volumes for all time points of patient 4 from raters 1, 2 and the automatic segmentation.

Patient 4	Rater 1 (volume in ml)	Rater 2 (volume in ml)	SMM-iter (volume in ml)
Time point 1	2.29	3.50	1.30
Time point 2	2.19	3.12	2.93
Time point 3	1.95	3.52	1.45
Time point 4	2.28	2.81	1.06

However, as mentioned in Section 2.4.1, our technique chooses the most hyperintense class in the end of the iterative approach as the MS lesion class. This may exclude lesion voxels that do not comply with the hyperintense assumption, thus decreasing the TPR and increasing the VD. A comparison among ground truths 1, 2 and the automatic segmentation regarding the mean and standard deviation of MS lesions for all time points from patient 1 are shown in Table 8.

It can be seen that our technique has the smallest standard deviation in all time points, which indicates that our algorithm segmented a narrower portion of MS lesions. Also, except for time point 2, our technique had the biggest mean among all cases. The combination of a big mean and a small standard deviation is an indication that our proposed technique indeed segmented MS lesions, but it was not able to include some lesion voxels located in the "low end" of the lesion intensity profile. The same behavior was observed in all other time points from all other patients.

In Section 2.4, we mentioned that the Student's t distribution has a heavy tail, which makes it more robust to outliers and offers an advantage over GMM. Many works in the literature propose the use of GMM to perform image segmentation [41–43]. For the sake of comparison, we used the same methodology explained in Sections 2.3–2.5 to perform the segmentation of all time points from patient 1 using GMM and then compared the results with SMM-iter. The DSC values of the results are shown in Table 9. It can be seen that the SMM-iter had slightly better results than the GMM.

We also used Pearson's r coefficient to assess the volumetric correlation between our segmentation and the ground truths. For rater 1, $r=0.8813$, and for rater 2, $r=0.8871$. For the sake of comparison, we also calculated the volumetric correlation between raters. In this scenario, $r=0.9841$. For all cases, p *value* <0.00001. The scatter plots for this data are shown in Figs. 7–9.

The results for the Pearson's r coefficient indicated a strong positive correlation between the volumes of our automatic segmentation and the ground truths from both raters. Considering these results along with the ones presented in Tables 3 and 4 we can observe that our segmentation proposal had good spatial and volumetric agreement with the ground truths. Comparatively, the correlation between raters was higher than both correlations from the automatic segmentation and raters 1 and 2. This difference can be explained by the fact that our technique segmented a narrower portion of MS lesions, leaving out the ones in the low end of the intensity profile.

Table 8
Comparison of mean ± standard deviation of MS lesions among ground truth 1 (GT1), ground truth 2 (GT2) and the automatic segmentation for images from patient 1.

Time points	$\mu_{GT1} \pm \sigma_{GT1}$	$\mu_{GT2} \pm \sigma_{GT2}$	$\mu_{automatic} \pm \sigma_{automatic}$
Time point 1	139.88 ± 12.49	137.58 ± 12.44	145.32 ± 9.27
Time point 2	112.20 ± 8.71	110.80 ± 9.05	110.52 ± 7.45
Time point 3	103.23 ± 8.31	102.978 ± 8.65	109.02 ± 6.26
Time point 4	118.20 ± 9.72	116.84 ± 10.23	123.37 ± 7.56

An advantage of our technique compared to the works described in Section 1 is that we do not use any training stage. Each segmentation is done in a stand-alone manner, and while the number of iterations and clusters is empirically defined, this process does not require the same level of training as other techniques such as a random forest or a neural network. In [13, 15–17], the authors' proposals either needed training or healthy subjects scans (or both) to perform the segmentation. This can be a setback when dealing with images from different centers, since they would have to go through a normalization stage which may lead to biased results because the training dataset does not take into account anatomical and physiological variability between different subjects [44].

In [14], the authors used patches and multimodal information to segment MS lesions. This approach did not require a training stage and was able to obtain a high TPR and low FPR. However, the DSC was relatively low (0.55 ± 0.14) and the authors did not use spatial information to distinguish similar intensity voxels.

The winner of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge was Vaidya et al. [16]. The authors used neural networks and patches to segment MS lesions. The neural networks had to be trained before the segmentation could take place. Since in [16] the authors presented their results on just a few time points from one patient, it is not possible to fully compare their approach to our proposal. But even though their DSC values for the presented time points were higher than ours, our

Table 9
DSC values for all time points from patient 1 using ground truths from raters 1 and 2. The same methodology explained in Sections 2.3–2.5 was used to perform the segmentation using GMM. SMM-iter values are shown for comparison.

Time points	Rater 1 (SMM-iter)	Rater 2 (SMM-iter)	Rater 1 (GMM)	Rater 2 (GMM)
Time point 1	0.675	0.633	0.675	0.660
Time point 2	0.698	0.760	0.698	0.760
Time point 3	0.570	0.566	0.552	0.527
Time point 4	0.633	0.605	0.586	0.605
Mean ± std dev	0.644 ± 0.05	0.641 ± 0.08	0.627 ± 0.06	0.638 ± 0.09

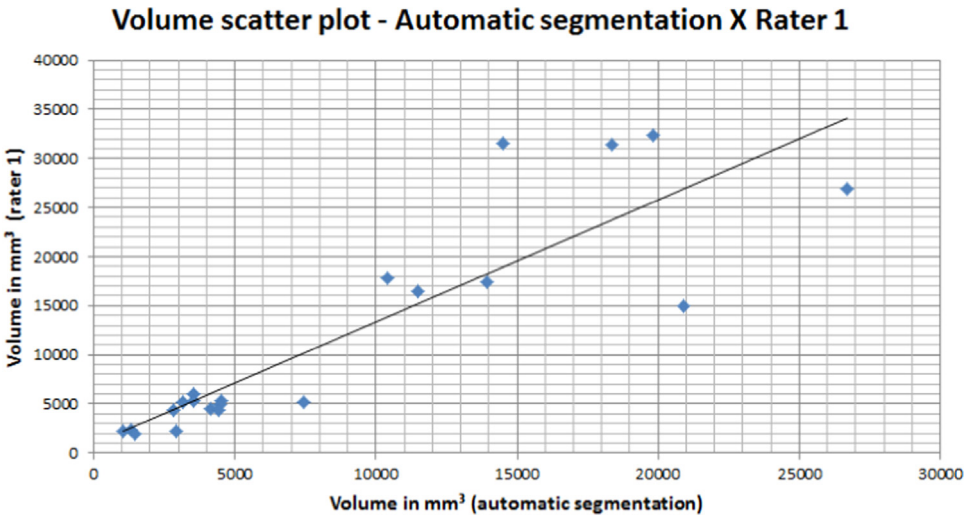


Fig. 7. Scatter plot for volumes from the automatic segmentation (horizontal axis) and rater 1 (vertical axis). For this data, $r=0.8813$ (p *value* <0.00001).

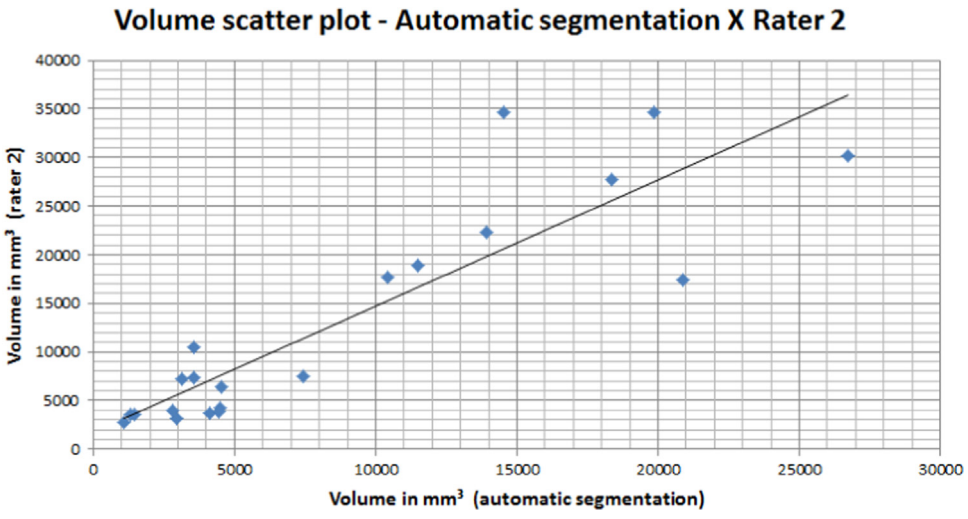


Fig. 8. Scatter plot for volumes from the automatic segmentation (horizontal axis) and rater 2 (vertical axis). For this data, $r=0.8871$ (p *value* <0.00001).

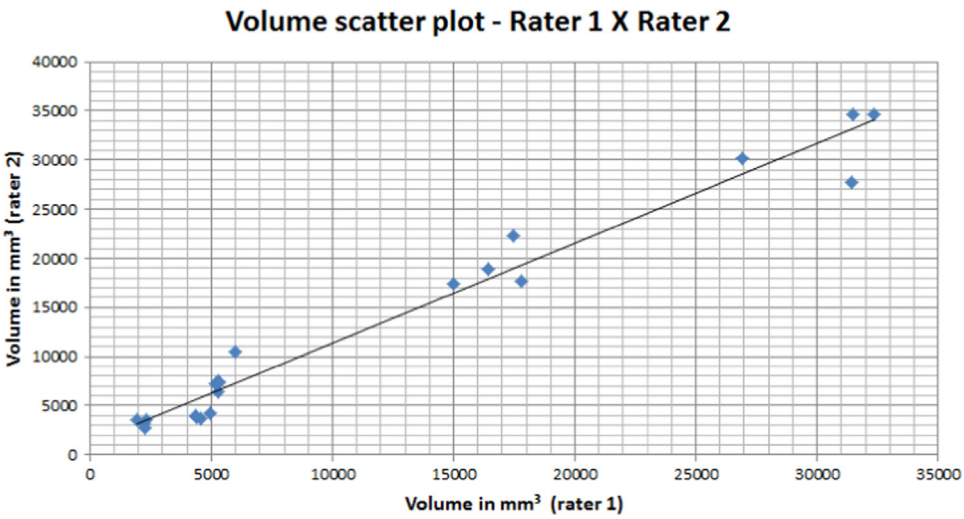


Fig. 9. Scatter plot for volumes from rater 1 (horizontal axis) and rater 2 (vertical axis). For this data, $r=0.9841$ (p *value* <0.00001).

Table 10
Segmentation proposals that used the 2015 ISBI Longitudinal Challenge. For comparison, our proposal's results are shown in the last row. GT 1 and GT 2 are the ground truths of raters 1 and 2, respectively.

Authors	Technique	Results	Image weights	Ground truth
Jesson and Arbel	Random forest	DSC=0.70/0.68 TPR=0.61/0.50 FPR=0.13/0.12	T1, T2, FLAIR	GT 1/GT 2
Maier and Handels	Random forest	DSC=0.70/0.65 TPR=0.53/0.38 FPR=0.48/0.43	T1, T2, PD, FLAIR	GT 1/GT 2
Vaidya et al.	Neural network	DSC=0.80 TPR=0.40 FPR=0.59	T1, T2, PD, FLAIR	GT 1 (only patient 2 time points)
Tomas-Fernandez and Warfield	GMM	DSC=0.62 TPR=0.53 FPR=0.48	T1, T2, FLAIR	Not clear
Prados et al.	Fast patch matching	DSC = 0.55 ± 0.14 TPR = 0.72 ± 0.10 FPR = 0.08 ± 0.02	T1, T2, PD, FLAIR	GT1 ∩ GT 2
Freire and Ferrari	Iterative SMM	DSC=0.65/0.59 TPR=0.58/0.51 FPR=0.21/0.18	FLAIR	GT 1/GT 2

approach had a higher TPR in some time points and lower FPR for some others. This indicates that our technique had a smaller misclassification ratio. It also shows that analysis using different metrics is important to obtain a good overview of automatic segmentation proposals.

The mean DSC, TPR and FPR values for the works described in Section 1 and our proposal are presented in Table 10. GT 1 and GT 2 are the ground truths from raters 1 and 2, respectively. The metric values were calculated taking into account every time point, except when stated otherwise.

Overall, our technique showed a good agreement with the manual delineations (especially the ones from rater 1) and was able to keep segmentation consistency among images with different lesion volumes. Regarding processing time, the average time to go through the preprocessing, segmentation and post-processing stages for each time point in the 2015 ISBI Longitudinal Challenge was around 20 min (Core i7 with 16GB RAM). Also, the fact that our technique does not require any training stage is an advantage when images from different centers (or with different parameters acquisition) have to be segmented because there is no bias introduced from training images.

4. Conclusions and future work

This paper presented an automatic iterative approach to segment MS lesions in FLAIR images. Our technique uses probabilistic anatomical atlases to constrain the segmentation process to the WM region (thus reducing the number of false positives and voxels to cluster) and segment the image using Student's t mixture models. Our method is performed in a refinement manner by iteratively segmenting brain tissues into smaller classes until MS lesions are grouped as the most hyperintense one.

We compared the results from our automatic technique to

2 raters' manual delineations and to other works that used the same dataset using the DSC, TPR, FPR and VD metrics. Results indicated that our technique successfully segments MS lesions in FLAIR images and provides consistent results on images with a different lesion volume.

A difficulty faced during the development of our proposed technique regarded the high inter-variability between the raters. This fact created difficulties in providing a solid foundation that could help us work on the adjustment of the parameters of our algorithm.

One limitation of our technique lies in the image registration stage. Since the alignment between fixed and moving images is not perfect, lesions located in areas that are not well aligned are likely to be pruned in the image segmentation constraint stage. On the other hand, this constraint also reduces the number of false positives by eliminating voxels that have intensity profiles close to those of MS lesions but are outside the WM region.

Another drawback of our technique is that lesion voxels that are on the "low end" of the MS lesion intensity profile are likely to be left out of the MS cluster. This happens because our proposal follows the assumption that MS lesions are the most hyperintense class in FLAIR images. If the lesions have a significant intensity variability, the less intense ones are prone not to be included in the cluster.

In future works, we aim to extend the approach presented in this paper to other MRI weights such as T2 and also to develop a model to automatically select the best number of clusters and iterations for our algorithm and circumvent the "low end" intensity problem. We also intend to apply our segmentation proposal to different field strengths and data acquired on different scanners to assess its performance in different scenarios.

Conflict of interest

None declared.

Acknowledgments

The authors would like to thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the financial support of this research (Grant nos. 2014/00019-6, 2012/16546-0 and 2012/03100-3).

References

[1] Multiple Sclerosis International Federation, 2013 Atlas of MS, ⟨http://www.msif.org/about-us/advocacy/atlas/⟩, 2013 (accessed on 12.01.2015).
[2] A. Compston, A. Coles, Multiple sclerosis, Lancet 372 (9648) (2008) 1502–1517.
[3] M. Filippi, M. Horsfield, S. Bressi, C. Baratti, P. Reganati, Intraobserver and interobserver variability in schemes for estimating volume of brain lesions on MR images in multiple sclerosis, Am. J. Neuroradiol. 19 (2) (1998) 239–244.
[4] C. Tur, X. Montalban, M. Tintoré, C. Nos, J. Río, F. Aymerich, L. Brieva, N. Téllez, H. Perkal, M. Comabella, I. Galán, D. Calle, J. Sastre-Garriga, A. Rovira, Interferon beta-1b for the treatment of primary progressive multiple sclerosis: five-year clinical trial follow-up, Arch. Neurol. 68 (11) (2011) 1421–1427.
[5] S. Jain, D. Simaa, A. Ribbensa, M. Cambronb, A. Maertensa, W. Heckea, J. Meyb, F. Barkhofc, M. Steenwijkc, M. Daamsc, F. Maesd, S. Huffele, H. Vrenkenc, D. Smeetsa, Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images, NeuroImage 8 (2015) 367–375.
[6] N. Weiss, D. Rueckert, A. Rao, Multiple sclerosis lesion segmentation using dictionary learning and sparse coding, Med. Image Comput. Comput.-Assist. Interv. 1 (2013) 735–742.
[7] E. Sweeney, R. Shiohara, N. Shiee, F. Mateen, A. Chudgar, J. Cuzzocreo, P. Calabresi, D. Pham, D. Reich, C. Crainiceanu, OASIS is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI, Neuroimage: Clin. 2 (2013) 402–413.
[8] E. Sweeney, R. Shinohara, C. Shea, D. Reich, C. Crainiceanu, Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI, Am. J. Neuroradiol. 34 (1) (2013) 68–73.
[9] S. Roy, Q. He, A. Carass, A. Jog, J. Cuzzocreo, D. Reich, J. Prince, D. Pham, Example based lesion segmentation, in: S. Ourselin, Martin A. Styner (Eds.), Proceedings of SPIE, vol. 9034, 2014.
[10] E. Geremia, O. Clatz, B. Menze, E. Konukoglu, A. Criminsi, N. Ayache, Spatila decision forests for MS lesion segmentation in multi-channel magnetic resonance images, NeuroImage 57 (2011) 378–390.
[11] T. Nguyen, Q. Wu, Robust student's t-mixture model with spatial constraints and its application in medical image segmentation, IEEE Trans. Med. Imaging 31 (1) (2012) 103–116.
[12] H. Zhang, Q. Wu, T. Nguyen, Image segmentation by a new weighted student's t-mixture model, Inst. Eng. Technol.—Image Process. 7 (3) (2013) 240–251.
[13] A. Jesson, T. Arbel, Hierarchical MRF and random forest segmentation of MS lesions and healthy tissue in brain MRI. Available at: ⟨http://www.iacl.ece.jhu.edu/w/images/7/72/Andrew_Jesson.pdf⟩, 2015.
[14] F. Prados, M. Cardoso, N. Cawley, O. Ciccarelli, C. Wheeler-Kingshott, S. Ourselin, Multi-contrast patchmatch algorithm for multiple sclerosis lesion detection. Available at: ⟨http://www.iacl.ece.jhu.edu/w/images/3/33/Ferran_Prados_Carrasco.pdf⟩, 2015.
[15] O. Maier, H. Handels, MS lesion segmentation in MRI with random forests. Available at: ⟨http://www.iacl.ece.jhu.edu/w/images/d/d7/Oskar_Maier.pdf⟩, 2015.
[16] S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks. Available at:⟨http://www.iacl.ece.jhu.edu/w/images/1/19/Suthirth_Vaidya.pdf⟩, 2015.
[17] X. Tomas-Fernandez, S. Warfield, Longitudinal multiple sclerosis lesion segmentation challenge: model of population and subject (MOPS) segmentation. Available at: ⟨http://www.iacl.ece.jhu.edu/w/images/d/d6/Xavier_Tomas-Fernandez.pdf⟩, 2015.
[18] V. Fonov, A. Evans, R. McKinstry, C.R. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage 47 (2009) S102.
[19] L. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.
[20] J. Bartko, Measurement and reliability: statistical thinking considerations, Schizophr. Bull. 17 (3) (1991) 483–489.
[21] M. Bland, An Introduction to Medical Statistics, 3rd edition, University Press, Oxford, UK, 2000.
[22] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2 (2005) 60–65.
[23] M. Lindenbaum, M. Fischer, A. Bruckstein, On Gabor contribution to image enhancement, Pattern Recognit. 27 (1) (1994) 1–8.
[24] P. Perona, J. Malik, Scale space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell. 12 (7) (1990) 629–639.
[25] R. Ferrari, Off-line determination of the optimal number of iterations of the robust anisotropic diffusion filter applied to denoising of brain MR images, Med. Biol. Eng. Comput. 51 (1–2) (2013) 71–88.
[26] J. Arnold, J. Liow, K. Schaper, J. Stern, J. Sled, D. Shattuck, A. Worth, M. Cohen, R. Leahy, J. Mazziotta, D. Rottenberg, Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects, NeuroImage 13 (5) (2001) 931–943.
[27] N. Tustison, J. Gee, N4ITK: Nick's N3 ITK implementation for MRI bias field correction, Penn Image Computing and Science Laboratory.
[28] M. Modat, G. Ridgway, Z. Taylor, M. Lehman, J. Barnes, D. Hawkes, N. Fox, S. Ourselin, Fast free-form deformation using graphics processing units, Comput. Methods Prog. Biomed. 98 (3) (2010) 278–284.
[29] N. Shiee, P. Bazin, A. Ozturk, D. Reich, P. Calabresi, D. Pham, A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions, NeuroImage 49 (2010) 1524–1535.
[30] D. Titterington, A. Smith, Statistical Analysis of Finite Mixture Distributions, Wiley, Hoboken, USA, 1985.
[31] D. Peel, G. McLachlan, Robust mixture modelling using the t distribution, Stat. Comput. 10 (1) (2000) 339–348.
[32] G. Sfikas, C. Nikou, N. Galatsanos, Robust image segmentation with mixtures of student's t-distributions, in: IEEE International Conference on Image Processing, San Antonio, USA, vol. 1, 2007, pp. 273–276.
[33] S. Chatzis, D. Kosmopoulos, T. Varvarigou, Robust sequential data modeling using an outlier tolerant hidden Markov model, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1657–1669.
[34] K. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, Cambridge, Cambridge, USA, 2012.
[35] D. Gerogiannis, C. Nikou, A. Likas, The mixtures of student's t-distributions as a robust framework for rigid registration, Image Vision. Comput. 27 (1) (2009) 1285–1294.
[36] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions: Series in Probability and Statistics, John Wiley & Sons, Hoboken, USA, 1997.
[37] M. Grebol, Atlas-based segmentation of multiple sclerosis lesions in magnetic resonance imaging (Ph.D. thesis), Universitat de Girona, 2013.
[38] J. Zhang, The mean field theory in EM procedures for Markov Random Field, Trans. Signal Process. 40 (10) (1992) 2570–2583.
[39] O. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edition, Wiley-Interscience, Hoboken, USA, 2000.
[40] G. Ruppert, L. Teverovskiy, C. Yu, A. Falcão, Y. Liu, A new symmetry-based method for mid-sagittal plane extraction in neuroimages, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2011,

