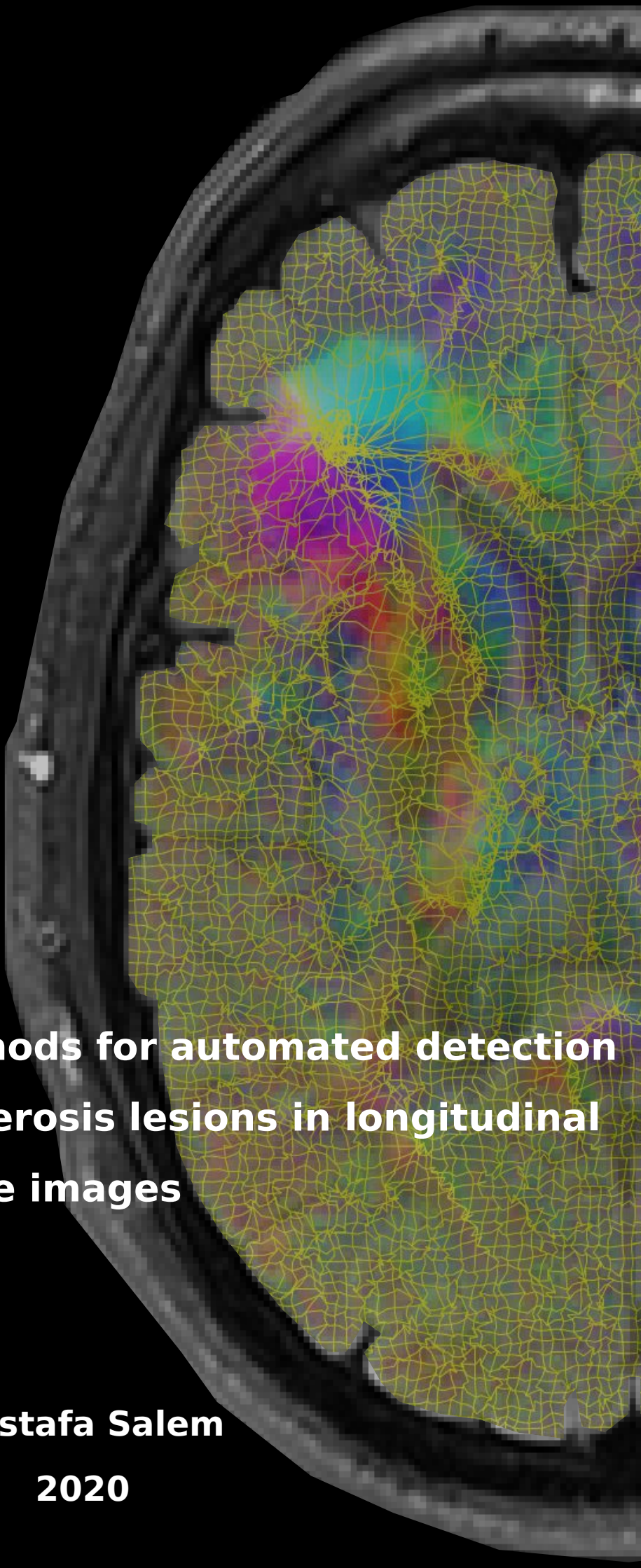
Universitat de Girona



PhD Thesis:

Deep learning methods for automated detection of new multiple sclerosis lesions in longitudinal magnetic resonance images

Mostafa Salem

2020

DOCTORAL THESIS

Deep learning methods for automated detection of new multiple sclerosis lesions in longitudinal magnetic resonance images

Mostafa Salem

2020

DOCTORAL PROGRAM in TECHNOLOGY

Supervised by:
Prof. Joaquim Salvi
Prof. Xavier Lladó

Work submitted to the University of Girona in partial fulfillment of the requirements for the degree of Doctor of Philosophy


قل إن صلاتي ونسكي ومحياي ومماتي لله رب العالمين

To my grandfather, Abdelmajed Salem
To my father, Abobaker A. Salem
To my mother, Sabah A. Mostafa
To my mother-in-law, Samia A. Hussein
To my brothers, Ahmed and Ibrahim
To my sister, Asmaa

To my lovely wife, ♥♥♥Safaa ♥♥♥
To my beloved son, ♥Marawan ♥
To my beloved daughter, ♥Khadija ♥
To my beloved son, ♥Mazen ♥

ACKNOWLEDGMENTS

First and foremost, I would like to thank God Almighty for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

This is the end of my PhD thesis journey. However, this is also just a beginning. This thesis would not have been accomplished without the support and encouragement of numerous people including my supervisors, colleagues, friends, family and various institutions. At the end of my thesis I would like to thank all those who made this thesis possible.

First of all, I am infinitely grateful to my supervisors, Prof. Joaquim Salvi and Prof. Xavier Lladó, for giving me the opportunity to work in this project and for promoting me relying on my work and my commitment with this amazing research group. I really appreciate their enthusiasm, insight, unconditional support and friendship specially when difficulties arose. I owe them a lot for continuously encourage me not to stop forward. Without Joaquim and Xavier, this doctoral thesis would simply not be possible. I am also extremely indebted to Dr. Arnau Oliver. His hypercritical spirit along with his ideas certainly brought this thesis to a higher level. Words fail me to express my appreciation to all of them.

I am indebted with so many colleagues and friends for inspiring me with their work. I would like to thank all the permanent staff, post-doctoral researchers, PhD students and administrative staff of the VICOROB team for their help during these years. I am specially indebted with Dr. Sergi Valverde and Dr. Mariano Cabezas for their great support, expert advices and the valuable discussions about the medical image analysis field. I also want to thank the rest of my office colleagues Richa Agarwal, Sandra González, Kaisar Kushibar, Jose Bernal, Albert Clérigues, Muhammed Habib, and Konstantin Korotkov for their revealed patience and friendship. I would like to mention the great job of the administrative staff Joseta, Mireia and Anna.

Most of the results described in this thesis would not have been obtained without

a close collaboration with few institutions. I want to express my gratitude to the Egyptian Ministry of Higher Education for awarding me with the research grant, which has been used to fund this doctoral thesis. Most of the images used in this work have been gently facilitated by different research hospital centers in Catalunya. Furthermore, I want to thank to Dr. Lluís Ramió-Torrentà, Dr. Hector Perkal, René Robles and Dr. Brigitte Beltrán from the Hospital Dr. Josep Trueta of Girona, Dr. Joan Carles Vilanova from the Hospital Santa Caterina of Girona, Dr. Deborah Pareto and Àlex Rovira from the Vall d'Hebron research hospital of Barcelona and Dr. Jaume Sastre-Garriga from the Multiple Sclerosis Center of Catalunya for their support, continuous reviewing processes and patience answering our medical questions.

After having mentioned all who walked along in the professional field, I would like to thank those who have suffered in the other side, those who have perfectly bear my unreasonable bad attitude at stressed days helping me to fight these adversities, I would like to deeply thank my parents, my brothers and my sister. During these years, I have forgotten the number of times that their love and faith in me have been my light in the darkness. This doctoral thesis is entirely dedicated to them, whose sacrifice has always nourished my dreams.

Last but not least, I wish to thank to Safaa for her advice, support, daily infinite love, patience during the four years and also for her smile that reminds me that this world is a very nice place to stay. Thanks to Marawan, Khadija, and Mazen for the happiness they used to give when returning back from the office after a very hard work day.

PUBLICATIONS

Journals

- **Mostafa Salem**, Mariano Cabezas, Sergi Valverde, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage: Clinical*, 25, 102149, 2020. Quality index: [JCR N IF 3.943, Q1(3/14)].
- **Mostafa Salem**, Sergi Valverde, Mariano Cabezas, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira and Xavier Lladó. Multiple Sclerosis Lesion Synthesis in MRI using an encoder-decoder U-NET. *IEEE Access*. 7, 25171-25185, 2019. Quality index: [JCR CSIS IF 4.098, Q1(23/155)].
- Sergi Valverde, **Mostafa Salem**, Mariano Cabezas, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Joaquim Salvi, Arnau Oliver and Xavier Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21, 101638, 2019. Quality index: [JCR N IF 3.943, Q1(3/14)].
- **Mostafa Salem**, Mariano Cabezas, Sergi Valverde, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage: Clinical*, 17, 607-615, 2018. Quality index: [JCR N IF 3.943, Q1(3/14)].