pp. 285–288.
[41] K. Blekas, A. Likas, N. Galatsanos, I. Lagaris, A spatially constrained mixture model for image segmentation, IEEE Trans. Neural Netw. 16 (2) (2005) 494–498.
[42] R. Khayati, M. Vafadust, F. Towhidkhah, S. Nabavi, Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model, Comput. Biol. Med. 38 (2008) 379–380.
[43] A. Bijar, R. Khayati, Segmentation of MS lesions using active contour model, adaptive mixtures method and mrf model, in: 2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA), 2011, pp. 159–164, ⟨http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6046599⟩.
[44] I. Despotovic, B. Goossens, W. Philips, MRI segmentation of the human brain: challenges, methods, and applications, Comput. Math. Methods Med. 2015 (2015) 23 p.

Midsaggital plane detection in magnetic resonance images using phase congruency, Hessian matrix and symmetry information: a comparative study

Conference Paper · May 2018

CITATION
1

READS
97

5 authors, including:

Paulo Guilherme de Lima Freire
Universidade Federal de São Carlos
17 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Bruno César Gregório da Silva
Universidade Federal de São Carlos
17 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Carlos Henrique Villa Pinto
Universidade Federal de São Carlos
15 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Camilo Aparecido Ferri Moreira
Universidade Federal de São Carlos
5 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Midsaggital plane detection in magnetic resonance images using phase congruency, Hessian matrix and symmetry information: a comparative study

Paulo. G. L. Freire, Bruno C. Gregório da Silva, Carlos H. Villa Pinto, Camilo A. Ferri Moreira, and Ricardo J. Ferrari

Departament of Computer Science, Federal University of São Carlos, SP 13565-905, Brazil,
{paulo.freire, rferrari}@dc.ufscar.br,
Home page: http://bipgroup.dc.ufscar.br

Abstract. The human brain has two hemispheres that are separated by a midsaggital plane (MSP) in the medial longitudinal fissure. By using the MSP, it is possible to assess clinically relevant information such as brain atrophy and asymmetry. In this paper, we evaluate the performance of three techniques (phase congruency, Hessian-based and symmetry-based approaches) for MSP detection using the normal angle difference and average Z-distance metrics to assess accuracy. The datasets were comprised of simulated and clinical magnetic resonance (MR) images with different weights, levels of noise and intensity inhomogeneities. Our results indicate the phase congruency approach is the most accurate of the three techniques, with an average normal angle difference of 0.446° and 0.395 average Z-distance for the simulated database and 0.705° and 0.753 for the clinical database. The Hessian-based approach has the best trade-off between accuracy and execution time; on average, the Hessian-based algorithm takes roughly 9 seconds to detect the MSP, 10-fold faster than the other two techniques. The symmetry-based approach presented the worst results regarding accuracy and execution time. Given these findings, it is possible to conclude that the phase congruency MSP detection approach is the appropriate choice in scenarios where precision is of paramount importance. However, in situations where accuracy is essential, but not crucial, and there are hundreds of images to be processed, the Hessian-based technique is indicated to promptly detect the plane.

Keywords: midsaggital plane, phase congruency, hessian matrix, symmetry information, magnetic resonance imaging

1 Introduction

The human brain has two hemispheres that can be separated by a midsaggital plane (MSP). Both hemispheres usually present a fair match of similar regions

and counterparts, and, therefore, analysis of these parts provides clinically relevant information in studies of neurodegenerative diseases, such as level of brain atrophy and asymmetry. The brain's longitudinal fissure (LF) is commonly used as a reference to guide manual delineation of the MSP, and although the separation surface is not exactly a plane surface, a planar approximation is enough for many applications [19].

Many techniques to automatically detect the MSP in magnetic resonance images (MRI) have been proposed in the literature over the years [8, 13, 18]. There are mainly two different classes of algorithms used to detect MSPs. The first one searches for the plane that maximizes a specific symmetry measure derived from information of the left and right brain hemispheres; the second one focus on estimating a plane from points that lie on the LF region.

To this day, there is no consensus regarding which algorithm works best for all scenarios. In this work, we used three different approaches for MSP detection. The first method [6] makes use of phase congruency (PC) information [10] to find a suiting plane, whereas the second approach is based on Hessian matrix and gradient features [3]. These two techniques fall under the category of algorithms that use the LF to estimate the MSP. For the sake of comparison, we used a publicly available symmetry-based MSP detection algorithm from Rupert *et al.* [19] to serve as a reference.

We applied all three algorithms to two databases, one with simulated [1] and the other with clinical images [2]. The purpose of using a simulated database was to generate different levels of noise and intensity inhomogeneity to assess how well the techniques would perform in adverse scenarios. Similarly, the clinical database was comprised of images from subjects with multiple sclerosis (MS). Since MS lesions have no well-defined shape, size or location, these images presented a certain degree of asymmetry, thus allowing us to verify if that would pose a challenge to MSP detection.

To quantitatively assess the algorithms, we used two well-known metrics for MSP detection: angle difference and average Z-distance. The first is used to calculate the difference in angle between a ground truth and an automatically detected MSP. Though widely used, this metric has the disadvantage of not taking into account specific situations such as when both planes are parallel but translated apart from each other. To circumvent this issue, we used the average Z-distance proposed in[19], which calculates the MSP estimation error as a function of the distances between corresponding voxels of each plane. These two metrics combined provide an excellent way to analyze any given MSP detection and compare one algorithm to another.

Our results show the PC and Hessian-based approaches outperformed the reference algorithm on angle difference and average Z-distance on both databases. The PC technique was the most accurate of all three algorithms, while the Hessian algorithm was the fastest. We also observed the symmetry-based approach fail on some images of the simulated database.

This paper is divided as follows: in Section 2 we describe our methodology, including the PC-, Hessian- and symmetry-based approaches as well as the

databases and metrics used in this work. We present and discuss our results in Section 3 and also provide images used to visually assess the algorithms' outputs. Our final considerations are made in Section 4.

2 Material and methods

In this section we elaborate on the PC, Hessian matrix and symmetry techniques and how they were used for the purpose of MSP detection. We also describe the databases and metrics used in this paper.

2.1 Phase congruency

The PC technique is an intensity and contrast invariant approach used to highlight relevant visual features in an image [17, 10]. The main idea of the PC technique is to search for image patterns having a similar order in the phase components of the Fourier transform. However, instead of working directly on the Fourier components, the PC technique, which can be proved to be proportional to the local energy of a signal, is implemented via convolution of the original image with a bank of complex band-pass filters in quadrature. For its implementation, Kovesi [10, 9, 11] proposed the use of log-Gabor filters, since they have zero mean components for arbitrarily large bandwidth that can be tuned to generate filters with minimal spatial extent. For efficiency, in this work the convolution of an input image with each complex filter in the bank was implemented in the frequency domain.

In this paper, we used the MSP-PC algorithm proposed by Ferrari *et al.*[6], where the MSP was estimated using the PC-technique proposed in [5] and a weighted least-squares algorithm [16] (Chapter 5) on voxels enhanced by a sheetness measure. The PC of an image is calculated, and then a sheetness measure is used on the PC output to enhance voxels that are likely to be located in the LF region. This measure is calculated based on the eigenvalues of a second-order moments local matrix of the 3D PC responses centered around a given spatial location $\boldsymbol{x}$, given by

$$M_{PC}(\boldsymbol{x}) = \begin{bmatrix} M_{200}(\boldsymbol{x}) & M_{110}(\boldsymbol{x}) & M_{101}(\boldsymbol{x}) \\ M_{110}(\boldsymbol{x}) & M_{020}(\boldsymbol{x}) & M_{011}(\boldsymbol{x}) \\ M_{101}(\boldsymbol{x}) & M_{011}(\boldsymbol{x}) & M_{002}(\boldsymbol{x}) \end{bmatrix}, \tag{1}$$

where $M_{pqr}(\boldsymbol{x})$ is the local moment of order $p+q+r$ calculated from the PC map.

Ratios can be devised to relate to certain geometric structures, assuming the eigenvalues of $M_{PC}(\boldsymbol{x})$ are sorted as $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$. These relations are shown in Table 1. Since a MSP is represented by a sheet-like structure, R_{sheet} is the ratio of interest in this case.

Two parameters related to the sheetness measure were set to different values from those proposed in [6]. Using the same notation as the authors, $\alpha = \delta = 0.10$ instead of $\alpha = \delta = 0.15$. This subtle change proved to yield better results for the

Table 1. Eigenvalue ratios and their relations with local geometric structures.

Eigenvalue ratios	Sheet	Tube	Blob	Noise
$R_{sheet}(\boldsymbol{x}) = \lambda_2/\lambda_3$	0	1	1	-
$R_{blob}(\boldsymbol{x}) = (2\lambda_3 - \lambda_2 - \lambda_1)/\lambda_3$	2	1	0	-
$R_{noise}(\boldsymbol{x}) = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}$	λ_3	$\lambda_3\sqrt{2}$	$\lambda_3\sqrt{3}$	0
$R_{tube}(\boldsymbol{x}) = (\lambda_3 - \lambda_2 + \lambda_1)/\lambda_3$	1	0	1	-

databases analyzed in this paper. Every other parameter related to the sheetness measure, PC and bank of filters remained the same. We recommend referring to [6] for an in-depth explanation of this technique.

2.2 Hessian matrix

A local Hessian matrix can be seen as a local shape descriptor which its eigenvalues carry important information about a structured object. Given a particular point $\boldsymbol{x}$ of an image I with dimension k, the local Hessian matrix is defined as

$$H_\sigma(I;\boldsymbol{x}) = \begin{bmatrix} I_{\sigma x_1^2} & I_{\sigma x_1 x_2} & \cdots & I_{\sigma x_1 x_k} \\ I_{\sigma x_2 x_1} & I_{\sigma x_2^2} & \cdots & I_{\sigma x_2 x_k} \\ \vdots & \vdots & \ddots & \vdots \\ I_{\sigma x_k x_1} & I_{\sigma x_k x_2} & \cdots & I_{\sigma x_k^2} \end{bmatrix}, \tag{2}$$

where σ in H_σ indicates a multiscale framework that is calculated by convolving the original image with Gaussian kernels $G(\boldsymbol{x}, \sigma)$. In this case, the second order derivatives in Equation 2 correspond to

$$I_{\sigma x_i^2} = I_0(\boldsymbol{x}) * \left(\sigma^{2\gamma} \frac{\partial^2}{\partial x_i^2} G(\boldsymbol{x};\sigma) \right), \tag{3}$$

$$I_{\sigma x_i x_j} = I_{\sigma x_j x_i} = I_0(\boldsymbol{x}) * \left(\sigma^{2\gamma} \frac{\partial^2}{\partial x_i x_j} G(\boldsymbol{x};\sigma) \right). \tag{4}$$

These equations are also called γ-parameterized normalized derivatives since the term γ is introduced to overcome the problem of increasingly smoothed responses [12]. A value of $\gamma = 1.25$ was used in this work, as suggested by Majer [15]. The factor 2 in 2γ follows the order of the derivative.

By examining the second-order variations in image intensities, a Hessian eigenvalue decomposition can be done to better understand local shape information. In this work, we followed the idea presented by Descoteaux *et al.*. [3], inspired by the works of [7, 14, 20]. By doing so, we were able to differentiate 3D tube-like, sheet-like and blob-like structures. In this sense, assuming our eigenvalues follow the ordering $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$, the same three ratios presented in Table 1 can be devised.

For sheet-like structures, the eigenvalues must be $\lambda_1 \approx \lambda_2 \approx 0$ and $|\lambda_3| \gg 0$, so measure R_{sheet} can easily differentiate sheet-like structures from others, while

R_{blob} reinforces sheetness features. Finally, the Frobenius norm, R_{noise}, helps to suppress random noise effects in the background.

In order to enhance the desired shaped structures, Descoteaux *et al.*. defined a sheetness measure, S, in which its maximum response over all scales σ corresponds to the acceptance level that a sheet-like structure is present in a particular image position and is given by

$$S(\boldsymbol{\lambda}) = \max_{\sigma \in \Sigma} S_{\sigma}(\boldsymbol{\lambda}), \tag{5}$$

where

$$S_{\sigma}(\lambda) = \begin{cases} 0, & \text{if } \lambda_3 < 0, \\ \exp\left(-\frac{R_{sheet}^2}{2\alpha^2}\right) \times & \\ \left(1 - \exp\left(-\frac{R_{blob}^2}{2\beta^2}\right)\right) \times & \\ \left(1 - \exp\left(-\frac{R_{noise}^2}{2c^2}\right)\right), & \text{otherwise.} \end{cases} \tag{6}$$

In Equation 6, $\Sigma = \{2.0, 4.0\}$ is a finite set of scales and α, β and c can be adjusted to control the sensitivity of the filter components. We performed a grid search over all possible combinations in range [0, 1] with steps of 0.05 to set $\alpha = 0.75$ and $\beta = 0.10$. Parameter c was calculated following the suggestion of Dzyubak and Ritman [4] as one-tenth of the maximum Laplacian image value.

An important observation must be made about the sign of λ_3 in Equation 6, since it dictates the intensity feature of the structure we are interested. In T1-weighted images, the LF is dark, so we use the equation as is. For T2- and PD-weighted images, however, the LF is bright, so in these cases, $S_{\sigma}(\lambda) = 0$ if $\lambda_3 > 0$.

As a result of the Hessian-based approach, voxel values closer to 1 indicate a higher probability of having sheet-like structures in a real-valued image. Similar to [6], plane estimation was performed by a weighted least-squares algorithm [16] in a coarse-to-fine manner applied to voxels enhance by the sheetness measure.

2.3 Symmetry analysis

In Ruppert *et al.* [19], a bilateral symmetry algorithm based on the maximization of symmetry measure was used to find the best plane that divides the brain into its hemispheres. The authors applied a 3D Sobel operator and thresholding to generate a binary-edge feature image, which was then used as input to the symmetry measure algorithm. Since evaluating all possible planes was unfeasible, the authors proposed a 3-stage coarse-to-fine multi-scale approach to reduce the search space. The plane with the highest symmetry score on a coarser scale was used as input to a finer scale, and then the region around the plane was analyzed to refine the MSP estimation.

Formally, let I^o be the binary image resulting from applying a threshold to the Sobel output and I^f its flipped copy with respect to a candidate plane. Then

the symmetry measure S is defined as

$$S = \frac{\sum_i^w \sum_j^h \sum_k^d I_{ijk}^o I_{ijk}^f}{\sqrt{\left(\sum_i^w \sum_j^h \sum_k^d I_{ijk}^o I_{ijk}^o\right)\left(\sum_i^w \sum_j^h \sum_k^d I_{ijk}^f I_{ijk}^f\right)}}, \tag{7}$$

where w, h, d are width, height and depth of the 3D image and I_{ijk}^f, similar to I_{ijk}^o, is the value in $\{0, 1\}$ of a voxel at a given coordinate (i, j, k).

They assessed their algorithm on MRI and computation tomography (CT) images using angle difference and average Z-distance and compared their results to other three approaches. Considering accuracy and execution time, the authors concluded their technique outperformed the other three algorithms.

2.4 Databases

The MSP algorithms investigated in this work were assessed using two different sets of MR images, described in details in this section.

BrainWeb We used 45 synthetic MR (T1-, T2- and PD-weighted) images from the publicly available McGill University BrainWeb MRI simulator [1]. We used images with five different noise levels (0%, 1%, 3%, 5% and 7%) and three intensity inhomogeneity percentages (0%, 20% and 40%), thus generating fifteen images for each MR weight. Every image had a voxel resolution of 1 mm^3 and dimensions $181 \times 217 \times 181$.

MS The MS database used in this paper came from the training data of the 2015 Longitudinal MS Lesion Segmentation Challenge [2]. This database was comprised of five MS subjects with a total number of 21 time-points and lesion volumes ranging from $2ml$ to $32ml$. Each time-point had T1-, T2-, PD-weighted and FLAIR images acquired on a 3T MR scanner. For this research, we used only T1-weighted images because of the high contrast and anatomical information such weight provides. Every T1-weighted image had a voxel resolution of 1 mm^3 and dimensions $181 \times 217 \times 181$.

2.5 Metrics

We used the angle difference (in degrees) and average Z-distance (in voxels) to assess how accurately the algorithms detected the MSP. These metrics are described in the following sections.

Angle difference The angle difference measures the angle between normal vectors of two given planes as

$$\alpha = \arccos\left(\frac{< \mathbf{u}, \mathbf{v} >}{\|\mathbf{u}\|.\|\mathbf{v}\|}\right) \times \frac{180}{\pi}, \tag{8}$$

where α is the angular difference in degrees, $\mathbf{u}$ and $\mathbf{v}$ are two given normal vectors, $<,>$ is the canonical inner product and $\|\|$ is the vector norm function. The closer α is to zero, the smaller is the angle between planes.

Average Z-distance The average Z-distance was proposed in [19] to measure the distance in voxels between two planes. This metric improves the quantitative analysis of the MSP detection, since the angle difference alone may be misleading in situations where both planes are parallel but translated from one another, for example.

To find this distance, the z coordinate of each plane is calculated using the plane equations and their x and y coordinates. The absolute differences between the z coordinates from the ground truth, and the estimated plane are summed up and divided by $dim(x) \times dim(y)$, where dim is the image dimension of a given axis, thus providing the average Z-distance in voxels between both planes. Formally,

$$Z\text{-distance} = \frac{\sum_{(x,y)} \left(|z_{coord}(GT) - z_{coord}(Aut)|\right)}{dim(x) \times dim(y)}, \tag{9}$$

where GT and Aut are the ground truth and automatically estimated MSPs, respectively.

3 Results and discussion

In this section we present the MSP detection results for both BrainWeb and MS databases and discuss the advantages and drawbacks of each technique.

3.1 BrainWeb database

The results for the PC, Hessian and symmetry approaches on the BrainWeb database are shown in Tables 2, 3 and 4.

Of the three techniques assessed in this paper, the symmetry approach was not able to correctly detect the MSP for some images in the BrainWeb database (indicated by *). These cases were not taken into account when calculating the average angle difference and Z-distance of the three techniques since they would artificially increase the average and standard deviation results of the symmetry approach.

The symmetry results shown in Tables 2, 3 and 4 indicate that this technique had a rather poor behaviour in situations where images were noisy and presented a high-intensity inhomogeneity level. Examples of MSPs detected on T2 image weights with 7% noise and 40% intensity inhomogeneity levels are shown in Figure 1. These results can be explained by the fact that this technique uses a bilateral symmetry measure and searches for a plane that maximizes it. Part of the algorithm consists of smoothing and applying a Sobel operator to the input image to detect edges. The output of the Sobel operator is then used to calculate the symmetry measure and find the best plane. Since we had some images with

Table 2. MSP detection results for PC-, Hessian- and symmetry-based algorithms on the BrainWeb database with intensity inhomogeneity = 0%. Rows in bold were not taken into account to calulate mean and standard deviation values.

	Phase Congruency		Hessian		Symmetry	
Image	Angle diff. (°)	Z-dist. (voxels)	Angle diff. (°)	Z-dist. (voxels)	Angle diff.(°)	Z-dist. (voxels)
PD 0%	0.196	0.156	0.719	1.013	0.318	1
PD 1%	0.185	0.147	0.719	1.012	0.318	1
PD 3%	0.290	0.238	0.016	0.013	0.318	1
PD 5%	0.703	0.558	0.016	0.013	0.318	1
PD 7%*	**1.194**	**1.134**	**0.017**	**0.014**	**90**	**0**
T1 0%	0.144	0.114	1.639	1.347	0.318	1
T1 1%	0.142	0.113	1.674	1.380	0.318	1
T1 3%	0.147	0.117	1.819	1.506	0.318	1
T1 5%	0.146	0.116	1.700	1.396	0.265	1
T1 7%	0.132	0.105	1.608	1.320	0.265	1
T2 0%	0.759	0.666	0.388	0.422	0.318	1
T2 1%	0.750	0.657	0.386	0.418	1.591	2.052
T2 3%	0.802	0.711	0.384	0.415	0.318	1
T2 5%*	**0.835**	**0.747**	**0.385**	**0.416**	**0**	**51.5**
T2 7%	0.867	0.784	0.383	0.413	0	0.5
Mean (std)	0.404 (0.310)	0.344 (0.278)	0.881 (0.696)	0.821 (0.553)	0.383 (0.373)	1.042 (0.333)

a high level of intensity inhomogeneity, we believe the level of smoothing used by this technique was not enough, leading the Sobel algorithm to detect noise as edges, which in turn caused the MSP detection to become rather erratic.

The PC approach had the smallest average Z-distance in both scenarios of intensity inhomogeneity and a very low angle difference on average. We can state that, on average, the PC-based algorithm was less than 0.5 degree and 0.5 voxels off compared to the ground truth. It is also possible to note that this technique was somewhat invariant to noise and intensity inhomogeneity given any image weight, which is strongly related to the frequency-domain aspect of PC [10]. But high accuracy and invariance come with a cost, since it took roughly 91 seconds to detect the MSP of each image in this database using phase congruency.

The Hessian-based algorithm was the least accurate approach on this database considering angle difference. However, it outperformed the symmetry technique on average Z-distance in all three intensity inhomogeneity scenarios. Since the Hessian algorithm is heavily based on image gradients, it is not as robust as phase congruency and is susceptible to highlight high-gradient areas that are not related to the LF. These areas interfere with the MSP detection and can cause the plane to be slightly dislocated compared to the ground truth. This drawback can be seen in T1-w results since this weight presents a high-contrast

Table 3. MSP detection results for PC-, Hessian- and symmetry-based algorithms on the BrainWeb database with intensity inhomogeneity = 20%. Rows in bold were not taken into account to calulate mean and standard deviation values.

	Phase Congruency		Hessian		Symmetry	
Image	Angle diff. (°)	Z-dist. (voxels)	Angle diff. (°)	Z-dist. (voxels)	Angle diff.(°)	Z-dist. (voxels)
PD 0%	0.183	0.145	0.016	0.013	0.636	1.5
PD 1%	0.158	0.125	0.016	0.013	0	1.5
PD 3%	0.181	0.144	0.017	0.013	0	1.5
PD 5%	0.275	0.223	0.017	0.014	0	1.5
PD 7%*	**1.266**	**1.197**	**1.142**	**1.016**	**90**	**0**
T1 0%	0.144	0.114	2.542	2.254	0.318	1
T1 1%	0.145	0.115	2.871	2.544	0.318	1
T1 3%	0.149	0.118	2.559	2.261	0.636	1.5
T1 5%	0.156	0.124	2.463	2.162	0.265	1
T1 7%	0157	0.125	0.981	0.782	0	0.5
T2 0%	0.711	0.615	0.456	0.475	0.318	1
T2 1%*	**0.685**	**0.586**	**0.456**	**0.474**	**0**	**51.5**
T2 3%*	**0.749**	**0.658**	**0.454**	**0.471**	**0**	**52.5**
T2 5%*	**0.813**	**0.727**	**0.452**	**0.469**	**89.681**	**10536.776**
T2 7%	0.849	0.759	0.473	0.484	1.273	1.067
Mean (std)	0.282 (0.250)	0.237 (0.226)	1.128 (1.212)	1.001 (1.066)	0.342 (0.389)	1.187 (0.334)

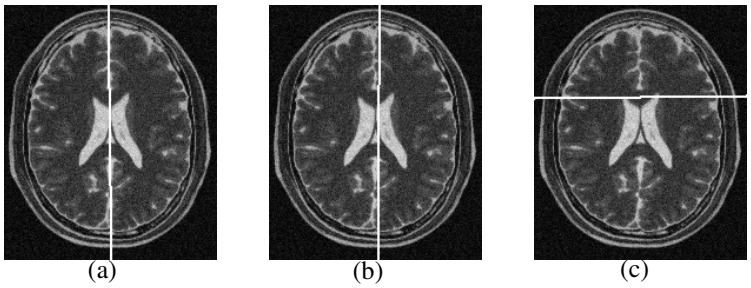


Fig. 1. BrainWeb MSP comparison (axial view) of the three approaches assessed in this paper on T2-weighted images with 7% noise and 40% intensity inhomogeneity. (a) PC-based MSP detection, (b) Hessian-based MSP detection and (c) Symmetry-based MSP detection.

Table 4. MSP detection results for PC-, Hessian- and symmetry-based algorithms on the BrainWeb database with intensity inhomogeneity = 40%. Rows in bold were not taken into account to calulate mean and standard deviation values.

	Phase Congruency		Hessian		Symmetry	
Image	Angle diff. (°)	Z-dist. (voxels)	Angle diff. (°)	Z-dist. (voxels)	Angle diff.(°)	Z-dist. (voxels)
PD 0%	0.168	0.133	0.017	0.014	0.636	1.5
PD 1%	0.188	0.149	0.017	0.014	0	1.5
PD 3%	0.168	0.134	1.064	0.907	0	1.5
PD 5%	0.391	0.313	1.070	0.913	0	1.5
PD 7%*	**1.074**	**1.017**	**0.017**	**0.014**	**90**	**0**
T1 0%	0.144	0.114	0.952	0.756	0.318	1
T1 1%	0.145	0.115	0.949	0.753	0.318	1
T1 3%	0.148	0.118	0.932	0.740	0.265	1
T1 5%	0.16	0.127	0.952	0.756	0.265	1
T1 7%	0.134	0.107	0.292	0.307	0	0.5
T2 0%	0.593	0.499	0.646	0.609	0.318	1
T2 1%*	**0.619**	**0.523**	**0.646**	**0.610**	**0**	**44.5**
T2 3%*	**0.673**	**0.576**	**0.654**	**0.615**	**89.364**	**7504.408**
T2 5%*	**0.779**	**0.685**	**0.676**	**0.627**	**90**	**0**
T2 7%*	**0.797**	**1.067**	**0.668**	**0.599**	**89.363**	**5255.808**
Mean (std)	0.223 (0.149)	0.180 (0.127)	0.689 (0.423)	0.577 (0.341)	0.212 (0.210)	1.150 (0.337)

between tissues. On the other hand, the Hessian technique was faster than the other two algorithms by a 10-fold, since the MSP detection took approximately 9 seconds using this approach. Qualitative analysis indicates the automatically detected MSPs correctly divided the brain hemispheres into two, though the planes did not precisely halve the LF. Examples of MSP detections on T1-w images using the Hessian algorithm are shown in Figure 2.

It is worth pointing out that both PC and Hessian approaches detected the MSPs regardless of noise and intensity inhomogeneity. However, we cannot state this for the symmetry approach, since the number of failures increased as the intensity inhomogeneity level increased. We understand that certain levels of noise and inhomogeneity - for instance, 7% and 40%, respectively - are no longer encountered in research or clinical work, but we used them to check the algorithms' robustness in extreme situations.

3.2 MS database

The results for the PC, Hessian and symmetry approaches on the MS database are shown in Tables 5, 6 and 7.

We can see that the PC approach had the best results (both for angle differences and Z-distance), followed by Hessian matrix and symmetry information.

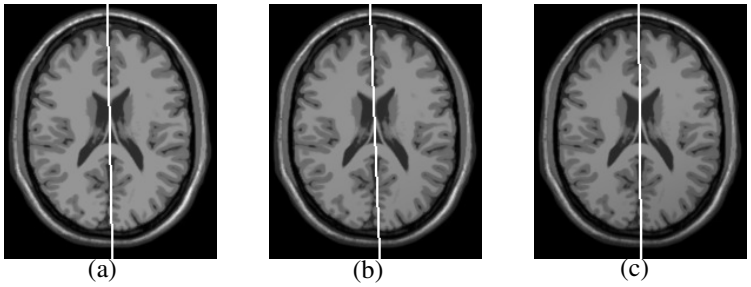


Fig. 2. BrainWeb MSP comparison (axial view) of MSPs detected by the Hessian approach on T1 images with 0% noise and (a) 0% and (b) 20% intensity inhomogeneity levels. Note that the hemispheres are correctly split, but the planes do not precisely halve the LF.

It is worth remembering that though this database had T1-, T2-, PD-weighted and FLAIR images, we used only T1 images to detect the MSP. Compared to PC, the Hessian matrix and symmetry techniques had approximately 36% and 42% difference in the angle measurement and 28% and 36% difference on the Z-distance metric, respectively. But all three techniques, on average, were less than 1.3 degrees and 1.2 voxels off compared to the ground truth. Based on the results shown in Tables 5, 6 and 7 indicate, all three techniques had a significant average performance regardless of lesion volume. Qualitatively, it was not possible to distinguish one method from another for any given detected MSP in this database. The average execution time for the PC, Hessian-based and symmetry-based algorithms on this database were 92 seconds, 9 seconds and 293 seconds, respectively, on an Intel Core i7 3.2 GHz with 16GB of RAM. Note that these times were very similar to those mentioned in Section 3.1. Examples of planes detected by the three techniques compared to the manually delineated planes in an MS image with high lesion load is shown in Figure 3.

Table 5. PC-based MSP detection results for the MS database.

Subject number	Time points	Avg. lesion load (ml)	Avg. angle diff. ($^\circ$)	Avg. Z-distance (voxels)
1	4	17.877	0.734	0.648
2	4	31.167	0.512	0.712
3	5	6.616	1.041	0.831
4	4	2.711	0.233	0.809
5	4	4.257	0.923	0.794
			Mean (std)	Mean (std)
			0.705 (0.339)	0.753 (0.125)

Table 6. Hessian-based MSP detection results for the MS database.

Subject number	Time points	Avg. lesion load (ml)	Avg. angle diff. ($^\circ$)	Avg. Z-distance (voxels)
1	4	17.877	1.341	1.190
2	4	31.167	1.192	1.076
3	5	6.616	1.105	0.897
4	4	2.711	0.187	1.025
5	4	4.257	1.643	1.328
			Mean (std)	Mean (std)
			1.094 (0.579)	1.044 (0.321)

It is important to note that visual assessment of the planes detected by all three approaches in the MS database indicated the brain hemispheres were correctly separated. However, some points need to be highlighted. For instance, compared to the Hessian-based algorithm, PC requires a bank of filters to work, and while it is true that this bank is created only once and then used on every image, the convolution of the image with all filters in the bank, performed in the frequency domain, is time-consuming. Besides, there is also an overhead for generating the bank of filters. In this work, we designed it to work best with image dimensions $181 \times 217 \times 181$, which were the image dimensions for both BrainWeb and MS databases. Another significant highlight is that we observed the Hessian-based approach runs 10-fold faster than PC. The same goes for the symmetry approach; similar to the Hessian matrix, the symmetry technique does not require a bank of filters, but it is approximately 32 times slower than the Hessian-based MSP detection. This result is an enormous advantage when working with hundreds of images in real-life scenarios such as clinical trials.

Finally, to visually assess the robustness of the three techniques even further, we applied them to MR FLAIR images either with glioma or post-brain tumor resection to analyze how well they would perform in a strong asymmetric scenario. The results are shown in Figures 4 and 5.

Table 7. Symmetry-based MSP detection results for the MS database.

Subject number	Time points	Avg. lesion load (ml)	Avg. angle diff. ($^\circ$)	Avg. Z-distance (voxels)
1	4	17.877	2.339	2.078
2	4	31.167	0.662	0.626
3	5	6.616	0.791	0.792
4	4	2.711	0.270	1.035
5	4	4.257	2.115	1.679
			Mean (std)	Mean (std)
			1.214 (0.903)	1.172 (0.644)

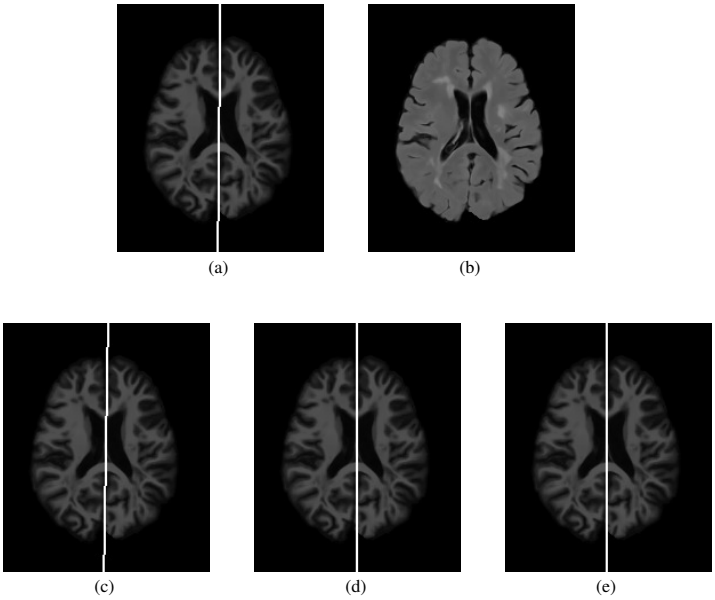


Fig. 3. Comparison between MSP detection techniques on an image with an average lesion load of $31.167ml$. (a) Ground truth, (b) FLAIR image where lesions are more easily identified, (c) PC-based MSP detection, (d) Hessian-based MSP detection, (e) Symmetry-based MSP detection.

In Figure 4, the MR image came from a subject with glioma. In Figure 5, the subject had undergone a medical removal procedure which left asymmetric holes in the brain.

As it can be seen in Figures 4 and 5, the PC technique outperformed the other two approaches; however, the same observation made earlier regarding execution time also holds true for these strong asymmetric cases. So in a scenario where accuracy is of paramount importance, the PC algorithm would be the best choice. On the other hand, if time is of the essence and accuracy is essential, but not crucial, then the Hessian-based MSP detection would be a good alternative.

4 Conclusions

This paper presented and compared three automatic MSP detection techniques applied to simulated and clinical databases. One of the techniques was a publicly available MSP detection algorithm based on symmetry analysis, while the other two were based on phase congruency and Hessian matrix, respectively, to find the brain's LF and estimate a plane from it. We assessed the algorithms' performance using the angle difference and average Z-distance metrics.