Conferences

- **Mostafa Salem**, Mariano Cabezas, Sergi Valverde, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. Detecting the appearance of new T2-w multiple sclerosis lesions in longitudinal studies using Deep

convolutional neural networks. *ECTRIMS 2019. Multiple Sclerosis*, September 2019, Stockholm, Sweden. Quality index: [JCR CN IF:5.649 Q1(23/199)].

- **Mostafa Salem**, Mariano Cabezas, Sergi Valverde, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. Lesion synthesis for extending MRI training datasets and improving automatic multiple sclerosis lesion segmentation. *ECTRIMS 2019. Multiple Sclerosis*, September 2019, Stockholm, Sweden. Quality index: [JCR CN IF:5.649 Q1(23/199)].

- Jose Bernal, Kaisar Kushibar, Sergi Valverde, Mariano Cabezas, Sandra González-Villà, **Mostafa Salem**, Joaquim Salvi, Àlex Oliver, Xavier Lladó. Six-month infant brain magnetic resonance image tissue segmentation using multi-atlas segmentation with joint label fusion and convolutional neural networks. *MICCAI Grand Challenge on 6-month infant brain MRI segmentation iSeg-2019. MICCAI 2019*. October 2019. China.

- Sergi Valverde, **Mostafa Salem**, Mariano Cabezas, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, Joaquim Salvi, Xavier Lladó. Manual delineation of only one image in unseen databases is sufficient for accurate performance in automated multiple sclerosis lesion segmentation. *ECTRIMS 2018. Multiple Sclerosis*, October 2018, Berlin, Germany. Quality index: [JCR CN IF:5.649 Q1(23/199)].

- **Mostafa Salem**, Mariano Cabezas, Sergi Valverde, Deborah Pareto, Àlex Rovira, Arnau Oliver, Joaquim Salvi and Xavier Lladó. Supervised detection of newly appearing T2-w multiple sclerosis lesions with subtraction and deformation fields features. *ECTRIMS 2017. Multiple Sclerosis*, October 2017, Paris, France. Quality index: [JCR CN IF:4.840 Q1(27/194)].

- Jose Bernal, **Mostafa Salem**, Kaisar Kushibar, Albert Clèrigues, Sergi Valverde, Mariano Cabezas, Sandra González-Villà, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. MR Brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors. *MR Brain tissue segmentation Challenge in Medical Imaging. MICCAI Workshop*, 2018. 16 September 2018, Granada, Spain.

- Mariano Cabezas, Sergi Valverde, Sandra González-Villà, Albert Clèrigues, **Mostafa Salem**, Kaisar Kushibar, Jose Bernal, Arnau Oliver, Joaquim Salvi and Xavier Lladó. Survival prediction using ensemble tumor segmentation and transfer learning. *Multimodal Brain Tumor Segmentation Challenge 2018 (BRATS) in Medical Imaging. MICCAI Workshop*, 2018. 16 September 2018, Granada, Spain.

- **Mostafa Salem**, Mariano Cabezas, Sergi Valverde, Joaquim Salvi, Àlex Rovira, Xavier Lladó. "Detección supervisada de lesiones activas de esclerosis múltiple usando substracción de imágenes y campos de deformación". *Congreso Nacional Sociedad Española de Radiología Médica (SERAM)*. May 2018, Pamplona, Spain.

- Jose Bernal, Kaisar Kushibar, Sergi Valverde, Mariano Cabezas, Sandra González-Villà, **Mostafa Salem**, Joaquim Salvi, Arnau Oliver, Xavier Lladó. Six-month infant brain tissue segmentation using three dimensional fully convolutional neural networks and pseudo-labelling. *MICCAI Grand Challenge on 6-month infant brain MRI segmentation iSeg-2017. MICCAI 2017.* 14 September 2017, Quebec, Canada.

- Xavier Lladó, Sergi Valverde, Mariano Cabezas, Sandra González-Villà, **Mostafa Salem**, Kaisar Kushibar, Jose Bernal, Jordi Freixenet, Joaquim Salvi, Arnau Oliver. Neuroimatge de la Neurodegeneració: situació actual i futur. *Jornades d'Esclerosis Múltiple del Mediterrani.* 2017, Girona, Spain.

- Sergi Valverde, Mariano Cabezas, Jose Bernal, Kaisar Kushibar, Sandra González-Villà, **Mostafa Salem**, Joaquim Salvi, Arnau Oliver, Xavier Lladó. White matter hyperintensities segmentation using a cascade of three convolutional neural networks. *MICCAI Grand Challenge on White Matter Hyperintensties Segmentation. MICCAI 2017.* 14 September 2017, Quebec, Canada.

ACRONYMS

CNS Central Nervous System
GM Gray Matter
WM White Matter
CSF Cerebrospinal Fluid
MS Multiple Sclerosis
CIS Clinically Isolated Syndrome
RRMS Relapsing Remitting Multiple Sclerosis
SPMS Secondary Progressive Multiple Sclerosis
PRMS Progressive Relapsing Multiple Sclerosis
PPMS Primary Progressive Multiple Sclerosis
RIS Radiologically Isolated Syndrome
MRI Magnetic Resonance Imaging
MR Magnetic Resonance
SNR Signal-to-Noise
CNR Contrast-to-Noise
CDMS Clinically Definite Multiple Sclerosis
DIS Dissemination In Space
DIT Dissemination In Time
DF Deformation Field
CNN Convolutional Neural Network
FCNN Fully Convolutional Network
DNN Deep Neural Network
LR Logistic Regression
RF Radio Frequency
TR Repetition Time
TE Echo Time
SE Spin Echo
GE Gradient Echo
cMRI Conventional Magnetic Resonance Imaging

T1-w T1-weighted
T2-w T2-weighted
PD-w Proton Density-weighted
FLAIR Fluid Attenuated Inversion Recovery
MP-RAGE Magnetization-Prepared Rapid Acquisition with Gradient Echo
DIR Double Inversion Recovery
PSIR Phase-Sensitive Inversion Recovery
HL Hyperintense Lesion
BH Black Holes Lesion
NAWM Normal Appearing White Matter
EL Enhancing Lesion
BET Brain Extraction Tool
BSE Brain Surface Extractor
ROBEX Robust Brain Extraction Tool
BEaST Brain Extraction based on nonlocal Segmentation Technique
DOF Degree-Of-Freedom
FSL FMRIB Software Library
FAST FMRIB's Automated Segmentation Tool
FLIRT FMRIB's Linear Image Registration Tool
WML White Matter Lesion
SuBLIME Subtraction-Based Logistic Inference for Modeling and Estimation
ML Machine Learning
DL Deep Learning
AI Artificial Intelligence
ANN Artificial Neural Network
MLP Multi-Layered Perceptrons
GPU Graphics Processing Units
SVM Support Vector Machine
DBN Deep Belief Net
ReLU Rectified Linear Unit
LeakyReLU Leaky Rectified Linear Unit
MSE Mean Square Error **FC** Fully-Connected
BRATS Brain Tumor Segmentation Challenge
ISLES Ischemic Stroke Lesion Segmentation Challenge
MRBrains Magnetic Resonance Brain Image Segmentation Challenge
GT Ground Truth
TPF True Positive Fraction
FPF False Positive Fraction
DSC Dice similarity coefficient
TP True Positive
FP False Positive
TN True Negative
LR-NDFNB Logistic Regression without Deformation Field without Baseline
LR-NDF Logistic Regression without Deformation Field
LR-DFNB Logistic Regression with Deformation Field without Baseline

LR-DF Logistic Regression with Deformation Field
WMH White Matter Hyperintensity
MSE Mean Square Error
SSIM Structural Similarity Index
DA Data Augmentation
GAN Generative adversarial network

LIST OF FIGURES

LIST OF TABLES

CONTENTS

ABSTRACT

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system, which is characterized by the presence of lesions in the brain and the spinal cord. Magnetic resonance imaging (MRI) has become a core para-clinical tool for diagnosing and predicting long-term disability and treatment response in MS patients. It has been accepted that dissemination in time can be demonstrated by a new T2 or gadolinium-enhancing lesion(s) in follow-up MRI, with reference to a baseline scan. The manual longitudinal detection of change is not only time-consuming, but is also prone to intra- and inter-observer variability. Therefore, a reliable and robust automatic detection and quantification of these lesions could be used to help neuroradiologists to improve the diagnosis and follow-up evaluation of MS patients.