Our results show the PC approach was the most accurate technique on both databases, followed by the Hessian- and symmetry-based algorithms. The accu-

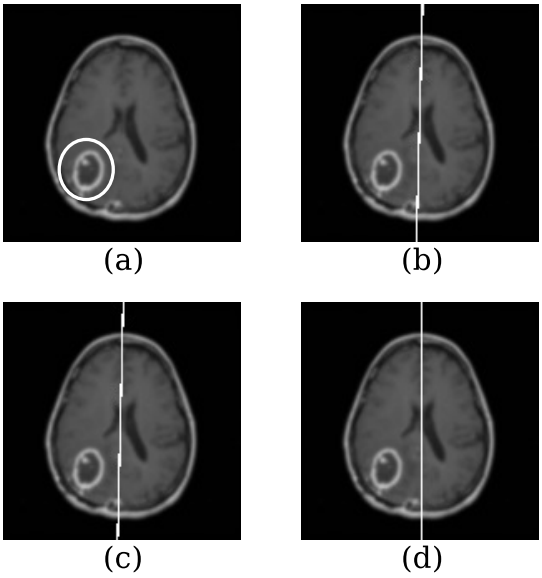


Fig. 4. Comparison between MSP detection techniques on an image with glioma. (a) Pathology indication, (b) PC-based MSP detection, (c) Hessian-based MSP detection, (d) Symmetry-based MSP detection.

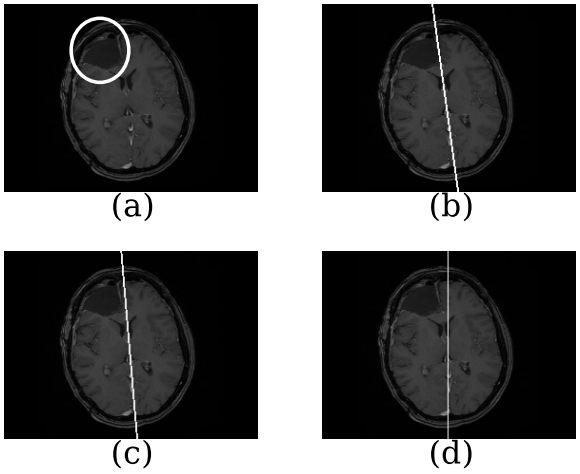


Fig. 5. Comparison between MSP detection techniques on a post-operation image with asymmetric holes. (a) Asymmetry indication, (b) PC-based MSP detection, (c) Hessian-based MSP detection, (d) Symmetry-based MSP detection.

racy of PC was strongly related to its frequency-domain aspect, which makes it contrast-invariant and more robust to noise and intensity variations.

The Hessian approach, though not as accurate as PC, proved to be the fastest of all techniques by a 10-fold. This fast execution time feature brings enormous advantage in scenarios where hundreds of images need to be processed such as in clinical trials. Visual assessment of MSPs detected by the Hessian algorithm indicated the brain hemispheres were correctly split into two, but the LFs were not necessarily halved.

We observed the symmetry-based algorithm was affected by high levels of noise and intensity inhomogeneity, which led to incorrect plane estimations. On the other hand, this approach was able to correctly estimate MSPs for the clinical database and images with glioma and tumor resection. However, it had the worst execution time compared to the other two approaches.

As mentioned in Section 1, there is no consensus regarding which MSP detection approach in MRI works best for all scenarios. Given the three techniques analyzed in this paper, we can conclude that if accuracy is of paramount importance, then the PC approach provides an excellent MSP estimation. But a trade-off between accuracy and execution time points to the Hessian-based algorithm as a better alternative to divide the brain hemispheres into two.

References

1. B. Aubert-Broche, M. Griffin, G.B. Pike, A.C. Evans, and D.L. Collins. Twenty new digital brain phantoms for creation of validation image data bases. *IEEE Transactions on Medical Imaging*, 25(11):1410–1416, 2006.
2. A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J Nguyen, F. Prados, C.H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C.A. Wheeler-Kingshott, Ourselin S., L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S.K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L.O. Iheme, D. Unay, S. Jain, D.M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.L. Bazin, P.A. Calabresi, C.M. Crainiceanu, L.M. Ellingsen, D.S. Reich, J.L. Prince, and D.L. Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *Neuroimage*, 148(1):77–102, 2017.
3. M. Descoteaux, M. Audette, K. Chinzei, and K. Siddiqi. Bone enhancement filtering: application to sinus bone segmentation and simulation of pituitary surgery. *Computer Aided Surgery*, 11(5):247–255, 2006.
4. O.P. Dzyubak and E.L. Ritman. Automation of hessian-based tubularity measure response function in 3D biomedical images. *International Journal of Biomedical Imaging*, 2011(920401):1–16, 2011.
5. R.J. Ferrari, S. Allaire, A. Hope, J. Kim, D. Jaffray, and V. Pekar. Detection of point landmarks in 3D medical images via phase congruency model. *Journal of the Brazilian Computer Society*, 17(2):117–132, 2011.
6. R.J. Ferrari, C.H.V. Pinto, and C.A.F. Moreira. Detection of the midsagittal plane in mr images using a sheetness measure from eigenanalysis of local 3D phase congruency responses. In *IEEE International Conference on Image Processing*, pages 2335–2339, Phoenix, AZ, USA, Sep 2016.

7. A.F. Frangi, W.J. Niessen, K.L. Vincken, and M.A. Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, volume 1496, pages 130–137, Cambridge, MA, USA, 1998. Springer-Verlag.
8. S.A. Jayasuriya, A.W. Liew, and N. Law. Symmetry plane detection in brain image analysis: A survey. *Current Medical Imaging Reviews*, 9(3):230–247, 2013.
9. P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):2–26, 1999.
10. P. Kovesi. Phase congruency: A low-level image invariant. *Psychological Research*, 64(2):136–148, 2000.
11. P. Kovesi. Phase congruency detects corners and edges. In *The Australian Pattern Recognition Society Conference: VIIth DICTA*, pages 309–318, Sydney, Australia, 2003.
12. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
13. Y. Liu, R.T. Collins, and W.E. Rothfus. Robust midsaggital plane extraction from normal and pathological 3-D neuroradiology images. *IEEE Transactions on Medical Imaging*, 20(3):175–192, 2001.
14. C. Lorenz, I.-C. Carlsen, T. M. Buzug, C. Fassnacht, and J. Weese. Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2D and 3D medical images. In *Proceedings of the First Joint Conference on Computer Vision, Virtual Reality and Robotics in Medicine and Medial Robotics and Computer-Assisted Surgery*, pages 233–242, London, UK, 1997. Springer-Verlag.
15. P. Majer. The influence of the gamma-parameter on feature detection with automatic scale selection. In *Proceedings of the Third International Conference on Scale-Space and Morphology in Computer Vision*, pages 245–254, Vancouver, BC, Canada, 2001. Springer.
16. D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression Analysis*. Series in Probability and Statistics. John Wiley & Sons, NJ, USA, fifth edition, 2012.
17. M.C. Morrone and D.C. Burr. Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London*, B(235):221–245, 1988.
18. S. Prima, Ourselin S., and N. Ayache. Computation of the mid-saggital plane in 3-D brain images. *IEEE Transactions on Medical Imaging*, 21(2):122–138, 2002.
19. G. Ruppert, L. Teverovskiy, C. Yu, A. Falcão, and Y. Liu. A new symmetry-based method for mid-sagittal plane extraction in neuroimages. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 285–288, Chicago, IL, USA, 2011.
20. Y. Sato, S. Nakajima, N. Shiraga, H. Atsumi, S. Yoshida, T. Koller, G. Gerig, and R. Kikinis. Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. *Medical Image Analysis*, 2(2):143–168, 1998.

Biomedical Signal Processing and Control 49 (2019) 338–348

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Multiple sclerosis lesion enhancement and white matter region estimation using hyperintensities in FLAIR images

Paulo G.L. Freire*, Ricardo J. Ferrari

Departamento de Computação, Universidade Federal de São Carlos, Rod. Washington Luis, Km 235, 13565-905 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 30 July 2018
Accepted 11 December 2018
Available online 26 December 2018

Keywords:
Multiple sclerosis
Lesion enhancement
White matter
Magnetic resonance image
Hyperintensities

ABSTRACT

FLAIR images are usually acquired in the context of multiple sclerosis (MS) because lesions appear hyperintense in this particular image weight. However, their intensity profile overlaps with white matter (WM) and gray matter (GM) tissues, posing difficulties for accurate segmentation. We propose a lesion enhancement technique to dim down WM and GM regions and highlight hyperintensities by creating a hyperintensity probability map. A byproduct of our proposal is the estimation of a mask encompassing WM and MS lesions. We observed that the lesion intensity profile in FLAIR was 25% and 19% brighter than WM and GM, respectively; comparatively, the same profile in our enhanced images was 444% and 264% brighter. On the mask estimation, we achieved a sensitivity of 99% and specificity of 98%. The results indicate significant improvement in the intensity distinction between lesions and tissues and can aid both experts and automated techniques in segmentation tasks.

1. Introduction

Multiple sclerosis (MS) is a demyelinating disease that attacks the central nervous system (CNS) and affects more than 2 million people worldwide [2]. It destroys neurons' myelin sheaths, causing many effects on one's body such as dizziness, confusion, memory problems and numbness of arms and legs [25]. The cause of MS is still unknown, and the disease itself has a devastating effect both for individuals and society. Since the onset of MS is typically around age 30, it affects subjects at the peak of their productivity in life [24]. In this context, it is essential to provide neurologists and radiologists with tools to quickly identify the disease and prescribe treatments to help patients lead a normal life.

Magnetic resonance imaging (MRI) is commonly used in diagnosis and follow-up of MS due to its high contrast between soft tissues [7]. One widely used imaging protocol of MRI is the Fluid Attenuated Inversion Recovery (FLAIR), which, as the name suggests, attenuates the effect of fluids, mainly from the cerebral spinal fluid (CSF) region. FLAIR images are important in the MS context because MS lesions appear hyperintense in this particular image weight, thus making it easier for physicians to identify them [12].

Though MS lesions present a hyperintense profile in FLAIR images, their intensity range varies significantly between different patients and between various time points of the same subject [11]. As a result, both manual and automatic segmentation can be affected and undermine the accuracy of MS lesion detection, since they can be mistaken for other brain tissues – namely, white matter (WM) and gray matter (GM). And while it is true that gray level intensity is not the only characteristic that helps experts and algorithms differ MS lesions from other tissue classes, a sharper distinction between hyperintensities and normal tissues would provide extra leverage to separate each class more accurately.

Proper distinction between brain tissues and abnormalities, such as lesions, would be of great help for experts and automatic segmentation techniques alike. In [17], the authors conducted a comprehensive analysis of the importance of intensity normalization and its effect on MS lesion segmentation. They showed that applying the intensity normalization technique proposed by [14] to 21 images from subjects with MS increased the Dice Similarity Coefficient (DSC) [8] of lesion segmentation on automatic supervised approaches. However, a scatter plot analysis showed that despite normalization, lesions still had a significant intensity overlap with WM and GM tissues.

In [19], the authors proposed an algorithm to increase automatic lesion and brain tissue segmentation robustness by estimating a spatially global within-the-subject intensity distribution and a spatially local intensity distribution derived from a healthy reference population. This approach tried to circumvent the overlap between the whole brain signal intensity distribution of lesions and healthy tissue. A scatter plot analysis showed that local intensities from

* Corresponding author.
URL: http://www.bipgroup.dc.ufscar.br (R.J. Ferrari).

https://doi.org/10.1016/j.bspc.2018.12.021
1746-8094/

a reference population offered a better lesion intensity separation than by either global or local intensity distributions derived from a patient with MS. Though the authors' proposal performed better than six other segmentation techniques on a 31-subject database, it was still dependent on a healthy reference population and, consequently, on careful registration and intensity normalization steps across the healthy and lesion image sets.

Accordingly, we propose a technique to enhance hyperintensities in FLAIR images to better distinguish MS lesions from WM and GM. We built on the works of [16,15] to automatically generate an image that shows the probability of each voxel being a hyperintense one, herein called the hyperintensity map. The main advantage of this technique is that it requires only FLAIR images and enhances MS lesions so their intensity profile is much brighter than WM and GM compared with their profiles in FLAIR itself.

Regarding the effects that brain abnormalities, such as lesions, have on brain tissue segmentation, some works in the literature have explored the issue. In [1], the authors showed that segmentation-based methods for brain volume measurement suffer in the presence of lesions since they interfere with GM and WM depending on lesion size and intensity. To overcome this problem, they proposed filling lesions with intensities matching surrounding normal-appearing WM. Their approach helped reduce the impact lesions had on tissue segmentation, especially regarding GM, and improved the accuracy of tissue classification and brain volume measurement. However, the filling algorithm depends on having the lesion ground truths at hand and, as noted by the authors themselves, their approach can overestimate lesion holes depending on where they are located and the intensity variation in the surrounding area.

Similarly, in [21], the authors stated that the accuracy of automatic tissue segmentation methods is affected by the presence of MS lesions during the tissue segmentation process. They applied six well-known segmentation techniques to 30 T1-weighted images from subjects with MS and verified that GM volume was overestimated by all methods when lesion volume increased. This overestimation persisted even when masking out or relabeling lesions during segmentation. This particular finding was significant because it provided evidence to show that tissue atrophy measurements are likely to be altered when the subject's lesion load is high.

In [22], the authors conducted a study to determine the effect ground truth annotations had on the assessment of automatic brain tissue segmentation accuracy. More specifically, they applied ten different brain tissue segmentation methods to the Internet Brain Segmentation Repository (IBSR),[1] since this dataset considered Sulcal Cerebrospinal Fluid (SCSF) voxels as gray matter. Though this dataset comprised only images from healthy subjects, the authors were able to observe that the performance and accuracy of the methods on IBSR images varied significantly when not considering SCSF voxels. This finding indicates that not only brain abnormalities, such as lesions, have the potential to interfere with segmentation techniques, but labeling is also a concern and can lead to a misguided analysis of accuracy.

Finally, in [23] the authors proposed an automated T1-weighted/FLAIR tissue segmentation approach designed to deal with images from subjects with WM lesions. They suggested a partial volume tissue segmentation with WM outlier rejection and filling, along with intensity, probabilistic and morphological prior maps, to segment brain tissues. This approach required no manual annotations of lesions, which is an advantage compared with other works that use lesion filling or masking based on expert ground truths. The authors applied their algorithm to two databases. One of them comprised only images from subjects with MS, and the authors' proposal achieved competitive results compared with five other segmentation techniques. However, the MS database was not publicly available, thus making it difficult to directly compare their results with those of other works in the literature.

To this end, a byproduct of our technique is the estimation of a white matter mask based on the hyperintensity map. This approach is relevant because, as previously noted, MS lesions interfere with brain tissue segmentation due to similarities in their intensity profiles. Our white matter mask estimation relies on the hyperintensity map to fill lesion holes that were left out during an automatic segmentation process. Our technique requires no manual annotations, only making use of FLAIR images and probabilistic anatomical atlases to obtain such estimation. To verify its accuracy, we extracted DSC and the percentage of MS lesions that were included in the WM mask during the process to confirm how well our approach performed in obtaining a reasonable WM region estimate.

This paper is divided as follows. In Section 2, we describe our methodology and the database we used to apply our technique on; we present our results in Section 3 and discuss them in Section 4. Finally, we present our final considerations in Section 5 and indicate our future works.

2. Materials and methods

In this section, we describe the databases we used, the pipeline for enhancing multiple sclerosis lesions in FLAIR images and the algorithm for estimating the white matter region mask.

2.1. Databases

2.1.1. Clinical images

We used the training dataset of the Longitudinal MS Lesion Segmentation Challenge[2] made available during the 2015 International Symposium on Biomedical Imaging [5]. This dataset comprised images of five patients, one male and four females, with a total of 21 time-points. The mean age of the patients was 43.5 years and the mean time between follow-up scans was one year. Eligibility criteria and more details on the patients and how they were chosen are described in [5].

Each scan was imaged and pre-processed in the same manner, with data acquired on a 3 Tesla MRI scanner (Philips Medical Systems, Best, The Netherlands). The imaging sequences were adjusted to produce T1-weighted, T2-weighted, proton density (PD) and FLAIR images.

Each subject underwent the following pre-processing: the baseline (first time-point) magnetization prepared rapid gradient echo (MPRAGE) was inhomogeneity-corrected using N4 [20], skull-stripped [6,4] and dura stripped [18], followed by a second N4 inhomogeneity correction and rigid registration to a 1 mm isotropic MNI template. Since the baseline MPRAGE was in MNI space, it was used as a target for the remaining images, which included the baseline T2-w, PD-w, and FLAIR, as well as the scans from each of the follow-up time-points. These images were then N4 corrected and rigidly registered to the 1 mm isotropic baseline MPRAGE in MNI space. In the end, image dimensions were $181 \times 217 \times 181$.

It is important to note that the training dataset also included manual MS lesion delineations by two experts for each time-point. More details about time-points and average MS lesion volume for each subject are summarized in Table 1.

[1] https://www.nitrc.org/projects/ibsr.

[2] http://iacl.ece.jhu.edu/index.php/MSChallenge/data.

Table 1
Number of time-points and average lesion volume for each patient.

	Number of time-points	Mean lesion volume (in ml) Expert 1	Mean lesion volume (in ml) Expert 2
Patient 1	4	16.67	19.07
Patient 2	4	30.52	31.80
Patient 3	5	5.40	7.81
Patient 4	4	2.17	3.23
Patient 5	4	4.55	3.96

2.1.2. Probabilistic anatomical atlases

To estimate white matter regions and identify gray matter clusters, we used the probabilistic atlases from the ICBM project [10]. The atlases spatial resolution was $1 \times 1 \times 1$ mm and their initial dimensions were $256 \times 256 \times 256$. However, they were registered to each clinical time-point to provide accurate spatial information. A T1-weighted image initially registered to both white matter and gray matter atlases was used as a moving image, while T1-weighted images from each time-point were used as reference images. The registration took place using the NiftyReg tool [13] with free-form B-Spline deformation model and multi-resolution approach for non-rigid registration. The transformation was then applied to the atlases, thus making them aligned to each time-point and with dimensions $181 \times 217 \times 181$.

2.2. Metrics

A total of five metrics were used to assess the intensity profile distinction between lesions and other brain tissues and to compare the estimated white matter mask with the ground truth.

The intensity profile distinction (IPD) is calculated as

$$IPD = \left(\frac{\text{average(lesions)}}{\text{average(tissue)}} - 1 \right) \times 100, \tag{1}$$

in which we simply divide the average intensity of lesions by the average intensity of a tissue of interest (white matter or gray matter) in a particular image and then scale it in terms of percentage. Thus, we can determine how much brighter, percent-wise, the lesion cluster is compared with other tissues.

In the assessment of the estimated white matter mask, we used four different metrics. The Dice Similarity Coefficient (DSC) [8] was used to determine the overall overlap between our estimation and the ground truth. The DSC is defined as

$$DSC = \frac{2 \times TP}{FP + FN + 2 \times TP}, \tag{2}$$

where *TP*, *FP*, and *FN* are the true positives, false positives and false negatives. DSC values fall in the interval [0, 1], and the closer they are to 1, the better.