The main goal of this PhD thesis is to develop novel and fully automated methods for the detection of new MS lesions in longitudinal brain MRI. In order to fulfill this goal, firstly, we analyzed and evaluated the state-of-the-art on MS lesion detection approaches. Our analysis showed that the tissue transformation, which is the effect of a lesion that does not always appear as an intensity change on the tissue where it is located, can also influence the appearance of surrounding tissues (tissue deformation). Moreover, we observed the importance of using prior knowledge to guide the lesion detection and segmentation. Supervised approaches that rely on similar segmented cases usually outperform unsupervised strategies. In the second stage, a novel fully automated logistic regression (LR) based framework has been proposed and evaluated for the detection and segmentation of new T2-w lesions. The framework was based on intensity subtraction and deformation field (DF). The DF were obtained using the multi-resolution Demons registration approach from ITK v.4. In the third stage, we focused on the use deep learning (DL) techniques, which simplify the feature extraction process, and could gather unknown patterns to help in the desired task. We proposed a fully convolutional neural network (FCNN) approach to detect new T2-w lesions in longitudinal brain MR images. The model was trained end-to-end and simultaneously learned both the DFs and the new T2-w lesions. We qualitatively and quantitatively evaluated the proposed methods (DL-based and

LR-based) using an in-house clinical dataset from our collaborating hospitals and compared it with other state-of-the-art methods. Finally, we proposed and evaluated a deep learning based approach for MS lesion synthesis. The proposed pipeline can generate synthetic images with MS lesions. We used the generated synthetic MS lesion images as data augmentation to improve the lesion detection and segmentation performance in both cross-sectional and longitudinal analysis.

RESUMEN

La Esclerosis Múltiple (EM) es una enfermedad inflamatoria del sistema nervioso central, que se caracteriza por la presencia de lesiones en el cerebro y la médula espinal. La resonancia magnética (RM) se ha convertido en una herramienta paraclínica central para diagnosticar y predecir la discapacidad a largo plazo y la respuesta al tratamiento en pacientes con EM. Se ha aceptado que la diseminación en el tiempo puede ser demostrada con la aparición de una nueva lesión, o lesiones que mejoran con gadolinio en la RM de seguimiento, con referencia a una exploración basal del paciente. La detección manual del cambio en estudios longitudinales no sólo conlleva mucho tiempo, sino que también es propensa a la variabilidad intra e interobservador. Por lo tanto, una detección y cuantificación automática fiable y robusta de estas lesiones podría utilizarse para ayudar a los neurorradiólogos a mejorar el diagnóstico y la evaluación del seguimiento de los pacientes con EM.

El objetivo principal de esta tesis doctoral es desarrollar métodos novedosos y totalmente automatizados para la detección de nuevas lesiones de EM en estudios longitudinales de resonancia magnética cerebral. Para cumplir este objetivo, en primer lugar, analizamos y evaluamos el estado del arte de los métodos de detección de lesiones de EM. Nuestro análisis mostró que la transformación del tejido, que es el efecto de una lesión que no siempre aparece como un cambio de intensidad en el tejido donde se encuentra, también puede influir en la apariencia de los tejidos circundantes (deformación tisular). Además, observamos la importancia de utilizar conocimientos previos para guiar la detección y segmentación de las lesiones. Los enfoques supervisados que se basan en casos segmentados similares suelen superar a las estrategias no supervisadas. En la segunda etapa, se propuso y evaluó un nuevo método basado en la regresión logística totalmente automatizada (LR) para la detección y segmentación de nuevas lesiones T2-w. La propuesta se basaba en la intensidad de la sustracción de las imágenes y el campo de deformación (DF). Los DF se obtuvieron utilizando la técnica de registro de *Demons* de ITK v.4. En la tercera etapa, nos centramos en el uso de técnicas de aprendizaje profundo (DL), que simplifican el proceso de extracción de características, y que permiten encontrar

patrones desconocidos para ayudar en la tarea deseada. Propusimos un enfoque de red neural completamente convolucional (FCNN) para detectar nuevas lesiones T2-w en las imágenes longitudinales de RM cerebral. El modelo fue entrenado de principio a fin y simultáneamente aprendió tanto los DFs como las nuevas lesiones T2-w. Evaluamos cualitativa y cuantitativamente los métodos propuestos (basados en DL y LR) utilizando un conjunto de datos clínicos internos de nuestros hospitales colaboradores y los comparamos con otros métodos de última generación. Finalmente, propusimos y evaluamos un enfoque basado en el aprendizaje profundo para la síntesis de lesiones de EM. La propuesta realizada permite generar imágenes sintéticas con lesiones de EM. Se utilizaron las imágenes sintéticas generadas de lesiones de EM para aumentar los datos de entrenamiento y mejorar así la detección y la segmentación de las lesiones, tanto en análisis transversales como longitudinales.

RESUM

L'Esclerosi Múltiple (EM) és una malaltia inflamatòria del sistema nerviós central, que es caracteritza per la presència de lesions al cervell i a la medul·la espinal. La ressonància magnètica (RM) s'ha convertit en una eina paraclínica bàsica per al diagnòstic i predicció de la discapacitat a llarg termini i la resposta al tractament en pacients amb EM. S'ha acceptat que la difusió en el temps es pot demostrar amb una nova lesió o lesions de gadolini observades en la RM de seguiment, en referència a una exploració basal. La detecció manual del canvi longitudinal no només requereix de temps, sinó que també és propensa a la variabilitat intra i interobservador. Per tant, es podria utilitzar una detecció i quantificació automàtica fiable i robusta d'aquestes lesions per ajudar els neuroradiòlegs a millorar el diagnòstic i l'avaluació del seguiment dels pacients amb EM.

L'objectiu principal d'aquesta tesi de doctorat és desenvolupar mètodes nous i totalment automatitzats per a la detecció de noves lesions d'EM en imatges longitudinals de RM del cervell. Per assolir aquest objectiu, primer hem analitzat i avaluat l'estat de l'art dels mètodes de detecció de lesions d'EM. L'anàlisi realitzat va demostrar que la transformació del teixit, que és l'efecte d'una lesió que no apareix sempre com a canvi d'intensitat al teixit on es troba, també pot influir en l'aparició de teixits circumdants (deformació del teixit). A més, es va observar la importància d'utilitzar coneixements a priori per guiar la detecció i segmentació de lesions. Els enfocaments supervisats que es basen en casos segmentats similars solen superar les estratègies no supervisades. En la segona etapa, s'ha proposat i avaluat una nova proposta basada en una regressió logística (LR) per a la detecció i segmentació de noves lesions T2-w. El proposta es basava en les intensitats de la resta d'imatges i amb el camp de deformació (DF). Els DF es van obtenir mitjançant la tècnica de registre *Demons* de resolució múltiple d'ITK v.4. En la tercera etapa, ens vam centrar en l'ús de tècniques d'aprenentatge profund (DL), que simplifiquen el procés d'extracció de característiques i poden descobrir patrons desconeguts per ajudar en la tasca desitjada. Es va proposar un mètode basat en una xarxa neuronal completament convolucional (FCNN) per detectar noves T2-w lesions en imatges de

RM longitudinals del cervell. El model es va entrenar *end-to-end* i alhora es van aprendre tant els camps de deformació (DF) com les noves lesions T2-w. Es van avaluar qualitativament i quantitativament els mètodes proposats (basats en DL i basats en LR) mitjançant un conjunt de dades clíniques internes dels nostres hospitals col·laboradors i ho vam comparar amb altres mètodes actuals de l'estat de l'art. Finalment, es va proposar i avaluar una tècnica basada en l'aprenentatge profund per a la síntesi de lesions d'EM. El pipeline proposat pot generar imatges sintètiques amb lesions d'EM. Hem utilitzat les imatges generades de lesions sintètiques d'EM com a augmentació de les dades d'entrenament per així millorar el rendiment de detecció i segmentació de lesions tant en estudis transversals com longitudinals.

CHAPTER 1

INTRODUCTION

1.1 Multiple sclerosis

1.1.1 What is multiple sclerosis?

The central nervous system (CNS) and the peripheral nervous system are the two parts of the human nervous system. The CNS consists of the brain and the spinal chord, and the peripheral nervous system connects the CNS with the sense organs [1]. CNS is mainly composed of two tissues: gray matter (GM), which consists of neuronal cell bodies, and white matter (WM) tissue, which is mainly composed of myelinated axon tracts [2]. The brain itself is composed mostly of GM and WM, both surrounded by the cerebrospinal fluid (CSF), which provides basic mechanical and immunological protection to the brain inside the skull [2].

Multiple sclerosis (MS) is the most common chronic immune-mediated disabling neurological disease of the central nervous system. Nowadays, MS is the most frequent nontraumatic neurological disease causing the most disability in young adults. It follows a similar behavior to other putative autoimmune diseases [3]. It has a low incidence in childhood, but the probability increases rapidly in young adulthood reaching a peak between 25 and 35 years, and then slowly declines, becoming rare at 50 and older [4]. Recent epidemiological studies show that 2.3 million people have been diagnosed with MS worldwide, of which almost three times more women than men are affected. The causes are still unknown, but interaction with multiple genetic and as-yet-unidentified environmental factor(s) are potential candidates [5]. Moreover, geographic studies show the prevalence of MS around the world (see Figure 1.1), affirming that migration from high to lower-prevalence in areas before the age of 15, reduces the likelihood of developing MS. Looking at the map in Figure 1.1, we can see that Europe, the United States, Canada, New

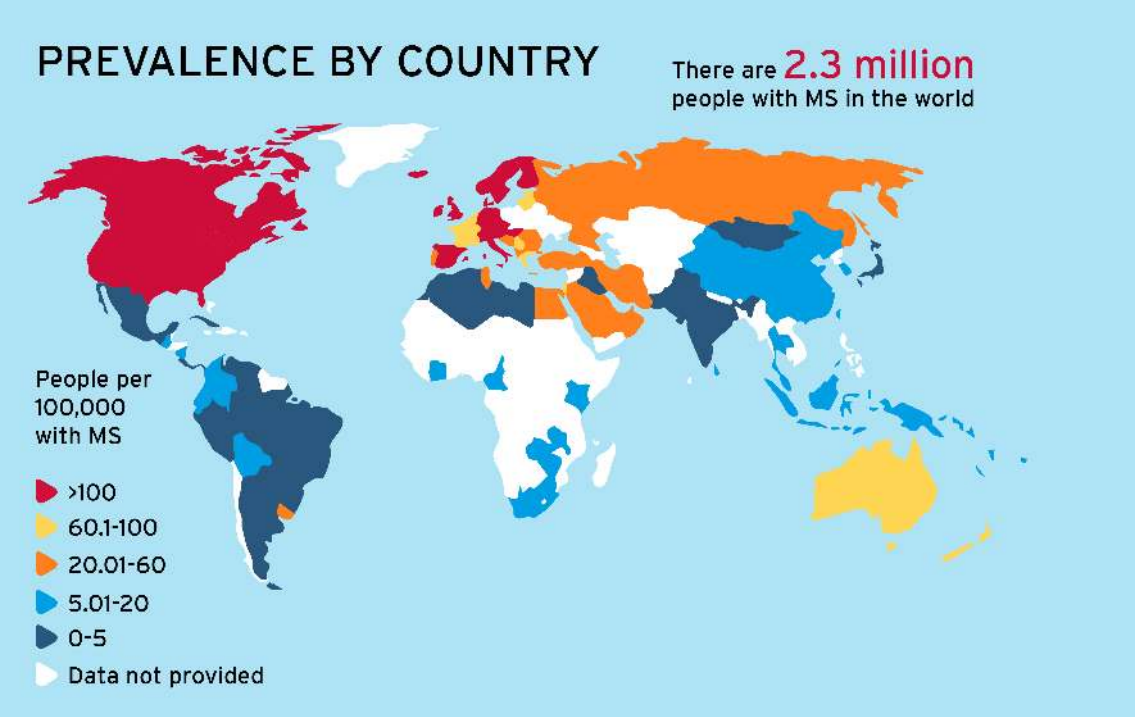


Figure 1.1: MS prevalence around the world. Image extracted from http://www.oysterhc.co.uk/blogs/ms-horrible/ (accessed 01.09.2019).

Zealand, and sections of Australia have more MS sufferers than Asia and the tropics.

Pathologically, MS is an inflammatory-demyelinating and neurodegenerative disease, clinically defined by demyelinating lesions and characterized by areas of inflammation, demyelination (i.e. damage of the myelin), axonal loss, and gliosis scattered throughout the CNS [6, 7]. Therefore, demyelination in the brain and spinal cord leads to a disruption of the communication within the brain and between the brain and the body. Partially demyelinated axons can cause delay and demyelinated axons can discharge spontaneously. Affecting different sites within the brain or spinal cord, depending on the site, MS can cause cognitive impairment, painful loss of vision, tremors, clumsiness and poor balance, vertigo, impaired speech and swallowing, weakness, stiffness and painful spasms, bladder dysfunction as well as many other impairments [6].

1.1.2 MS phenotypes and clinical course

MS takes several forms, with new symptoms either occurring in isolated attacks (relapsing forms) or building up over time (progressive forms). The initial presentation of the disease varies according to both the location of the lesions and the type of symptom onset (relapsing or progressive). The majority of patients who develop MS begin with a single episode, called clinically isolated syndrome (CIS), that involves the optic nerve, brainstem, or spinal cord, and resolves over time.

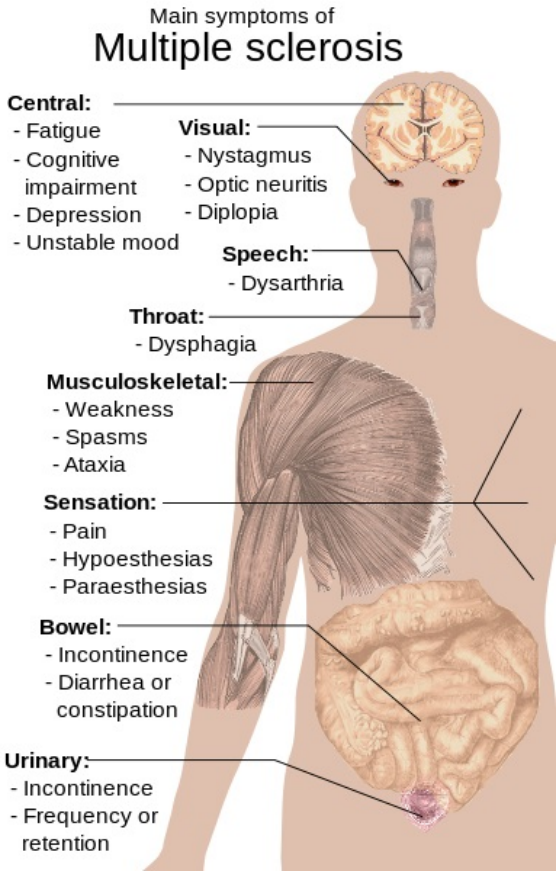


Figure 1.2: Main symptoms of MS. Image extracted from http://neurosciencenews.com/neurology-ms-gray-matter-393/ (accessed 01.09.2019).

Patients that have at least two relapses are described as having Relapsing Remitting MS (RRMS). RRMS is characterized by exacerbation times where symptoms are present. These periods are followed by periods of remission, where the patient recovers partially or totally from the disease's symptoms. Sufferers are relatively symptom-free for periods of time that are interrupted by attacks that can put them in hospital for weeks, or even months, at a time. These attacks worsen the symptoms (see Figure 1.2) and are followed by full, partial, or no recovery of some function or another. The interval between relapses varies, there can be many years between the first manifestation and the first relapse. On average, 65% of people with RRMS develop secondary-progressive MS (SPMS), this progressive part may begin shortly after the onset or may occur even decades later.

The last two types of this condition are less common and usually affect people who develop MS after age 40. The progressive remitting (PRMS) form is typified by an increase in the relapse times with significant recovery but with worsening symptoms in new relapse intervals. The Primary Progressive (PPMS) form is characterized by a severe decrease of remission times with special localization in the brain. Moreover, patients with incidental MRI findings consistent with MS are classified as suffering from radiologically isolated syndrome (RIS). A third of patients

(a) (b) (c)

Figure 1.3: MRI scanners. The most common MRI scanners are (a) Philips , (b) Siemens and (c) General Electric.

with RIS will develop clinical symptoms of MS within 5 years of follow-up, either a relapse or progressive symptoms [8].

As a result, diagnosing and monitoring the progression of this disease is vital for MS patients. In this sense, in order to improve the quality of the diagnostic assessment and to provide a rapid and sensitive measure of treatment, magnetic resonance imaging (MRI) techniques have been widely used and have become a key tool for clinical purposes.