We also used the well-known sensitivity and specificity metrics to measure the accuracy of our estimation. Sensitivity (SS) is defined as

$$SS = \frac{TP}{(TP + FN)} \tag{3}$$

and in the context of this paper measures the proportion of WM and lesion voxels that were correctly identified as such during estimation.

Specificity (SP) is defined as

$$SP = \frac{TN}{(FP + TN)}, \tag{4}$$

where *TN* represents the number of true negatives. In the context of this paper, *SP* measures the proportion of non-WM and non-lesion voxels that were correctly identified as such. Both SP and SS fall in the interval [0, 1] and similar to DSC, the closer they are to 1, the better.

Another metric we used to assess the white matter mask estimation was the lesion intersection (LI), defined as

$$LI = \frac{|\text{Lesion}_{GT} \cap \text{Mask}_{estim}|}{|\text{Lesion}_{GT}|} \times 100. \tag{5}$$

Similar to IPD, LI is also calculated in terms of percentage and provides a quantitative tool to analyze the lesion load that was kept during the white matter mask estimation.

2.3. Pre-processing

The pre-processing stage used in this work was comprised of three steps: noise reduction, intensity normalization and intermediate enhancement.

The noise reduction step was performed using the non-local means approach [3] with $\sigma = 15$. Reducing image noise is important because it helps eliminate part of intensity variations within the image, creating a smoother profile for each brain tissue. Given the nature of our lesion enhancement technique, which is heavily based on gray level intensities, noise reduction is relevant to mitigate effects inherent in the image acquisition procedure and thus improve the enhancement of MS lesions while dimming out other brain tissues.

Intensity normalization was done in order to assure that every image had the same intensity range. We applied the normalization proposed in [16], which rescales intensities as $v' = \frac{v}{\mu + 3\sigma}$, where v' is the new value of a given voxel, v is the voxel value in FLAIR and μ and σ represent the mean and standard deviation of the whole brain, respectively.

An edge detection step using Sobel [9] is required as input to generate the intermediate enhanced image with increased contrast between MS lesions and their surroundings, namely white matter tissue. Following the proposal in [16], this intermediate image was created as follows. Let $s = \{x, y, z\}$ be a particular spatial location, I_s the FLAIR intensity at s and g_s the gradient in the Sobel image at s. The edge and intensity information are combined as

$$h(i) = \frac{1}{N} \sum_{s \in \{s | I_s = i\}} Prob(g \leq g_s), \tag{6}$$

where N is the total number of voxels with intensity i. In other words, Eq. (6) goes over every FLAIR image intensity i and, given the probability density function (PDF) of the Sobel image, sums up the histogram bins that have a smaller frequency than g_s and then normalizes them.

After calculating $h(i)$, we compute the cumulative distribution function (CDF) of h as proposed in [16]:

$$q(i) = \sum_{k=1}^{i} h(k). \tag{7}$$

In the end, each $q(i)$ is used to replace each intensity i. An example of such intermediate image is shown in Fig. 1.

2.4. Hyperintensity probability map

The hyperintensity probability map is calculated based on the intermediate image generated during the pre-processing stage detailed in Section 2.3. We devised an algorithm that automatically generates such map and does not depend on parameters that must be set by experimental observations, as opposed to [16,15].

The central principle behind this map is to compare each voxel neighborhood intensity with patches across different points in the image. The more times the voxel's neighborhood mean intensity is

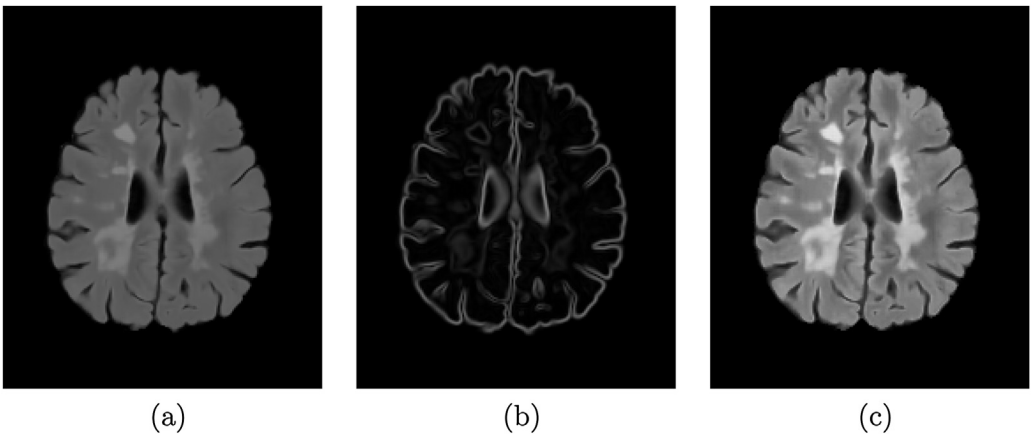


Fig. 1. Intermediate image generation. (a) FLAIR image, (b) Sobel image, and (c) intermediate image.

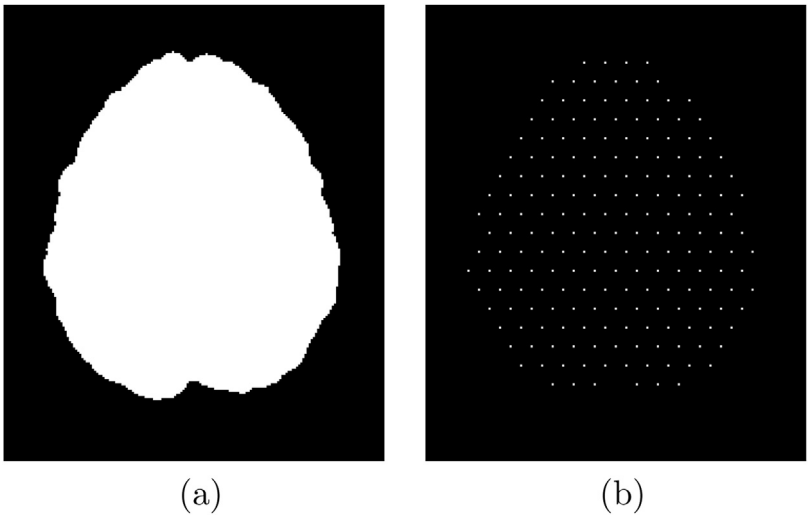


Fig. 2. Brain mask (a) and cropped point net (b). Each patch of this particular slice is centered around one point in (b).

higher than the patches', then the more likely it is for that particular voxel to stand out and more likely for it to have a high hyperintense probability.

The first step to create the map is to define where each patch will be centered. To do that, we define a point net for each slice following the algorithm proposed in [15], which uses the combination of sines and cosines to evenly distribute points across a slice. Let $p = \{x, y, z\}$ be the coordinates of a candidate point. We then create new points $p' = \{x', y', z\}$ with

$$\begin{aligned} x' &= x + r\cos\theta \\ y' &= y + r\sin\theta \end{aligned}, \tag{8}$$

where θ is the angle and $r = 10$ is the radius. We set θ to zero and increase it by 60 degrees six times to complete a whole circumference. The six newly defined points become candidate points, and the process is repeated until no new point is found.

After defining such points, we crop the net using a brain mask to only keep points that are inside our ROI. This procedure is done by simply purging points outside the brain mask. An example of a final point net set P for a particular slice is shown in Fig. 2.

Now, let μ and σ be the mean and standard deviation of the whole intermediate image within the brain mask. Then, for each voxel, we calculate its neighborhood mean intensity as

$$\mu_v = \frac{1}{N_v}\sum_{k=1}^{N_v} i_k, \tag{9}$$

where μ_v is the mean neighborhood intensity of voxel v, N_v is the number of neighbors of v and i_k is the intensity of neighbor k. The neighborhood size was defined as $3 \times 3 \times 3$ in order to maintain a good trade-off between sharpness and smoothness. The same rationale is used for the patches: the mean intensity is calculated as in Eq. (9), thus creating μ_p for each patch.

Finally, we create a score S_v as

$$S_v = \frac{1}{|P|}\sum_{p \in P} \delta(\mu_v, \mu_p), \tag{10}$$

where $|P|$ is the cardinality of the patch set and

$$\delta(\mu_v, \mu_p) = \begin{cases} 1, & \text{if } \mu_v - \mu_p \geq \sigma \\ 0, & \text{otherwise} \end{cases}. \tag{11}$$

In other words, if the difference of intensity between a candidate voxel neighborhood and a patch is greater than or equal to the standard deviation of the whole image, then it is a hit. Otherwise, it is a

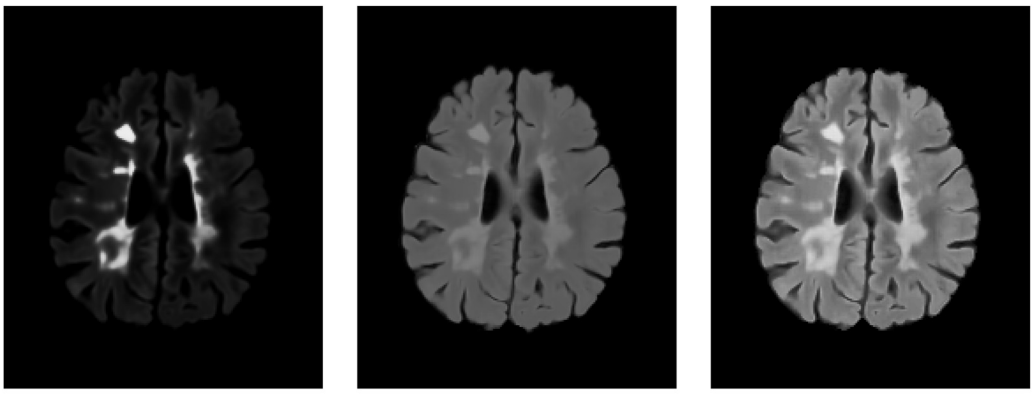


Fig. 3. Hyperintensity map (a) compared to the original FLAIR image (b) and the intermediate image (c).

miss. By doing so, voxels with bright neighborhoods are enhanced while other regions and tissues are dimmed out. Moreover, since we normalize the score S_v, each voxel remains in the range [0, 1], which also serves as a hyperintensity probability indicator. An example of the map is shown in Fig. 3.

It is important to note that our approach requires no hard threshold for Eq. (11) or a fixed number of patches as it happens in [16]. Instead, the threshold is calculated automatically with respect to the standard deviation. Though simple, this is a significant improvement, since the main problem of using a hard threshold is that even normalized, intensities inherently vary from image to image. Accordingly, a soft threshold such as the one we propose in this work offers a better option because the enhancement of hyperintensities for it can adapt to each image intensity profile.

2.5. White matter mask estimation

The white matter region usually comprises most MS lesions [7]. An automatic brain segmentation into three clusters (WM, GM, and CSF) based on gray level intensities is most certainly going to mix lesions and cluster them as GM, WM or both [1,21]. In this context, being able to estimate a mask that encompasses both white matter tissue and MS lesions can help narrow down the ROI and increase the accuracy of lesion segmentation. To do so, we leveraged the fact that the map described in Section 2.4 can also be interpreted as a probability map and used it to obtain an estimate of such mask.

In this work, we made use of the Student t mixture model proposed in [11] and used T1-weighted and FLAIR images from each time-point to segment the brain into three different clusters and obtain an initial WM mask, herein referred to as $WM_{initial}$. To automatically identify the WM cluster from others, we used the WM probability map described in Section 2.1.2, averaged it over each cluster and selected the one with the highest WM probability.

Since the 2015 Longitudinal MS Lesion Segmentation Challenge provided no WM ground truths, we created our own using a straightforward approach. Given any $WM_{initial}$, we simply merged it with the lesions ground truth to get the whole WM region in one single mask, herein referred to as WM_{whole}. Considering that each time-point had two different lesions ground truth, we created two WM masks for each time-point as well.

The actual WM estimation was conducted as follows. First, we calculated the mean (μ_{HI}) and standard deviation (σ_{HI}) of the region defined by $WM_{initial}$ on the hyperintensity map image and the mean (μ_{prob}) of the region defined by $WM_{initial}$ on the WM probability atlas. The idea was to expand $WM_{initial}$ by considering voxels that are not part of the mask yet and analyzing $3 \times 3 \times 3$ neighborhoods centered around these voxels to determine their potential for being included. The expansion itself occurred by incorporating voxels that seemed as outliers; more precisely, voxels with mean neighborhood values greater than $\mu_{HI} + 1 \times \sigma_{HI}$ in the hyperintensity map and greater than μ_{prob} in the probability atlas. A pseudo-algorithm for estimating the white matter mask is presented in Algorithm 1 and an example of the output of this estimation is shown in Fig. 4.

Algorithm 1. White matter mask estimation algorithm

Input: $WM_{initial}$, HI_{map}, WM_{prob}
Output: WM_{estim}
  $i \longleftarrow 3$
  $WM_{estim} \longleftarrow WM_{initial}$
  $\mu_{HI}, \sigma_{HI} \longleftarrow MeanAndSigma(WM_{estim}, HI_{map})$
  $\mu_{prob} \longleftarrow Mean(WM_{estim}, WM_{prob})$
  $t_{prob} \longleftarrow \mu_{prob}$
  $t_{HI} \longleftarrow \mu_{HI} + i \times \sigma_{HI}$
  for each voxel **not** in WM_{estim} **do**
    $WM_{estim} \longleftarrow ExpandWM(WM_{estim}, HI_{map}, WM_{prob}, t_{HI}, t_{prob})$
  end for

2.6. Pure white matter and gray matter clusters

To estimate the intensity profiles of white matter and gray matter clusters without lesions, we did the following. For each time-point segmentation, we automatically identified the white and gray matter clusters by analyzing their mean intensities on white and gray matter probabilistic atlases. The cluster with highest white matter atlas mean intensity was taken as the white matter cluster ($WM_{initial}$); the same rationale was used for the gray matter cluster ($GM_{initial}$). Then, for each expert annotation, we simply excluded every voxel that had any intersection with the lesion ground truth and these two clusters. Formally, $WM_{pure} = WM_{initial} - (WM_{initial} \cap GT_{E \in \{\text{Expert1,Expert2}\}})$ and $GM_{pure} = GM_{initial} - (GM_{initial} \cap GT_{E \in \{\text{Expert1,Expert2}\}})$. By doing so, we were able to obtain so-called "pure" WM and GM clusters, which were then used to calculate their intensity profiles and compare them to lesion profiles in Section 3.1.

3. Results

In this section, we present the results regarding the brightness intensity profile of MS lesions compared to gray matter and white matter on FLAIR, intermediate and hyperintensity map images for each patient.

For the sake of comparison, all images were rescaled to the range [0, 1] and the results in Section 3.1 are shown in percentage; that is, how much brighter, percent-wise, the lesion profile was compared to other brain tissues.

Fig. 4. White matter estimation output. (a) Hyperintensity map, (b) $WM_{initial}$, (c) $WM_{ground\ truth}$ and (d) WM_{estim}.

We also present the Dice, sensitivity, specificity and lesion intersection metrics regarding the white matter mask in Section 3.2 to provide a quantitative analysis of the mask estimation.

It is important to note that each patient had a number of time-points made available. For the sake of understandability and inter-patient comparison, every result presented in Sections 3.1 and 3.2 represents the average of all time-points of a particular patient.

3.1. Brightness profile

Since the 2015 Longitudinal MS Lesion Segmentation Challenge provided two ground truths for each time-point, we extracted the intensity profiles for both annotations. We rescaled all images to the [0, 1] interval, averaged the white matter, gray matter and lesion profiles and also calculated the standard deviation for each patient. The results are shown in Figs. 5 and 6 for experts 1 and 2, respectively.

Each bar in Figs. 5 and 6 represents the mean lesion intensity over the mean intensity of a given tissue (gray matter or white matter) in a particular image type (FLAIR, intermediate and hyper-intensity map). For instance, the FLAIR WM bar in Fig. 5(a) must be interpreted as "the average MS lesion profile in FLAIR images from Patient 1 is approximately 25% brighter than the average WM tissue intensity for the same image type and patient". This result allows a direct comparison between tissues and images.

3.2. White matter mask comparison

We compared the white matter mask estimated in Section 2.5 with our WM ground truths using the Dice coefficient. We also extracted the percentage of lesions (intersection) present in each estimated mask to analyze the lesion load that was kept during the estimation. Again, since there were two lesion ground truths for each time-point, we extracted metrics for both experts. The results are presented in Table 2 and shown in Figs. 8 and 9.

We also calculated the sensitivity and specificity of the estimated masks. The results for both experts are shown in Table 3.

4. Discussion

The results in Figs. 5 and 6 indicate a significant difference in the MS lesion intensity profile in the hyperintensity map compared to FLAIR and the intermediate image described in Section 2.3. This result is a significant outcome, because it provides quantitative background to show the discriminative features of the HI map.

It is possible to note that the lesion intensity profile was more similar to gray matter than to white matter. The difference in intensity between MS lesions and these two tissues in FLAIR and intermediate images was minimal compared with the HI map, which showed, in the worst case (Patient 1, Expert 2, GM), a 141% brightness gap. In contrast, this same case presented an 18% and 32% IPD for FLAIR and intermediate images.

At the same time, the standard deviation in the HI map was far higher than in FLAIR and intermediate images, which is an indication that the map has a rather spread out MS lesion intensity profile. While this is a concern that must be addressed when using the map to segment lesions, whether manually or automatically, the overall difference in intensity between lesions and other brain tissues is still significant and provides enough leverage to overcome, at least partially, the wide standard deviation variation.

Moreover, a distinct drawback of the HI map is that it is highly dependent on the gray level intensity in FLAIR. This fact poses two problems that can be observed in Fig. 7. The first one is presented in Fig. 7(a)–(c) and concerns the natural intensity variation within the lesion profile. In this case, the lesion enclosed by the rectangle was not as enhanced as the one enclosed by the elipse. As observed in Fig. 7(a), the rectangle lesion did not present a profile as hyperintense as the elipse one in FLAIR, so this difference was propagated to the HI map. The other problem is shown in Fig. 7(d)–(f). Areas enclosed by rectangles indicate regions that have no MS lesions and yet are enhanced in the map. Again, this occurs because these regions presented rather high-intensity profiles in FLAIR and thus were enhanced in the map. Both these problems interfere with lesion segmentation accuracy and indicate that the HI map should not be used as a stand-alone feature in manual and automatic segmentation techniques.

While there are some works in the literature [17,19,1] that mention MS lesion intensity profiles and how they relate to other tissues, none of them provide a quantitative analysis regarding percentage, making it difficult to compare our results with theirs objectively. The databases are also different. However, by analyzing scatter plots in [17,19], it is possible to observe that lesion intensity profiles present a significant overlap with other brain tissues. Hence, the HI map can undoubtedly help distinguish lesions more easily.

Regarding the white matter mask estimation, the results shown in Tables 2 and 3 and in Figs. 8 and 9 indicate high DSCs and significant intersection with lesions. It also points to very high sensitivity and specificity metrics. A relevant observation to be made is that the LI metric presented a consistent level of intersection regardless of lesion volumes, which is an indication of robustness of our technique.

In Fig. 8, it is possible to observe that patient 4 presented very different results on lesion intersection. The reason for this is that the expert annotations for this patient had the lowest DSC (0.612 ± 0.0019) among all patients, as mentioned in [11]. In other words, experts did not have a high agreement coefficient on lesion segmentation for this particular case, which consequently made

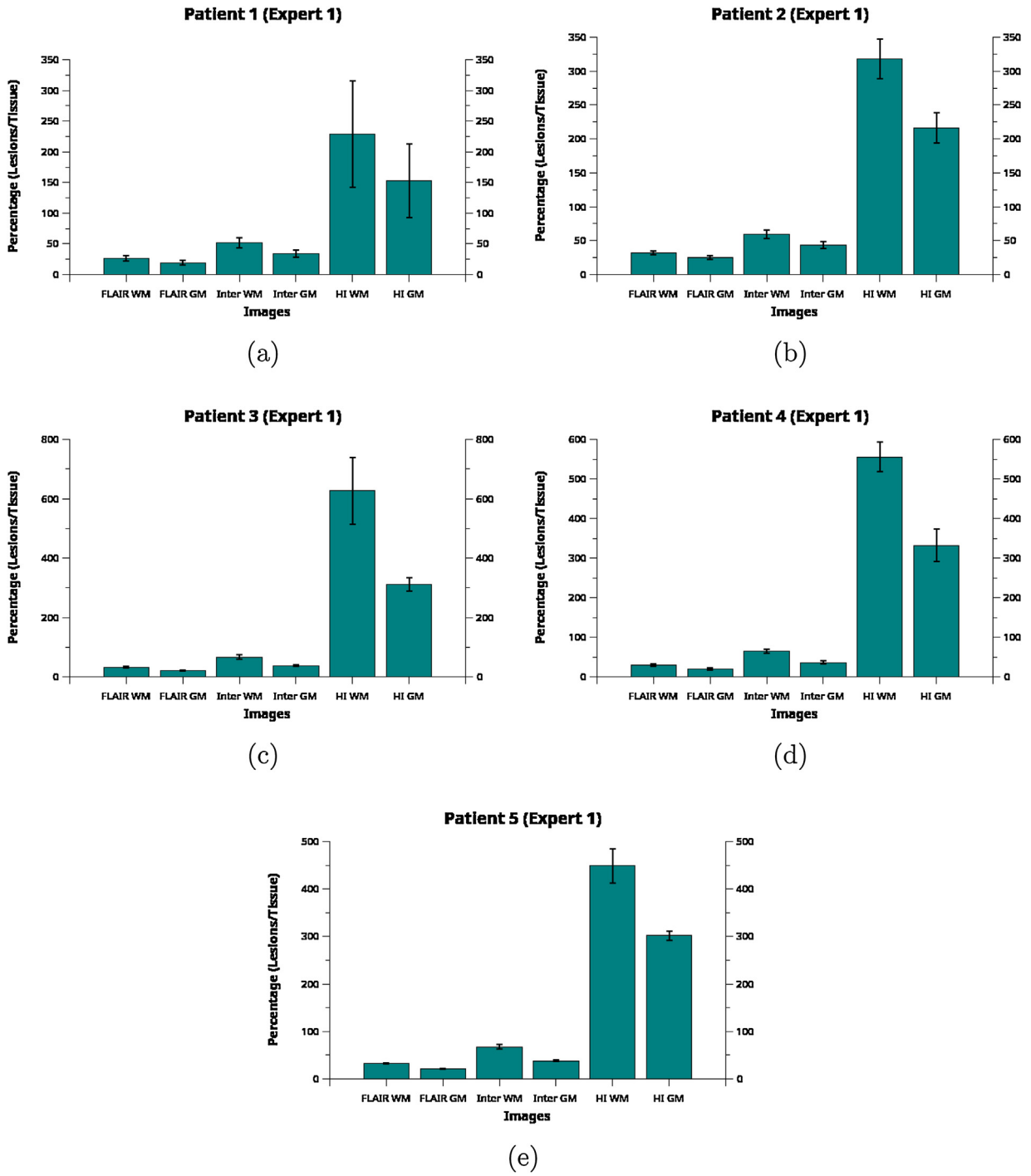


Fig. 5. Lesion intensity profile compared with white matter and gray matter tissues using ground truths from expert 1. Here, "Inter" is the intermediate image and "HI" is the hyperintensity map. Each bar represents the mean lesion intensity over the mean intensity of a given tissue (white matter or gray matter) in a particular image type.

Table 2
Lesion intersection (LI) and Dice coefficients for the white matter mask estimation for both expert ground truths.

	LI (%) Expert 1 ($\mu \pm \sigma$)	LI (%) Expert 2 ($\mu \pm \sigma$)	Dice Expert 1 ($\mu \pm \sigma$)	Dice Expert 2 ($\mu \pm \sigma$)
Patient 1	78.56 ± 6.70	77.65 ± 8.30	0.9763 ± 0.0003	0.9764 ± 0.0007
Patient 2	89.60 ± 1.59	88.20 ± 1.71	0.9786 ± 0.0021	0.9775 ± 0.0020
Patient 3	83.88 ± 0.94	79.08 ± 2.51	0.9834 ± 0.0007	0.9829 ± 0.0005
Patient 4	76.83 ± 1.20	56.95 ± 4.03	0.9860 ± 0.0012	0.9853 ± 0.0013
Patient 5	73.73 ± 3.55	71.00 ± 2.16	0.9828 ± 0.0017	0.9826 ± 0.0017

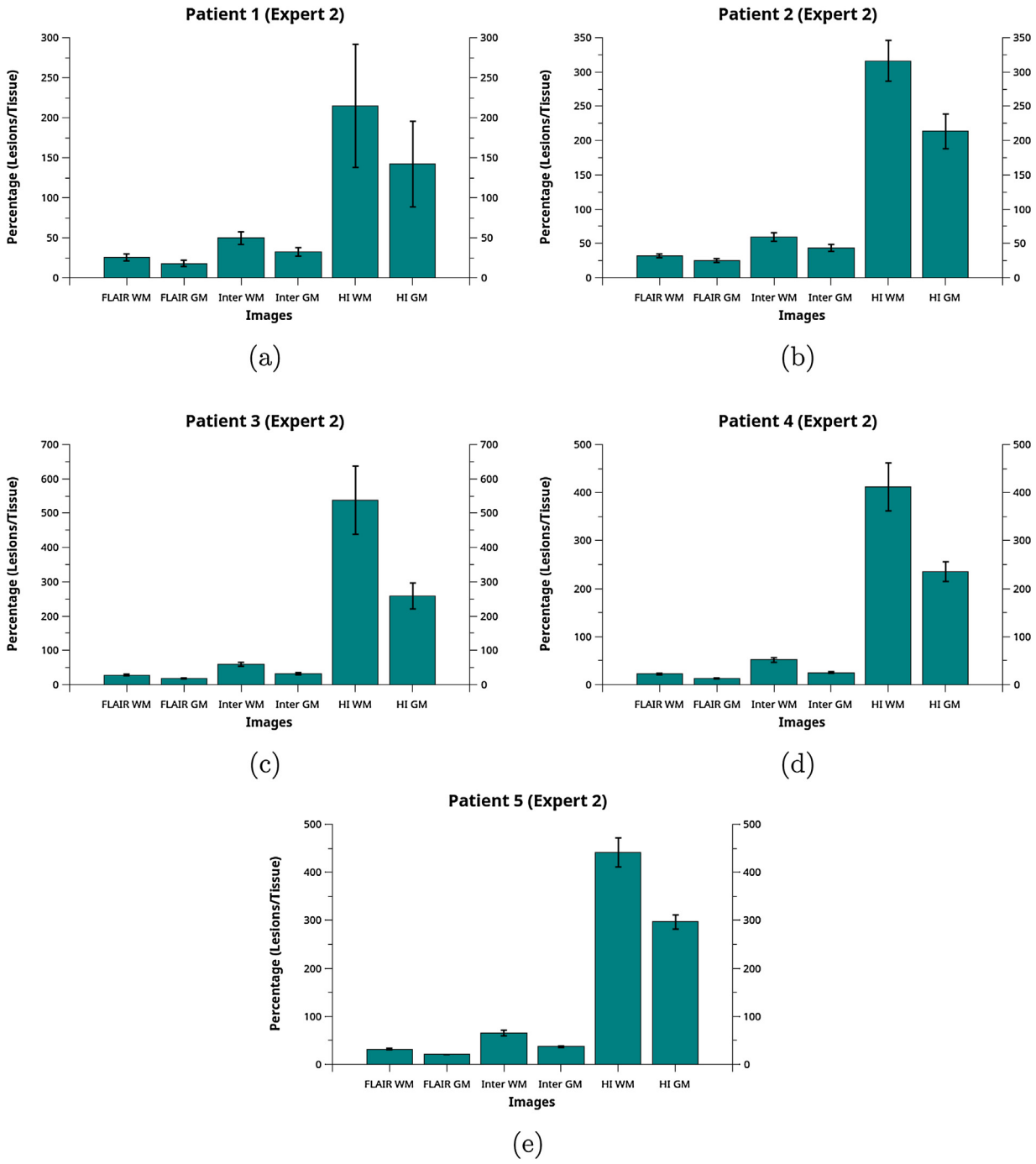


Fig. 6. Lesion intensity profile compared with white matter and gray matter tissues using ground truths from expert 2. Here, "Inter" is the intermediate image and "HI" is the hyperintensity map. Each bar represents the mean lesion intensity over the mean intensity of a given tissue (white matter or gray matter) in a particular image type.

Table 3
Sensitivity (SS) and specificity (SP) values for the white matter mask estimation for both expert ground truths.

	SS Expert 1 ($\mu \pm \sigma$)	SS Expert 2 ($\mu \pm \sigma$)	SP Expert 1 ($\mu \pm \sigma$)	SP Expert 2 ($\mu \pm \sigma$)
Patient 1	0.9931 ± 0.0024	0.9920 ± 0.0028	0.9786 ± 0.0012	0.9793 ± 0.0008
Patient 2	0.9928 ± 0.0012	0.9915 ± 0.0014	0.9823 ± 0.0023	0.9818 ± 0.0016
Patient 3	0.9983 ± 0.0001	0.9968 ± 0.0009	0.9816 ± 0.0010	0.9818 ± 0.0011
Patient 4	0.9990 ± 0.0008	0.9975 ± 0.0002	0.9818 ± 0.0015	0.9818 ± 0.0016
Patient 5	0.9974 ± 0.0002	0.9974 ± 0.0002	0.9822 ± 0.0020	0.9820 ± 0.0020
Overall	0.9962 ± 0.0028	0.9813 ± 0.0021	0.9951 ± 0.0030	0.9814 ± 0.0017

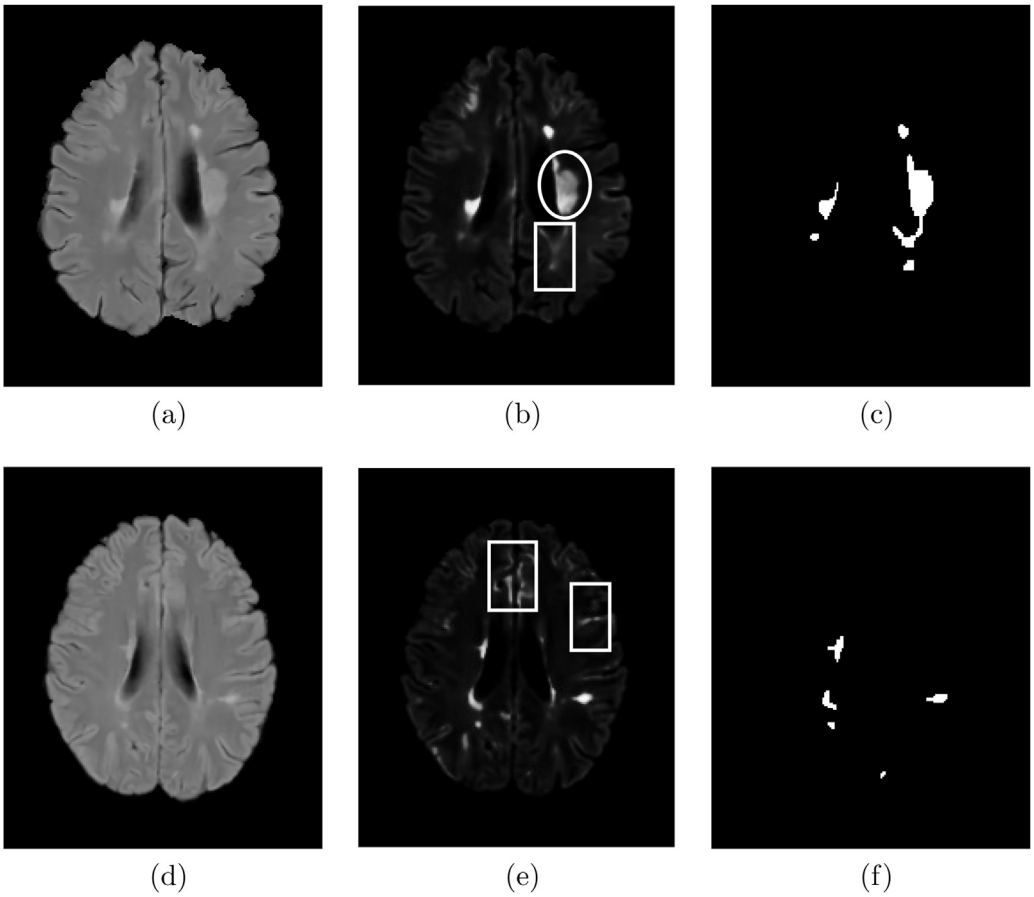


Fig. 7. Intensity problems caused by wide lesion intensity range (first row) and hyperintensities in regions other than lesions (second row).

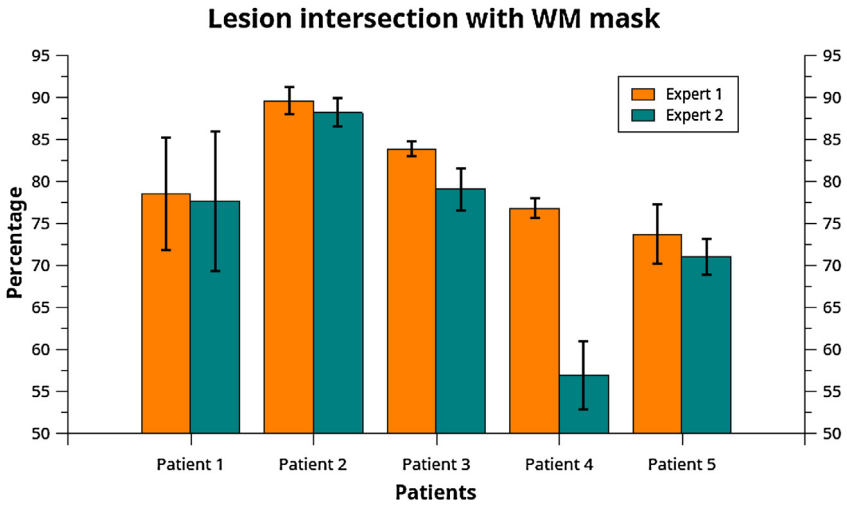


Fig. 8. Lesion intersection with the estimated white matter mask using ground truths from both experts.

our technique present very different intersection values for each annotation.

Another point to be made about Fig. 8 is that patient 1 presented the highest standard deviation of all. This results from the fact that this patient's lesion intensities faded across time-points, making the enhancement less effective. This fading phenomenon can also be observed in Figs. 5 and 6, since patient 1 had the highest standard deviation on the lesion intensity profile in the HI map compared with other patients.

As mentioned in Section 1, there are several works in the literature focused on automatic brain tissue segmentation [1,21–23]. Though a direct comparison is not possible due to different metrics being used and database access restrictions, the closest work to ours regarding white matter estimation was [23]. Contrary to the authors' approach, our proposal only requires FLAIR images (from which the HI maps are created) and WM probability atlases to fill lesion holes left out during an automatic segmentation process. It is important to note that our technique focuses only on the white matter region at this point, while all three major brain tissues are

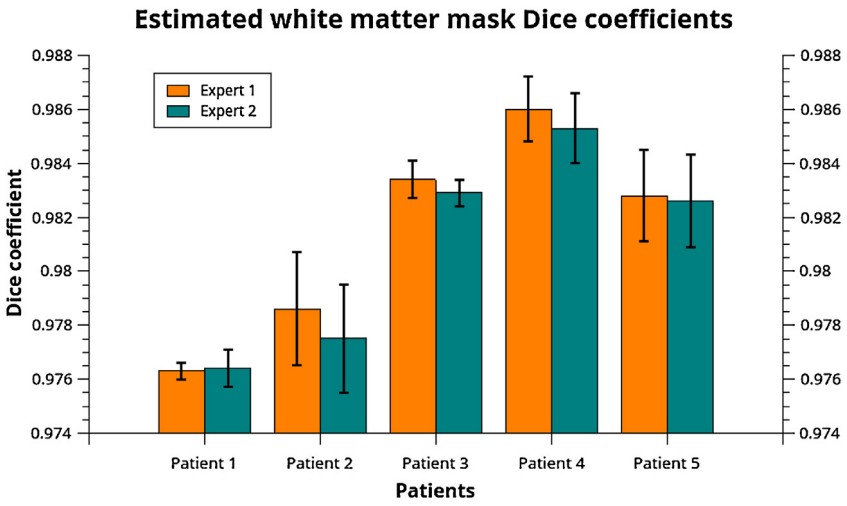


Fig. 9. Dice coefficients of the estimated white matter mask compared to ground truths created using experts lesion annotations and automatic brain segmentation.

segmented in [23]. However, we believe that it is possible to extend our work to also handle gray matter and cerebrospinal fluid tissues, thus allowing a thorough comparison between techniques in the future.

5. Conclusions

This work presented an automatic technique based on the works of [16,15] to enhance hyperintensities in FLAIR images, making it easier to distinguish multiple sclerosis lesions from other brain tissues, namely gray matter and white matter. By defining a metric called Intensity Profile Difference (IPD), we were able to analyze, percent-wise, how much brighter the lesion profile was compared with other tissues and image types on five patients from the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge.

The hyperintensity map, created by the enhancement process, provided a much more distinct lesion profile compared with FLAIR. On average, lesions presented a mean intensity profile 444.57% and 264.88% brighter than white matter and gray matter in the HI map, respectively. In FLAIR, the same profile was only 25% and 19% brighter considering the same tissues. This result serves as an essential aid for segmentation tasks on both manual and automatic segmentation techniques.