1.2 Magnetic resonance imaging

MRI is a noninvasive medical imaging technique used in radiology to generate image representations of different internal anatomical organs and physiological processes of the body. MRI scanners use strong magnetic fields and radio waves to acquire the 3D images without the use of damaging radiation. Over the last 40 years, MRI has evolved as a clinical modality [9], and, in particular, as an essential tool for the diagnosis and evaluation of CNS disorders such as MS [10], due to the high specificity and sensitivity visualization of structural MRI for the dissemination of WM lesions in time and space, which is a key factor in recent diagnostic criteria [11, 12]. There are different brands for MRI scanners, being the most common Philips, Siemens and General Electric. Figure 1.3 depict three MRI scanners from the 3 different brands. Types of MRI scanners can also be differentiated by their magnetic field strength (teslas (T)). Scanners use a magnet strength that can range from 0.5T up to 7.0T. A 1.5T MRI provides good image quality, fast scan times, and the evaluation of how specific structures in the body function. It is the most standard nowadays for the MS diagnosis. The 3.0T MRI scanner is ideal for visualizing very fine detail such as brain and heart vessels. With its higher signal-to-noise (SNR) and contrast-to-noise ratio (CNR) compared to lower field strengths, 7.0T MRI has current applications in the field of brain MRI, in clinical studies as well as clinical practice [13].

Brain MRI consists of a 3D model of the brain, which can be acquired in three different orientations: axial, coronal, and sagital (see Figure 1.4). The resolution

is given by a 3×3 matrix, each axis being one of its orientation. Axis z denotes the number of slices, which are 2D images of $x \times y$. Even though the final image resolution is given in voxels, the voxel spacing is described in mm. The voxel is isotropic if it has equal distance in all three directions. For example, isotropic images can be obtained at $0.5 \times 0.5 \times 0.5$ mm^3 or $1.0 \times 1.0 \times 1.0$ mm^3, meaning that each voxel in the image represents 0.125 mm^3 and 1.0 mm^3 of the real brain and there is 0.5 mm and 1.0 mm between each slice, respectively. Instead, anisotropic images can be acquired at $1.0 \times 1.0 \times 3.0$ mm^3, thus the gap between slices will be 3.0 mm, so the resolution would be lower, 3.0 mm^3 per voxel. A whole brain is illustrated in Figure 1.4, which has been acquired at 1.5T scanner in the axial orientation with a resolution of $1.0 \times 1.0 \times 3.0$ mm^3, i.e. 240×320 by 46 slices. In the chapter 2, we will explain in more details the MRI technology.

With MRI it is possible to detect contrast differences in soft tissues [14]. Additionally, it has been demonstrated that MRI is highly sensitive for detecting MS plaques. MRI techniques play a pivotal role in both diagnosing and monitoring the progression of MS and is used as a surrogate marker of drug efficacy in treatment trials [15]. For instance, as a CIS is an individual's first neurological episode caused by inflammation or demyelination of nerve tissue, MRI helps to confirm the diagnosis of MS after the second validated clinical event (clinically definite MS (CDMS)) and differential diagnosis with other neurological diseases [15, 16]. Moreover, a number of trials have reported that MRI is useful in monitoring early treatment of MS and offers an opportunity to reduce the disease's activity, slowing disability progression [17]. Consequently, MRI-derived metrics have become the most important paraclinical tool in diagnosing MS and in understanding the natural history of the disease as well as monitoring the efficacy of experimental treatments [15, 17, 18].

1.3 Longitudinal brain MRI analysis for MS

In practice, it is necessary to integrate the clinical, imaging, and laboratory findings for providing a definite diagnosis of MS. The diagnosis of MS requires objective evidence of CNS lesions disseminated in space (cross-sectional analysis, i.e. MS lesion segmentation in a single time-point) and time (longitudinal analysis, i.e. MS lesion segmentation between successive time points). In 2001, a panel of experts on MS came up with a set of diagnostic criteria that included MR images for the first time to provide evidence of CNS lesions [19]. The so called McDonald criteria have undergone several revisions in recent years [15, 20, 21], with increased certainty with successive versions, and have become the gold standard test for MS diagnosis.

The last revision (2017) defines lesion dissemination as follows:

- **Dissemination in space (DIS):** demonstrated by one or more lesions that are characteristic of MS in two or more of four areas of the CNS (periventricular, cortical or juxtacortical, and infratentorial brain regions, and the spinal cord).

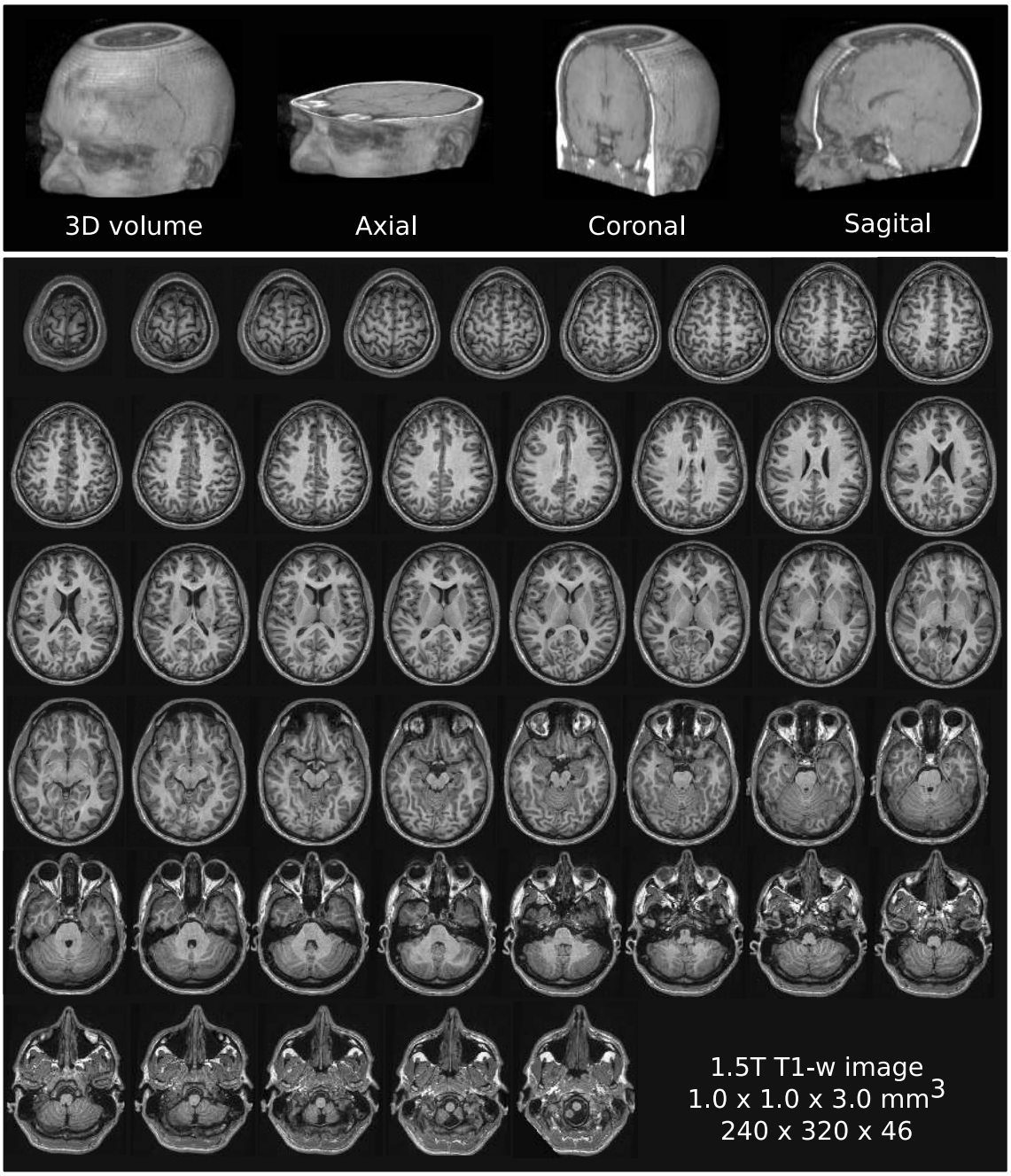


Figure 1.4: Brain MRI representation. The first row illustrates the 3D volume and its 3 different orientations (axial, coronal, and sagital respectively from left to right). The figure illustrates the volume's 46 slices in the axial orientation.

- **Dissemination in time (DIT):** demonstrated by the simultaneous presence of gadolinium-enhancing and nonenhancing lesions at any time or by a new lesion on follow-up MRI, with reference to a baseline scan, irrespective of the timing of the baseline MRI.