The HI map has two significant drawbacks. The first one regards the intensity variation within lesions, since the difference in intensity from one lesion to another in FLAIR is propagated to the map. The second drawback regards regions that have no lesions but also appear hyperintense in FLAIR. These regions are also enhanced in the map and lead to false positives. Therefore, the HI map should not be used as a stand-alone feature in applications such as tissue or lesion segmentation.

A byproduct of the HI map is an initial estimate of the white matter mask for a given time-point. Automatically segmenting a brain image with multiple sclerosis into three clusters (white matter, gray matter, and cerebral spinal fluid) will undoubtedly mix lesions with other tissues. In this sense, we can simply obtain the white matter cluster mask and fill regions that are not yet in the mask but are above a certain threshold in the map to obtain a "full" WM mask. The estimation of the white matter area is relevant to narrow down the ROI when segmenting lesions and also help with brain tissue segmentation and volume assessment.

We showed that lesions have an intensity profile that is brighter than WM than it is to GM, so we believe that restricting the segmentation area to a mask that excludes most gray matter region might increase lesion segmentation accuracy. However, the two problems mentioned before about the HI map also affected the white matter mask estimation. Part of the lesions was left out, as evidenced by the LI metric, and some regions not related to white matter were included in the estimation.

In conclusion, the results of this study showed that the hyperintensity map provides a much more distinct profile for multiple sclerosis lesions compared with white matter and gray matter tissues in FLAIR and such map can also be used to estimate an initial white matter mask. In future works, we aim to address the problems with the enhancement algorithm mentioned earlier and efficiently use it to increase both tissue and lesion segmentation accuracies in automatic techniques. We also plan on extending the WM estimation technique to the other two primary brain tissues (WM and cerebrospinal fluid).

Conflicts of interest

None declared.

Acknowledgement

This work was supported by the São Paulo Research Foundation (FAPESP) (grant 2016/15661-0).

References

[1] M. Battaglini, M. Jenkinson, N.D. Stefano, Evaluating and reducing the impact of white matter lesions on brain volume measurements, Hum. Brain Mapp. 33 (9) (2012) 2062–2071.
[2] P. Browne, D. Chandraratna, C. Angood, H. Tremlett, C. Baker, B. Taylor, A. Thompson, Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity, Neurology 83 (11) (2013) 1022–1024.
[3] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (2005) 60–65.
[4] A. Carass, J. Cuzzocreo, M.B. Wheeler, P.-L. Bazin, S.M. Resnick, J. Prince, Simple paradigm for extra-cerebral tissue removal: algorithm and analysis, NeuroImage 56 (4) (2011) 1982–1992.
[5] A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C.H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C.A. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S.K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L.O. Iheme, D. Unay, S. Jain, D.M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.L. Bazin, P.A. Calabresi, C.M. Crainiceanu, L.M. Ellingsen, D.S. Reich, J.L. Prince, D.L. Pham, Longitudinal multiple sclerosis lesion segmentation: resource and challenge, NeuroImage 148 (2017) 77–102.
[6] A. Carass, M.B. Wheeler, J. Cuzzocreo, P.-L. Bazin, S.S. Bassett, J. Prince, A joint registration and segmentation approach to skull stripping, in: Proceedings of

the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Arlington, VA, USA, IEEE, April 2007, pp. 656–659.
[7] A. Compston, A. Coles, Multiple sclerosis, Lancet 372 (October (9648)) (2008) 1502–1517.
[8] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.
[9] O.R. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd edition, Wiley-Interscience, Hoboken, USA, 2000.
[10] V.S. Fonov, A.C. Evans, R.C. McKinstry, C.R. Almli, D.L. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage 47 (2009) S102.
[11] P.G.L. Freire, R.J. Ferrari, Automatic iterative segmentation of multiple sclerosis lesions using Student's t mixture models and probabilistic anatomical atlases in FLAIR images, Comput. Biol. Med. 73 (2016) 10–23.
[12] R.H. Hashemi, W.G. Bradley, Y.D. Chen, J.E. Jordan, J.A. Queralt, A.E. Cheng, J.N. Henrie, Suspected multiple sclerosis: MR imaging with a thin-section fast FLAIR pulse sequence, Radiology 196 (2) (1995) 505–510.
[13] M. Modat, G.R. Ridgway, Z.A. Taylor, M. Lehman, J. Barnes, D.J. Hawkes, N.C. Fox, S. Ourselin, Fast free-form deformation using graphics processing units, Comput. Methods Programs Biomed. 98 (3) (2010) 278–284.
[14] L. Nyúl, J. Udupa, X. Zhang, New variants of a method of MRI scale normalization, IEEE Trans. Med. Imaging 19 (2) (2000) 143–150.
[15] P.K. Roy, A. Bhuiyan, A. Janke, P.M. Desmond, T.Y. Wong, E. Storey, W.P. Abhayaratna, K. Ramamohanarao, Automated segmentation of white matter lesions using global neighbourhood given contrast feature-based random forest and Markov random field, in: Proceedings of the 2014 IEEE International Conference on Healthcare Informatics, Verona, Italy, IEEE, September 2014, pp. 1–6.
[16] P.K. Roy, A. Bhuiyan, K. Ramamohanarao, Automated segmentation of multiple sclerosis lesion in intensity enhanced FLAIR MRI using texture features and support vector machine, in: Proceedings of the 20th IEEE International Conference on Image Processing, Melbourne, VIC, Australia, IEEE, September 2013, pp. 4277–4281.
[17] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. Arnold, D. Collins, T. Arbel, Evaluating intensity normalization on MRIs of human brain with multiple sclerosis, Med. Image Anal. 15 (2011) 267–282.
[18] N. Shiee, P.-L. Bazin, J.L. Cuzzocreo, C. Ye, B. Kishore, A. Carass, P.A. Calabresi, D.S. Reich, J.L. Prince, D. Pham, Robust reconstruction of the human brain cortex in the presence of the WM lesions: method and validation, Hum. Brain Mapp. 35 (7) (2014) 3385–3401.
[19] X. Tomas-Fernandez, S. Warfield, A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation, IEEE Trans. Med. Imaging 34 (6) (2015) 1349–1361.
[20] N. Tustison, J. Gee, N4ITK: Nick's N3 ITK Implementation for MRI Bias Field Correction, Penn Image Computing and Science Laboratory, 2009.
[21] S. Valverde, O.A. Dez, Y. Cabezas, M. Vilanova, J. Rami-Torrentà, A. Rovira, X. Lladó, Evaluating the effects of white matter multiple sclerosis lesions on the volume estimation of 6 brain tissue segmentation methods, Am. J. Neuroradiol. 36 (6) (2015) 1109–1115.
[22] S. Valverde, A. Oliver, M. Cabezas, E. Roura, X. Lladó, Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations, J. Magn. Reson. Imaging 41 (2015) 93–101.
[23] S. Valverde, A. Oliver, E. Roura, S. González-Villà, D. Pareto, J. Vilanova, L. Ramío-Torrentà, A. Rovira, X. Lladó, Automated tissue segmentation of MR brain images in the presence of white matter lesions, Med. Image Anal. 35 (2016) 446–457.
[24] S. Warren, K. Warren, Multiple Sclerosis, 2001 (last accessed in 28.11.17) http://bit.ly/2AbSVqj.
[25] World Health Organization, Atlas – Multiple Sclerosis Resources in the World, 2008 (last accessed in 28.11.17) http://bit.ly/14Ddglh.

REFERENCES

ABDULLAH, B.; YOUNIS, A.; JOHN, N. Multi-sectional views textural based SVM for MS lesion segmentation in multi-channels MRIs. *The open biomedical engineering journal*, v. 6, p. 56–72, 2012.

ADMIRAAL-BEHLOUL, F. et al. Fully automatic segmentation of white matter hyperintensities in mr images of the elderly. *Neuroimage*, v. 28, n. 3, p. 607–617, 2005.

AGENCY, E. M. *EMA's final opinion confirms restrictions on use of linear gadolinium agents in body scans*. 2017. Http://bit.ly/EuropeanMedicinesAgency2017.

AHMED, M. et al. A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *IEEE Transactions on Medical Imaging*, v. 21, n. 3, p. 193–199, 2002.

AJA-FERNÁNDEZ, S.; VEGAS-SÁNCHEZ-FERRERO, G. *Statiscial Analysis of Noise in MRI*. Switzerland: Springer International Publishing, 2016.

ALI, E.; BUCKLE, G. Neuroimaging in mutiple sclerosis. *Neurologic Clinics*, v. 27, n. 1, p. 203–219, 2009.

ANBEEK, P. et al. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Medical Image Analysis*, v. 8, n. 3, p. 205–215, 2004.

ARDAKANI, A. et al. Quantitative MRI Texture Analysis in Differentiating Enhancing and Non-enhancing T1-hypointense Lesions without Application of Contrast Agent in Multiple Sclerosis. *Neurologie a Neurochirurgie*, v. 80, n. 6, p. 700–707, 2017.

ARNOLD, J. et al. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *NeuroImage*, v. 13, n. 5, p. 931–943, 2001.

ASHBURNER, J.; FRISTON, J. Unified segmentation. *NeuroImage*, v. 26, n. 3, p. 839–851, 2005.

ASHBURNER, J.; FRISTON, K. Voxel-based morphometry - The methods. *NeuroImage*, v. 11, p. 805–821, 2000.

AUBERT-BROCHE, B. et al. Twenty new digital brain phantoms for creation of validation image data bases. *IEEE Transactions on Medical Imaging*, v. 25, n. 11, p. 1410–1416, 2006.

B, A. et al. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. *Journal of Magnetic Resonance Imaging*, v. 12, n. 6, p. 799–807, 2000.

BAKSHI, R. et al. MRI in multiple sclerosis: current status and future prospects. *Lancet Neurology*, v. 7, n. 7, p. 615–625, 2008.

BARTKO, J. Measurement and reliability: statistical thinking considerations. *Schizophrenia Bulletin*, v. 17, n. 3, p. 483–489, 1991.

BATTAGLINI, M.; JENKINSON, M.; STEFANO, N. Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Human Brain Mapping*, v. 33, n. 9, p. 2062–2071, 2012.

BEHRENBRUCH, C. et al. Image filtering techniques for medical image post-processing: an overview. *British Journal of Radiology*, v. 77, n. 2, p. S126–S132, 2004.

BENNETT, J.; LANNING, S. The Netflix Prize. In: *KDD Cup and Workshop in conjunction with KDD*. San Jose, California, USA: [s.n.], 2007.

BENTO, M. et al. Probabilistic Segmentation of Brain White Matter Lesions Using Texture-Based Classification. In: KARRAY, F.; CAMPILHO, A.; CHERIET, F. (Ed.). *Proceedings of the 2017 Internatinal Confere on Image Analysis and Recognition*. Cham: Springer International Publishing, 2017. p. 71–78.

BISHOP, C. *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press, 1995.

BISHOP, C. *Pattern recognition and machine learning*. Berlin, Germany: Springer, 2006.

BLAND, M. *An introduction to medical statistics*. 3. ed. Oxford, UK: Oxford University Press, 2000.

BLEI, D.; NG, A.; JORDAN, M. Latent Dirichlet allocations. *Journal of Machine Learning Research*, v. 3, n. 1, p. 993–1022, 2003.

BLEKAS, K. et al. A spatially constrained mixture model for image segmentation. *IEEE Transactions on Neural Networks*, v. 16, n. 2, p. 494–498, 2005.

BO, L. et al. Gray matter pathology in multiple sclerosis. *Acta Neurologica Scandinavica*, v. 183, p. 48–50, 2006.

BORSCH, T. et al. Deep 3D convolutional encoder networks with shortcuts for multiscale feature Integration applied to multiple sclerosis lesion segmentation. *Transactions on Medical Imaging*, 2015.

BOYKOV, Y.; VEKSLER, O.; ZABIH, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 23, n. 11, p. 1222–1239, 2001.

BRACEWELL, R. *The Fourier Transform and Its Applications*. New York, USA: McGraw Hill, 2000.

BREIMAN, L. et al. *Classification and regression trees*. USA: Wadsworth International Group, 1984.

BROSCH, T. et al. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, v. 35, n. 5, p. 1229–1239, 2016.

BROWNE, P. et al. Atlas of multiple sclerosis 2013: A growing global problem with widespread inequity. *Neurology*, v. 83, n. 11, p. 1022, 1024 2013.

BUADES, A.; COLL, B.; MOREL, J.-M. A non-local algorithm for image denoising. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 2, p. 60–65, 2005.

BURKE, L. et al. Self-reported gadolinium toxicity: a survey of patients with chronic symptons. *Magnetic Resonance Imaging*, v. 34, p. 1078–1080, 2016.

BUSHBERG, J. et al. *The Essential Physics of Medical Imaging*. 3. ed. Philadelphia, USA: Lippincott Williams & Wilkins, 2012.

CARASS, A. et al. Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis. *NeuroImage*, v. 56, n. 4, p. 1982–1992, 2011.

CARASS, A. et al. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, v. 148, p. 77–102, 2017.

CARASS, A. et al. A Joint Registration and Segmentation Approach to Skull Stripping. In: *Proceedings of the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Arlington, VA, USA: IEEE, 2007. p. 656–659.

CASTELLANO, G. et al. Texture analysis of medical images. *Clinical Radiology*, v. 59, n. 12, p. 1061–1069, 2004.

CHARD, D. et al. Brain atrophy in clinically early relapsing-remitting multiple sclerosis. *Brain*, v. 125, p. 327–337, 2002.

CHATZIS, S.; KOSMOPOULOS, D.; VARVARIGOU, T. Robust sequential data modeling using an outlier tolerant hidden Markov model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 31, n. 9, p. 1657–1669, 2009.

CHATZIS, S.; VARVARIGOU, T. A fuzzy clustering approach toward hidden markov random field models for enhanced spatially constrained image segmentation. *IEEE Transactions on Fuzzy Systems*, v. 16, n. 5, p. 1351–1361, 2008.

CHEN, L.; BENTLEY, P.; RUECKERT, D. A novel framework for sub-acute stroke lesion segmentation based on random forest. In: *Proceedings of Ischemic Stroke Lesion Segmentation Challenge, held in conjunction with International Conference on Medical Image Computing and Computer Assisted Intervention*. Munich, Germany: Springer, 2015.

CHEN, S.; ZHANG, D. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. *IEEE Transactions on Systems Man and Cybernetics*, v. 34, n. 4, p. 1907–1916, 2004.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

COMPSTON, A.; COLES, A. Multiple Sclerosis. *Lancet*, v. 372, n. 9648, p. 1502–1517, October 2008.

DALE, A.; FISCHL, B.; SERENO, M. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*, v. 9, p. 179–194, 1999.

DAMANGIR, S. et al. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. *Journal of Neurological Science*, v. 322, p. 211–216, 2012.

DATTA, S. et al. Segmentation and quantifications of black holes in multiple sclerosis. *NeuroImage*, v. 29, n. 2, p. 467–474, 2006.

DESPOTOVIC, I.; GOOSSENS, B.; PHILIPS, W. MRI segmentation of the human brain: Challenges, methods, and applications. *Computational and Mathematical Methods in Medicine*, v. 2015, n. 2015, p. 23 pages, 2015.

DICE, L. Measures of the amount of ecologic association between species. *Ecology*, v. 26, n. 3, p. 297–302, 1945.

DO, C.; BATZOGLOU, S. What is the expectation maximization algorithm? *Nature Biotechnology*, v. 26, n. 8, p. 897–899, 2008.

DOUGHERTY, G. *Digital image processing for medical applications*. 1. ed. The Edinburgh Building, Cambridge CB2 8RU, UK: Cambridge University Press, 2009.

DOUGHERTY, G. *Medical Image Processing - Techniques and Applications*. 1. ed. Springer New York Dordrecht Heidelberg London: Springer, 2011.

DOYLE, A. et al. Lesion Detection, Segmentation and Prediction in Multiple Sclerosis Clinical Trials. In: CRIMI, A. et al. (Ed.). *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing, 2018. p. 15–28. ISBN 978-3-319-75238-9.

DUDA, O.; HART, P.; STORK, D. *Pattern Classification (2Nd Edition)*. Hoboken, USA: Wiley-Interscience, 2000.

DYRBY, T. et al. Segmentation of age-related white matter changes in a clinical multi-center study. *Neuroimage*, v. 41, n. 2, p. 335–345, 2008.

EDELMAN, R.; HESSELINK, J.; ZLATKIN, M. *Clinical Magnetic Resonance Imaging*. 3. ed. Philadelphia: Saunders, 2006.

EGGER, C. et al. MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? *NeuroImage: Clinical*, v. 13, p. 264–270, 2017.

European Medicines Agency. *EMA reviewing gadolinium contrast agents used in MRI scans*. 2016. http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/referrals/Gadolinium-containing_contrast_agents/human_referral_prac_000056.jsp&mid=WC0b01ac05805c516f.

FATIMA, M.; PASHA, M. Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, p. 1–16, 2017.

FENG, C.; ZHAO, D.; HUANG, M. Segmentation of stroke lesions in multi-spectral MR images using bias correction embedded FCM and three phase level set. In: *Proceedings of Ischemic Stroke Lesion Segmentation Challenge, held in conjunction with International Conference on Medical Image Computing and Computer Assisted Intervention*. Munich, Germany: Springer, 2015.

FERNANDES, K.; CARDOSO, J. Deep local binary patterns. *Computing Research Repository*, abs/1711.06597, p. 1–10, 2017.

FERRARI, R. Off-line determination of the optimal number of iterations of the robust anisotropic diffusion filter applied to denoising of brain MR images. *Medical & Biological Engineering & Computing*, v. 51, n. 1-2, p. 71–88, 2013.

FILIPPI, M. et al. Glatiramer acetate reduces the proportion of new MS lesions evolving into black holes. *Neurology*, v. 57, n. 4, p. 731–733, 2001.

FISCHL, B.; SERENO, M.; DALE, A. Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, v. 9, p. 195–207, 1999.

FONOV, V. et al. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, v. 47, p. S102, 2009.

FOOD, U.; ADMINISTRATION, D. *FDA warns that gadolinium-based contrast agentes (GBCAs) are retained in the body; requires new class warnings*. 2017. Https://www.fda.gov/media/109825/download.

FOOD, U.; ADMINISTRATION, D. *Update on FDA approach to safety issue of gadolinium retention after administration of gadolinium-based contrast agents*. 2018. Https://www.fda.gov/media/116492/download.

FORBES, F.; PEYRAR, N. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 25, n. 9, p. 1089–1101, 2003.

FREIRE, P.; FERRARI, R. Do multiple sclerosis lesions affect the outcome of magnetic resonance image registration? *Revista de Informática Teórica e Aplicada*, v. 21, n. 2, p. 47–63, 2014.

FREIRE, P.; FERRARI, R. Automatic iterative segmentation of multiple sclerosis lesions using Student's t mixture models and probabilistic anatomical atlases in FLAIR images. *Computers in Biology and Medicine*, v. 73, p. 10–23, 2016.

FREIRE, P.; FERRARI, R. Multiple sclerosis lesion enhancement and white matter region estimation using hyperintensities in FLAIR images. *Biomedical Signal Processing and Control*, v. 49, p. 338–348, 2019.

FREIRE, P. et al. Midsaggital plane detection in magnetic resonance images using phase congruency, Hessian matrix and symmetry information: a comparative study. In: *Proceedings of the 18th International Conference of Computational Science and Its Applications*. Melbourne, Victoria, Australia: Springer, 2018. p. 245–260.

GALLOWAY, M. Texture analysis using run lengths. *Computer Graphics and Image Processing*, v. 4, n. 2, p. 172–179, 1975.

GARCÍA-LORENZO, D. et al. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical Image Analysis*, v. 17, n. 1, p. 1–18, 2013.

GAYEN, A. The frequency distribution of the product moment correlation coefficient in random samples of any size draw from non-normal universes. *Biometrika*, v. 38, p. 219–247, 1951.

GEREMIA, E. et al. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, v. 57, p. 379–390, 2011.

GEROGIANNIS, D.; NIKOU, C.; LIKAS, A. The mixtures of student's t-distributions as a robust framework for rigid registration. *Image and VIsion Computing*, v. 27, n. 1, p. 1285–1294, 2009.

GONZALEZ, R.; WOODS, R. *Digital Image Processing*. Upper Saddle River, USA: Prentice Hall, 2008.

GOVINDARAJAN, K. et al. Effect of in-painting on cortical thickness measurements in multiple sclerosis: a large cohort study. *Human Brain Mapping*, v. 36, p. 3749–3760, 2015.

GREBOL, M. *Atlas-based segmentation of multiple sclerosis lesions in magnetic resonance imaging*. Tese (Doutorado) — Universitat de Girona, 2013.

GRIFFANTI, L. et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, v. 141, p. 191–205, 2016.

HALL-BEYER, M. *GLCM Texture: A Tutorial*. Calgary, Alberta, Canada, 2017.

HALLIDAY, D.; RESNICK, R.; WALKER, J. *Fundamentals of Physics*. Hoboken, USA: Wiley, 2010.

HALME, H.; KORVENOJA, A.; SALLI, E. ISLES (SISS) Challenge 2015: Segmentation of stroke lesions using spatial normalization, random forest classification and contextural clustering. In: *International Workshop on Brainlesion: glioma, multiple sclerosis, stroke and traumatic injuries*. Berlin, Germany: Springer, 2015.

HANSON, L. *Introduction to Magnetic Resonance Imaging Techniques*. Copenhagen, Denmark, 2009. Disponível em: <http://eprints.drcmr.dk/37/>.

HARALICK, R.; SHANMUGAM, K.; DINSTEIN, I. Textural feature for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, v. 3, n. 6, p. 610–621, 1973.

HARTIGAN, J.; WONG, M. Algorithm as 136: a K-Means clustering algorithm. *Journal of the Royal Statistical Society*, v. 28, n. 1, p. 100–108, 1979.

HASHEMI, R. et al. Suspected mutiple sclerosis: MR imaging with a thin-section fast FLAIR pulse sequence. *Radiology*, v. 196, n. 2, p. 505–510, 1995.

HAVEI, M. et al. A convolutional neural network approach to brain tumor segmentation. In: *International Workshop on Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*. Munich, Germany: Springer, 2015. p. 195–208.

HERSKOVITS, E.; ITOH, R.; E.R., M. Accuracy for detection of simulated lesions: comparison of fluid-attenuated inversion-recovery, proton density–weighted, and T2-weighted synthetic brain MR imaging. *American Journal of Roentgenology*, v. 176, n. 5, p. 1313–1318, 2001.

HU, H. et al. Increased signal intensities in the dentate nucleus and globus pallidus on unenhanced T1-weighted images: evidence in children undergoing multiple gadolinium MRI exams. *Pediatric Radiology*, 2016.

HUBBARD, R. Blurring the distinctions between p's and a's in psychological research. *Theory Psychology*, v. 14, n. 3, p. 295–327, 2004.

JAIN, A.; DUIN, R.; MAO, J. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 4–37, 2000.

JUNTU, J. et al. Bias field correction for MRI images. In: ____. Berlin, Germany: Springer Berlin Heidelberg, 2005. (Advances in Soft Computing, v. 30), cap. 4, p. 543–551.

KAMNITSAS, K. et al. Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In: *Proceedings of Ischemic Stroke Lesion Segmentation Challenge, held in conjunction with International Conference on Medical Image Computing and Computer Assisted Intervention*. Munich, Germany: Springer, 2015.

KHADEMI, A.; VENETSANOPOULOS, A.; MOODY, A. Robuts white matter lesion segmentation in FLAIR MRI. *IEEE Transactions on Biomedical Engineering*, v. 59, n. 3, p. 860–871, 2012.

KRINIDIS, S.; CHATZIS, V. A robust fuzzy local information C-means clustering algorithm. *IEEE Transactions on Image Processing*, v. 5, n. 5, p. 1328–1337, 2010.

KROMREY, M. et al. Intravenous injection of gadobutrol in an epidemiological study group did not lead to a difference in relative signal intensities of certain brain structres after 5 years. *European Radiology*, p. 1–6, 2016.

LEITE, M. et al. 3D texture-based classification applied on brain white matter lesions on MR images. In: TOURASSI, S. G. A. I. G. D. (Ed.). *Proceedings of SPIE Medical Imaging 2016*. San Diego, California: SPIE, 2016. v. 9785, n. 97852, p. 97852N1–97852N6.

LEWIS, P. A.; SPILLANE, J. E. Chapter 7 - Multiple Sclerosis. In: LEWIS, P. A.; SPILLANE, J. E. (Ed.). *The Molecular and Clinical Pathology of Neurodegenerative Disease*. Academic Press, 2019. p. 221 – 251. ISBN 978-0-12-811069-0. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780128110690000070>.

LI, C.; GORE, J.; DAVATZIKOS, C. Multiplicative intrinsic component optimization (MICO) for MRI bias field estimation and tissue segmentation. *Magnetic Resonance Imaging*, v. 32, n. 7, p. 913–923, 2014.

LI, S. *Markov Random Field Modeling in Image Analysis*. Berlin, Germany: Springer-Verlag, 2009.

LINDENBAUM, M.; FISCHER, M.; BRUCKSTEIN, A. On Gabor contribution to image enhancement. *Pattern Recognition*, v. 27, n. 1, p. 1–8, 1994.

LITJENS, G. et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, v. 42, p. 60 – 88, 2017. ISSN 1361-8415. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1361841517301135>.

LOIZOU, C. et al. Quantitative texture analysis of brain white matter lesions derived from T2-weighted MR images in MS patients with clinically isolated syndrome. *Journal of Neuroradiology*, v. 2015, n. 42, p. 99–114, 2014.

LU, Q. et al. Cerebra-WML: a stand-alone application for quantification of white matter lesion. In: *Proceedings of Imaging Network Ontario Symposium*. Ontario, Canada: Sunnybrook Research Institute, 2014.

LUDERS, E. et al. A voxel-based approach to gray matter asymmetries. *NeuroImage*, v. 22, p. 656–664, 2004.

MACOVSKI, A. Noise in MRI. *Magnetic Resonance in Medicine*, v. 36, n. 3, p. 494–497, 1996.

MARTIN, D. et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings - IEEE International Conference on Computer Vision*, v. 2, p. 416–423, 2008.

Massachusetss General Hospital. *Data exchange, Center for Morphometric Analysis Massachusetts General Hospital*. 2012. http://www.cma.mgh.harvard.edu/ibsr/data.html. Acesso em 20/03/2015.

MAZZIOTTA, J.; TOGA, A.; EVANS, A. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society*, v. 356, n. 1412, p. 1293–1322, 2001.

MAZZIOTTA, J. et al. A probabilistic atlas of the human brain: theory and rationale for its development the international consortium for brain mapping (icbm). *NeuroImage*, v. 2, n. 2PA, p. 89–101, 1995.

MAZZOLA, A. Ressonância magnética: princípios de formação da imagem e aplicações em imagem funcional. *Revista Brasileira de Física Médica*, v. 3, n. 1, p. 117–129, 2009.

MCLACHLAN, G.; KRISHNAN, T. *The EM algorithm and extensions: series in probability and statistics*. Hoboken, USA: John Wiley & Sons, 1997.

MCLACHLAN, G.; PEEL, D. *Finite Mixture Models*. Hoboken, USA: Wiley, 2000.

MICCAI MS Lesion Segmentation Team. *MS lesion Segmentation challenge*. 2008. http://www.ia.unc.edu/MSseg/index.html. Acesso em 10/02/2015.

MICHOUX, N. et al. Texture analysis of T2-weighted MR images to assess acute inflammation in brain MS lesions. *PLoS One*, v. 12, 2015.

MODAT, M. et al. Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, v. 98, n. 3, p. 278–284, 2010.

MOORE, D.; MCCABE, G. *Introduction to the Practice of Statistics*. 3. ed. London, UK: W.H. Freeman, 1999.

MURPHY, K. *Machine Learning: A Probabilistic Perspective*. Cambridge, USA: The MIT Press Cambridge, 2012.

National Multiple Sclerosis Society. *How many people have multiple sclerosis?* 2013. http://www.nationalmssociety.org/What-is-MS/MS-FAQ-s#question-How-many-people-have-MS.

NGUYEN, T.; WU, Q. Robust student's-t mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, v. 31, n. 1, p. 103–116, 2012.

NOWAK, R. Wavelet-based Rician noise removal for magnetic resonance imaging. *IEEE Transactions on Image Processing*, v. 8, n. 10, p. 1408–1419, 1999.

NUZZO, R. Scientific method: statistical errors. *Nature*, v. 506, n. 7487, p. 150–152, 2014.

NYUL, L.; UDUPA, J.; ZHANG, X. New Variants of a Method of MRI Scale Standardization. *IEEE Transactions on Medical Imaging*, v. 19, n. 2, p. 143–150, 2000.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 7, p. 971–987, 2002.

OKUDA, T. et al. Brain lesions: when should fluid-attenuated inversion-recovery sequences be used in mr evaluation? *Radiology*, v. 212, n. 3, p. 793–798, 1999.

ORLHAC, F.; NIOCHE, C.; BUVAT, I. *Texture - User Guide*. 2019. https://www.lifexsoft.org/images/phocagallery/documentation/ProtocolTexture/UserGuide/TextureUserGui

OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems Man and Cybernetics*, v. 9, n. 1, p. 62–66, 1975.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12, n. 1, p. 2825–2830, 2011.

PEEL, D.; MCLACHLAN, G. Robust mixture modelling using the t distribution. *Statistics and Computing*, v. 10, n. 1, p. 339–348, 2000.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 27, p. 1226–1238, 2005.

PEREIRA, S. et al. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, v. 35, n. 5, p. 1240–1251, 2016.

PERONA, P.; MALIK, J. Scale space and edge detection using anisotropic difusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 12, n. 7, p. 629–639, 1990.

PHAM, D. Spatial models for fuzzy clustering. *Computer Vision Image Understanding*, v. 84, n. 2, p. 285–297, 2001.

PRINSTER, A. et al. Grey matter loss in relapsing-remitting multiple sclerosis: a voxel-based morphometry study. *NeuroImage*, v. 29, p. 859–867, 2006.

QUARANTELLI, M. et al. Brain tissue volume changes in relapsing-remitting multiple sclerosis: correlation with lesion load. *NeuroImage*, v. 18, n. 2, p. 360–366, 2003.

RAJAPAKSE, J.; KRUGGEL, F. Segmentation of MR images with intensity inhomogeneities. *Image and VIsion Computing*, n. 16, p. 3, 165-180 1998.

RAJPOOT, N. Texture classification using discriminant wavelet packet subbands. In: IEEE. *Proceedings of the 2002 IEEE 45th Midwest Symposium on Circuits and Systems*. Tulsa, OK, USA: IEEE, 2002. v. 3, p. III–300–III–303.

ROURA, E. et al. A toolbox for multiple sclerosis lesion segmentation. *Neuroradiology*, v. 57, n. 10, p. 1031–1043, 2015.

ROY, P. et al. Automated segmentation of white matter lesions using global neighbourhood given contrast feature-based random forest and Markov random field. In: *Proceedings of the 2014 IEEE International Conference on Healthcare Informatics*. Verona, Italy: IEEE, 2014. p. 1–6.

ROY, P.; BHUIYAN, A.; RAMAMOHANARAO, K. Automated Segmentation of multiple sclerosis lesion in intensity enhanced FLAIR MRI using texture features and support vector machine. In: *Proceedings of the 20th IEEE International Conference on Image Processing*. Melbourne, VIC, Australia: IEEE, 2013. p. 4277–4281.

RUPPERT, G. et al. A new symmetry-based method for mid-sagittal plane extraction in neuroimages. In: USA. *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Chicago, IL, 2011. p. 285–288.

SAJJA, B. et al. Unified approach for multiple sclerosis lesion segmentation on brain MRI. *Annals of Biomedical Engineering*, v. 34, p. 142–151, 2006.

SCHMIDT, P. et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, v. 59, n. 4, p. 3774–3783, 2012.

SFIKAS, G.; NIKOU, C.; GALATSANOS, N. Robust image segmentation with mixtures of student's t-distributions. In: . San Antonio, USA: IEEE International Conference on Image Processing, 2007. v. 1, p. 273–276.

SHAH, M. et al. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis*, v. 15, p. 267–282, 2011.

SHEN, H.; ZHANG, J.; ZHENG, W. Efficient symmetry-driven fully convolutional network for multimodal brain tumor segmentation. In: *IEEE International Conference on Image Processing (ICIP)*. Beijing, China: IEEE, 2017.

SHIEE, N. et al. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, v. 49, p. 1524–1535, 2010.

SHIEE, N. et al. Robust Reconstruction of the Human Brain Cortex in the Presence of the WM Lesions: Method and Validation. *Human Brain Mapping*, v. 35, n. 7, p. 3385–3401, 2014.

SIMÕES, R. et al. Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. *Magnetic Resonance Imaging*, v. 31, n. 7, p. 1182–1189, 2013.

SOUPLET, J. et al. An automatic segmentation of T2-FLAIR multiple sclerosis lesions. *MIDAS Journal*, p. 1–11, 2008.

STANFORD, D. *Fast automatic unsupervised image segmentation and curve detection in spatial point pattern*. Tese (Doutorado) — Department of Statistics, University of Washington, 1999.

SUETENS, P. *Fundamentals of medical imaging*. 2. ed. The Edinburgh Building, Cambridge CB2 8RU, UK: Cambridge University Press, 2009.

SUKISSIAN, L.; KOLLIAS, S.; BOUTALIS, Y. Adaptive classification of textured images using linear prediction and neural networks. *Signal Process*, v. 36, p. 209–232, 1994.

TITTERINGTON, D.; SMITH, A. *Statistical Analysis of Finite Mixture Distribuitions*. Hoboken, USA: Wiley, 1985.

TIWARI, P. et al. Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: a feasibility study. *American Journal of Neuroradiology*, v. 37, p. 2231–2236, 2016.

TOMAS-FERNANDEZ, X.; WARFIELD, S. A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, v. 34, n. 6, p. 1349–1361, 2015.

Translational Imaging Group. *NiftyReg*. University College London, 2014. Acesso em 20/04/2015. Disponível em: <http://sourceforge.net/projects/niftyreg/>.

TUSTISON, N.; GEE, J. N4ITK: Nick's N3 ITK implementation for MRI bias field correction. *Penn Image Computing and Science Laboratory*, 2009.

UNNIKRISHNAN, R.; PANTOFARU, C.; HEBERT, M. A measure for objective evaluation of image segmentation algorithms. *IEEE Conference on Computer Vision and Pattern Recognition*, v. 3, p. 34–41, 2005.

U.S. Food and Drug Administration. *FDA Drug Safety Communication: FDA evaluating the risk of brain deposits with repeated use of gadolinium-based contrast agents for magnetic resonance imaging (MRI)*. 2015. http://www.fda.gov/Drugs/DrugSafety/ucm455386.htm.

VAIDYA, S. et al. *Longitudinal multiple sclerosis lesion segmentation usign 3D convulutional neural networks*. 2015. Available at: `http://www.iacl.ece.jhu.edu/w/images/1/19/Suthirth_Vaidya.pdf`.

VAISHALI, S.; RAO, K.; RAO, G. S. A review on noise reduction methods for brain MRI images. In: *2015 International Conference on Signal Processing and Communication Engineering Systems (SPACES)*. Guntur, India: IEEE, 2015. p. 363–365.

VALVERDE, S. et al. Evaluating the effects of white matter multiple sclerosis lesions on the volume estimation of 6 brain tissue segmentation methods. *American Journal of Neuroradiology*, v. 36, n. 6, p. 1109–1115, 2015.

VERMA, R. et al. Differentiating enhancing multiple sclerosis lesions, glioblastoma, and lymphoma with dynamic texture parameters analysis (DTPA): A feasibility study. *Medical Physics*, v. 44, n. 8, p. 4000–4008, 2017.

VOVK, U.; PERNUS, F.; LIKAR, B. A review of methods for correction of intensity inhomogeneity in mri. *IEEE Transactions on Medical Imaging*, v. 26, n. 3, p. 405–421, 2007.

WARREN, S.; WARREN, K. *Mutiple sclerosis*. 2001. http://bit.ly/2AbSVqj. Last accessed in November 28th, 2017.

World Health Organization. *Atlas - Multiple Sclerosis Resources in the World*. 2008. http://bit.ly/14Ddglh. Last accessed in November 28th, 2017.

X., L. et al. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated apapproach. *Information Sciences*, v. 186, n. 1, p. 164–185, 2012.

ZHANG, H.; WU, Q.; NGUYEN, T. Image segmentation by a new weighted student's t-mixture model. *The Institue of Engineering and Technology - Image Processing*, v. 7, n. 3, p. 240–251, 2013.

ZHANG, J. The mean field theory in EM procedures for Markov Random Field. *Transactions on Signal Processing*, v. 40, n. 10, p. 2570–2583, 1992.

ZHANG, J. et al. Texture analysis of multiple sclerosis: a comparative study. *Magnetic Resonance Imaging*, v. 26, p. 1160–1166, 2008.

ZHANG, Y.; SMITH, S.; BRADY, M. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, v. 20, n. 1, p. 45–57, 2001.

ZHAO, Y. et al. An energy minimization method for MS lesion segementation from T1-w and FLAIR images. *Magnetic Resonance Imaging*, v. 39, p. 1–6, 2017.

ZIDJDENBOS, A.; FORGHANI, R.; EVAN, A. Automatic "Pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE Transactions on Medical Imaging*, v. 21, n. 10, p. 1280–1291, 2002.

ZOBEL, B. et al. Gadolinium-based contrast agents: did we miss something in the last 25 years? *Radiol Med*, 2016.