Consequently, it has been accepted that DIT can be demonstrated by either the simultaneous presence of asymptomatic gadolinium-enhancing and nonenhancing lesions (types of MS lesions will be explained in more detail in 2.1) in any MRI scan or in those patients who do not meet this criteria, a new T2 or gadolinium-enhancing lesion(s) in follow-up MRI, with reference to a baseline scan (longitudinal MRI analysis). Figure 1.5 depicts McDonald's diagnostic criteria example. The manual longitudinal detection of change is not only time-consuming, but is also prone to intra- and inter-observer variability. Therefore, with the need of using MRI-derived metrics, a reliable and robust automatic detection and quantification of these lesions could be used to diagnose the disease according to McDonald's criteria and to help neuroradiologists to improve the diagnosis and follow-up evaluation of MS patients.

1.4 Research background

The Computer Vision and Robotics group (VICOROB) of the University of Girona has been working on medical image analysis since 1996, mainly in the segmentation and registration of mammographic images. In 2009, the group started collaborating with several medical teams experts in MS, with the objective of developing new tools that could be transferred for clinical use. Thanks to the group prior knowledge acquired through previous medical projects, a new line of research emerged, focused on brain MRI analysis and the extraction of brain MRI biomarkers. This new line started with the segmentation of MS lesions and has expanded to other fields such as temporal analysis, registration (temporal and inter-subject), tissue segmentation, atrophy analysis and brain structure segmentation.

All these studies have been accomplished inside the framework of several research projects:

1. [2015 - 2017] NICOLE: "Herramientas de neuroimagen para mejorar el diagnosis y el seguimiento clínico de los pacientes con Esclerosis Múltiple". Awarded in 2014 by the spanish call Retos de investigación 2014. Ref: TIN2014-55710-R.

2. [2015 - 2019] BiomarkEM.cat: "New technologies applied to clinical practice for obtaining biomarkers of atrophy and lesions in magnetic resonance images of patients with multiple sclerosis". Awarded in 2015 by the Fundació la Marató de TV3.

3. [2018 - 2020] EVOLUTION: "Predictive models for multiple sclerosis using brain magnetic resonance imaging biomarkers". Awarded in 2017 by Ministerio de ciencia y tecnologia. RETOS 2017. Ref: DPI2017-86696-R.

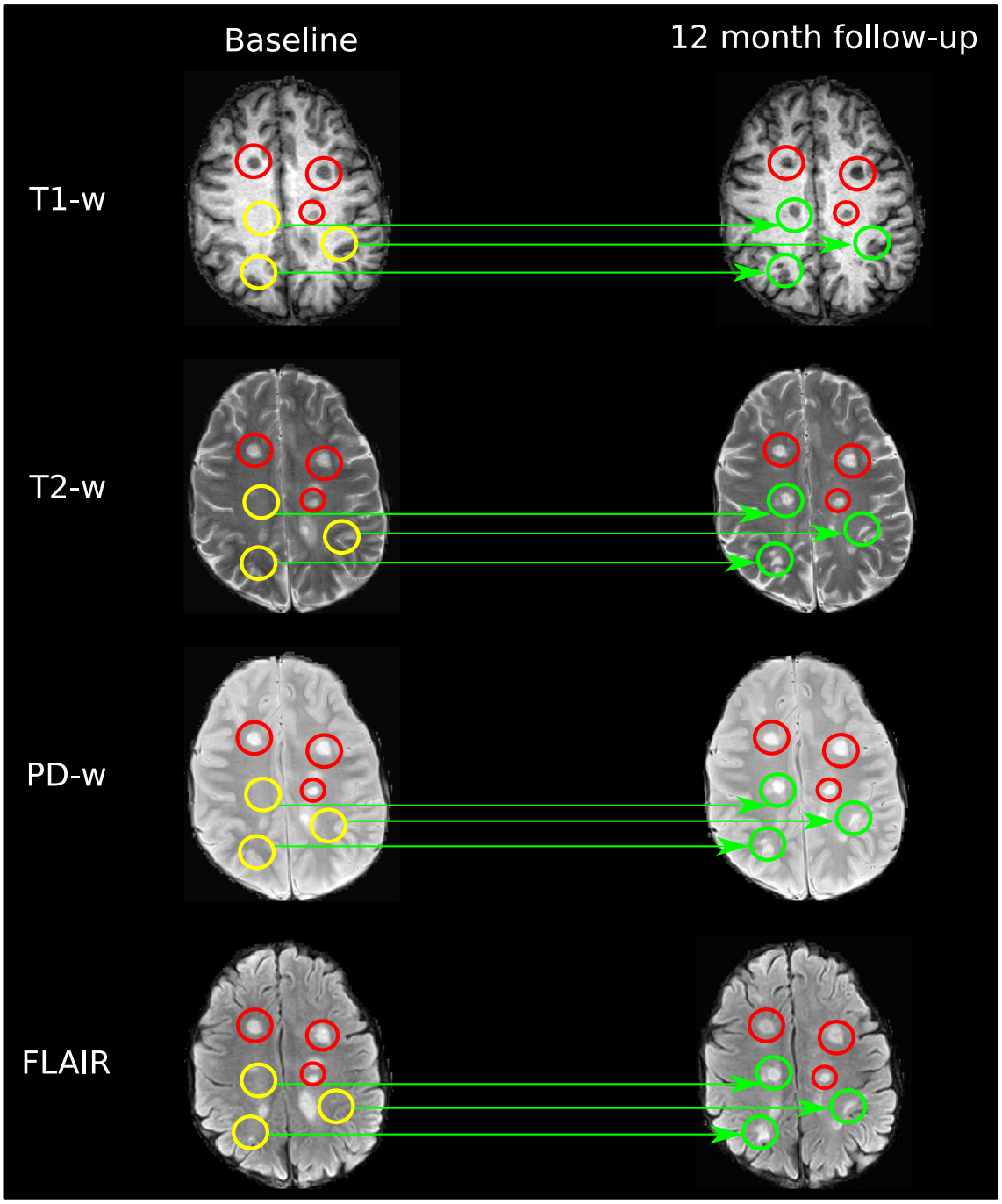


Figure 1.5: McDonald's diagnostic criteria example. Demonstration of dissemination in space (red circles in baseline and follow-up scans) and dissemination in time (yellow circles in the baseline scan and green circles in the follow-up scan).

Since then, the research group has published original contributions in different fields such as image preprocessing [22, 23], MS lesion segmentation [24, 25, 26, 27, 28, 29, 30], temporal analysis [31, 32, 33, 34], image registration [35, 36], and tissue segmentation [37, 38, 39, 40]. All the projects have been carried out in collaboration with different medical MS teams from:

- The Hospital Vall d'Hebron: Dr. Rovira, who is the director of the "Unitat de Ressonància Magnètica-Centre Vall d'Hebron" (URMVH) and has participated in numerous research projects funded by public and private institutions in the last years, as well as Dr. Pareto and technicians Huerga and Corral. This group is part of the MAGNetic resonance Imaging in MS (MAGNIMS) network, a European network of centers that share an interest in the MS study through MRI.
- The Hospital Josep Trueta: Dr. Ramió-Torrentà, who is the current coordinator of the "Unitat de Neuroimmunologia i Esclerosi Múltiple", as well as Dr. Robles and Dr. Beltrán, who work in the neurology and radiology units, respectively.
- The Clínica Girona / Hospital Santa Caterina: Dr. Vilanova and Dr. Barceló are the codirectors of the "Unitat de Ressonància Magnètica" at the Clínica Girona and are members of several national and international radiology societies.

1.5 Objectives

As part of the NICOLE, BiomarkEM.cat, and EVOLUTION research project frameworks, the main goal of this thesis is:

to develop novel and fully automated methods for the detection of new T2-w lesions in longitudinal MR images of multiple sclerosis patients.

Different sub-objectives have to be covered first in order to fulfill the main goal. All these stages can be considered as sub-objectives that allow us to gain a better knowledge of the problem that we want to overcome. In what follows, we detail these proposed sub-goals:

- **to propose and evaluate a fully automated supervised framework with intensity subtraction and deformation field for the detection of new T2-w lesions in multiple sclerosis**. This stage aims to propose a fully automated method for detecting new T2-w MS lesions. We aim to study how to incorporate the deformation field (DF) information extracted from a registration together with features extracted from intensity subtraction.

We plan to merge intensity- and deformation-based features in an automated multichannel supervised logistic regression (LR) classification. In contrast with the previous supervised approaches, we aim to use features not only from the baseline, follow-up, and subtraction images but also from the DF operators obtained from the nonrigid registration between longitudinal scans. In this stage, we aim to validate the accuracy of the proposed method using an in-house clinical dataset from our collaborating hospitals and comparing it with the state of the art in longitudinal MS detection.

- **to propose and evaluate a deep learning (DL) based approach for the detection of new T2-w lesions in multiple sclerosis**. This stage aims to propose fully automated deep learning based method for detecting new T2-w MS lesions. With this approach, we aim to avoid the definition of hand-crafting feature vectors to extract appearance information. A convolutional neural networks (CNNs) will learn a set of features that are specifically optimized for the lesion segmentation task directly from the image data. In this stage, we will qualitatively and quantitatively evaluate the proposed method using an in-house clinical dataset from our collaborating hospitals and comparing it with the state-of-the-art methods including our previous proposal using DF operators and the LR model.

- **to propose and evaluate a deep learning based approach model for MS lesion synthesis in MRI to improve the performance of cross-sectional and longitudinal MS lesion segmentation and detection approaches**. This goal aims to propose a deep learning based pipeline that will able to generate synthetic images with MS lesions. Our objective is to tackle one of the main limitations of deep learning methods which is the lack of available ground-truth data needed for the supervised MS lesion detection and segmentation strategies. Therefore, the generated synthetic MS lesion images could be used as data augmentation to improve the cross-sectional and longitudinal lesion detection and segmentation performance. For cross-sectional analysis, this will be done by synthesizing the lesions in new brain images, coming from either healthy subjects or from patients with lesions. For longitudinal analysis, this will be done by MS lesions could be added only to the follow-up scans keeping the baseline images untouched. We will qualitatively and quantitatively evaluate the proposed pipeline using MS patient data from an in-house clinical dataset, the public MICCAI 2016 challenge dataset, and the public ISBI2015 challenge dataset.

1.6 Document structure

The rest of this thesis is structured as follows:

- **Chapter 2. Thesis background.** After stating the problem in chapter 1, we present a general background about the main topics of this thesis. The chapter

is divided in 3 main sections covering details of MRI, brain image analysis methods in MS, and some general machine learning concepts for medical image analysis.

- **Chapter 3. A logistic regression model for new T2-w lesion detection in multiple sclerosis.** In this chapter, we present a new method for detecting new T2-w MS lesions. The method is a supervised framework with intensity subtraction and deformation field features. The method is evaluated qualitatively and quantitatively and compared with the state-of-the-art methods. Moreover, we analyze the performance of the method according to the different lesion sizes, and also the specificity of the method with patients with no new T2-w lesions.

- **Chapter 4. A deep learning model for new T2-w lesion detection in multiple sclerosis.** In this chapter, we propose a fully CNN approach to detect new T2-w lesions in longitudinal brain MR images. The proposed model combines intensity-based and deformation-based features within an end-to-end deep learning approach. The DFs and the new T2-w lesions are learned simultaneously using a combined loss function. The method is evaluated qualitatively and quantitatively and compared with the state-of-the-art methods. Moreover, we demonstrate the contribution of simultaneously learning both the DF and the segmentation of new T2-w lesions. We analyze the performance of the method according to the different lesion sizes, and also the specificity of the method with patients with no new T2-w lesions. We analyze also the performance of the proposed model and the state-of-the-art approaches on different brain regions. The analysis of the new MS lesion detection was divided into 4 types (periventricular, juxtacortical, infratentorial, and deep white matter) according to its location in the brain. Finally, we analyze the generalization and the performance of the proposed approach when tested in images from a different scanner and image acquisition protocol.

- **Chapter 5. Multiple sclerosis lesion synthesis on magnetic resonance imaging.** In this chapter, we propose generating synthetic MS lesions on MR images with the final aim to improve the performance of supervised machine learning algorithms. The pipeline is evaluated cross-sectionally and longitudinally on MS patient data from an in-house clinical dataset, the MICCAI 2016 challenge dataset, and the public ISBI 2015 challenge dataset. For cross-sectional analysis, the evaluation is based on measuring the similarities between the real and the synthetic images as well as in terms of lesion detection performance by segmenting both the original and synthetic images individually. Moreover, we demonstrate the usage of synthetic MS lesions generated on healthy images as data augmentation. We also analyze a scenario of limited training data (one-image training) to demonstrate the effect of the data augmentation. For longitudinal analysis, we present how to generate longitudinal synthetic datasets by generating cross-sectional MS lesion masks on only to the follow-up scans of the longitudinal datasets with no new lesions

keeping the baseline images untouched. Moreover, we use these longitudinal synthetic datasets for training the supervised MS lesion change detection and segmentation method demonstrating an increase in the performance of this model when using them as data augmentation.

- **Chapter 6. Conclusions and future work.** Lastly, the main conclusions based on the contributions of this thesis are presented. Moreover, we also point out different future investigations to improve and extend the work carried out for this PhD thesis.

CHAPTER 2

THESIS BACKGROUND

2.1 Magnetic resonance imaging in MS

2.1.1 MRI in details

In the chapter 1 we briefly described the importance of MRI. In what follows, we will go deeper into the acquisition process, the components of an MRI system, how MRI works, MRI parameters, and the obtained MRI sequences.

What is MRI?

The human body is composed of molecules that contain nuclei (or protons). MRI scanners make use of the electromagnetic activity of atomic nuclei and use strong magnetic fields and radio-waves in order to form images of the body. This is possible due to the fact that a large proportion of the human body is made up of fat and water, both of which contain lots of hydrogen atoms. The hydrogen nuclei are made up of protons and neutrons, both of which spin around their own axis, this motion induces a magnetic field. When no external magnetic field is applied, their axes are randomly aligned until they are exposed to an external magnetic field. The interaction between the two magnetic fields urges the nuclei to align with the magnetic field, and this movement creates magnetic moments. Tissues can be distinguished from each other by examining the sum of all the magnetic moments called, the net magnetization vector. For this purpose, a radio frequency (RF) that matches the center frequency of the system is applied to the net magnetization vector (resonance matching) [14].

By sending an RF pulse to the center frequency, with a certain strength (amplitude) and for a certain period of time, it is possible to flip the net magnetization

by any degree (flip angle) in the range from 1° to 180° (lifting the protons into a higher energy state), which is called the RF excitation process. However, as the protons would rather be in a low energy state, when the RF energy source is turned off, the net magnetization vector realigns with the axis of the external magnetic field. Realigning with the magnetic field simultaneously and independently, the longitudinal magnetization increases or recovers (T1 recovery, T1 relaxation or the so-called Spin-Lattice relaxation) and the transverse magnetization decreases or decays (T2 and T2* decays, T2 relaxation or the so-called Spin-Spin relaxation). Note that various tissues have different relaxation times that make them distinguishable. During the relaxation processes, the spins shed their excess energy in the shape of radio frequency waves. In order to produce an image, these waves are caught by a receiving coil positioned at right angles to the main magnetic field [14].

As shown in Figure 2.1, an MRI system consists of the following components:

- **Magnet:** the magnet is the most expensive part of the whole scanner. It is used to generate the magnetic field and shim coils to make the magnetic field as homogeneous as possible. This magnetic field aligns the hydrogen nuclei of the brain.
- **Gradient Coils:** it is used to provide spatial localization of the signals applying additional magnetic fields. These additional magnetic fields can be used to only generate detectable signals from specific locations in the body (spatial excitation) and/or to make magnetization at different spatial locations at different frequencies, which enables k-space encoding of spatial information. The gradient coils allow the different parts of the body to be scanned.
- **Radio Frequency (RF) coil:** it is used to transmit a radio signal into the body part being imaged. This radio signal is applied after aligning the hydrogen nuclei with the high magnetic field.
- **Receiver Coil:** it is used to detect the returning radio signals due to the nuclei relaxation.
- **Computer System:** it is used to reconstruct the radio signals into the final image.

MRI parameters and image contrast

The contrast in an MR image can be manipulated by changing the pulse sequence parameters. A pulse sequence sets the specific number, strength, and timing of the RF and gradient pulses. Repetition time (TR) and echo time (TE) are the two key parameters that set the timing of the RF and gradient pulses, both measured in milliseconds:

- **TR:** is the time between the application of the RF excitation pulse and the start of the next RF pulse.

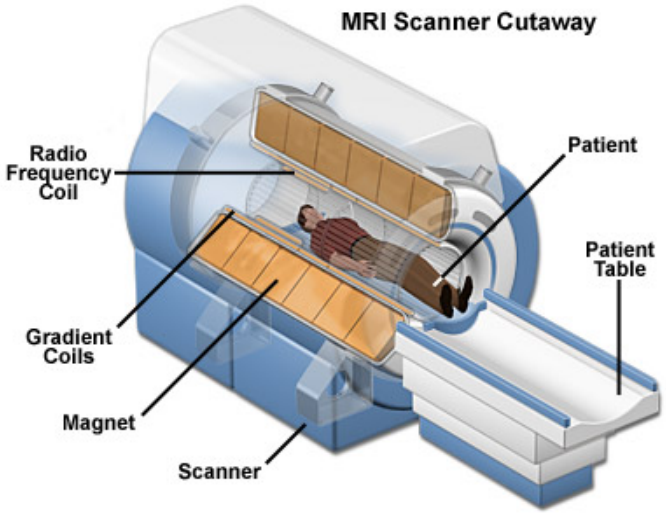


Figure 2.1: MRI scanner scheme. Image extracted from http://www.colinmcnulty.com/blog/wp-content/uploads/2011/08/mri-scanner-cutaway.jpg.

- **TE:** is the time between the application of the RF pulse and the peak of the echo detected.

For instance, the difference between fat and water can be detected at short TRs since the longitudinal magnetization (T1 recovery) recovers more quickly in fat than in water. On the other hand, differences in the T2 signal decay in fat and water can be detected at long TEs.

The spin echo (SE) and the gradient echo (GE) are two different MR pulse sequences that can be found in the daily practice of MRI. SEs are produced by pairs of RF pulses while GEs are generated by a single RF pulse in conjunction with a gradient reversal [41]. GE sequences can record the echo much more quickly, a fact that allows to reduce the TE. Moreover, when using low-flip-angle excitations (less than 90°) the TR can also be shorter. Hence, this kind of sequence is useful when fast scans are needed, although it does not correct for local magnetic field inhomogeneities. All other MRI sequences are variations of these two sequences obtained by different parameterization [14].

Conventional MRI

Conventional MRI (cMRI) sequences refers to techniques that are available and widely used in the diagnosis and treatment outcome measures in clinical trials [15]. The most common cMRI sequences are:

- **T1 weighted (T1-w):** related to TR (TR $<$ 1000ms, TE $<$ 30ms). Shorter TRs allow us to distinguish between fat and water.
- **T2 weighted (T2-w):** related to TE (TR $>$ 2000ms, TE $>$ 80ms). Longer TEs allow us to detect differences between fat and water.

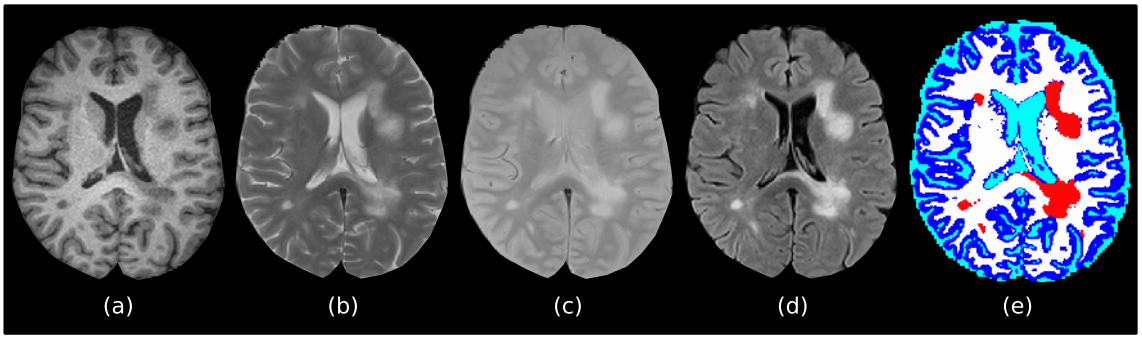


Figure 2.2: Different MR images of the brain: a) T1-w image, b) T2-w image, c) PD-w image and d) FLAIR image and their e) tissue segmentation: CSF appears cyan, GM appears dark blue, WM appears white and lesions appear red. Note that soft tissues are more distinguishable in the T1-w image, while lesions are usually better appreciated in the FLAIR one.

- **Proton Density-weighted (PD-w):** This is the result of a dual echo sequence on the T2w (TR $>$ 2000ms, TE $<$ 30ms).
- **Fluid Attenuated Inversion Recovery (FLAIR):** T2-w with the CSF signal suppressed, presenting a high contrast between tissue and lesions. An inversion recovery pulse is used to null the signal from the CSF.

In addition, other modalities such as Magnetization-Prepared Rapid Acquisition with Gradient Echo (MP-RAGE), Double Inversion Recovery (DIR), and Phase-Sensitive Inversion Recovery (PSIR) can also be helpful for disease diagnosis and follow-up.

2.1.2 What are MR images of MS patients like?

As shown at Figure 2.2, the high contrast between the main brain tissues (GM, formed by neuron nuclei, WM, formed by neuronal axons, and CSF which is the colorless bodily fluid that provides protection and cerebral autoregulation of cerebral blood flow) offered by cMRI modalities are clear. For instance, the CSF appears dark in both T1-w and FLAIR images, while its the brightest tissue in T2-w and has similar intensities to GM in PD-w images. On the other hand, WM is the brightest tissue in T1-w, has an intermediate gray level in FLAIR, similar to GM, and has the lowest signal in both PD-w and T2-w images. Finally, GM also appears with an intermediate gray level in T2-w and T1-w images in comparison with to the other two brain tissues. In MRI, MS plaques are well-delimited regions with hypointense signal intensity with respect to GM on T1-w, while hyperintense with respect to GM on T2-w, PD-w and FLAIR modalities.

Each sequence has its own advantages and drawbacks. For instance, while T1-w images depict the anatomy better, T2-w images provide better depiction of the disease due to fact that most tissues involved in a pathologic process have a higher

water content than normal tissues. On the other hand, PD-w sequences are capable of depicting both the anatomy and the disease entity [14]. Therefore, all sequences have some advantages and drawbacks in visualizing MS lesions in various parts of the brain.

At this point, we have seen why MRI has become a powerful technique in clinical practice for MS. Thanks to the presence of water molecules in the brain, and more precisely, hydrogen nuclei, MRI scanners can provide volumetric soft tissue information with a high contrast. Moreover, by tuning the MRI parameters, such as the pulse sequence or relaxation times, different volume sequences can be acquired. The most widely used images in MS trials are PD-w, T1-w, T2-w, and FLAIR. Focusing on MS lesions, they can be classified into 3 groups (T2-w lesions, T1-w lesions, and enhancing lesions) depending on their pathology and properties in other images (see Figure 2.3).

T2-w lesions

T2-w SE sequences consists of two sequences one with a short TE (PD-w) and one with a long TE (T2-w) images, and are called dual echo images [42]. In T2-w sequences, the characteristic appearance of MS is bright hyperintense lesions (HL), reflecting their increased water content (see Figure 2.3.a-c). The signal increase indicates edema, inflammation, demyelination, reactive gliosis and/or axonal loss in proportions that differ from lesion to lesion. They are typically discrete and focal in the early stages of the disease, but become confluent as the disease progresses.

These lesions are more frequent in periventricular areas and also typically seen in juxtacortical, infratentorial and temporal regions (see Figure 2.4). In PD-w images, the periventricular lesions are easily identified [43, 45] because of the better contrast between periventricular MS lesions and CSF when compared to T2-w images, but suffer more from flow artifacts, particularly in the posterior cranial fossa, which makes it difficult to identify infratentorial lesions. As FLAIR images produce heavily T2-w images by suppressing the signal from CSF, they can increase the noticeability of lesions, particularly those located in the periventricular area. However, they are less sensitive in the depiction of plaques involving the brainstem and cerebellum [43].

T1-w lesions

Unlike T2-w lesions, MS lesions in T1-w sequences can be both hyperintense and hypointense. Approximately 10% to 20% of T2-w HL are also visible as areas of low signal intensity compared with normal appearing white matter (NAWM) in T1-w images, so called black holes (BH) (see Figure 2.3.d) [17]. These T1-w BH have a different pathological substrate that depends, in part, on the lesion's age. Chronic black holes correlate pathologically with the most severe demyelination and axonal loss, indicating areas of irreversible tissue damage. T1-w images have a higher specificity than T2-w images for detecting lesions with irreversible tissue damage

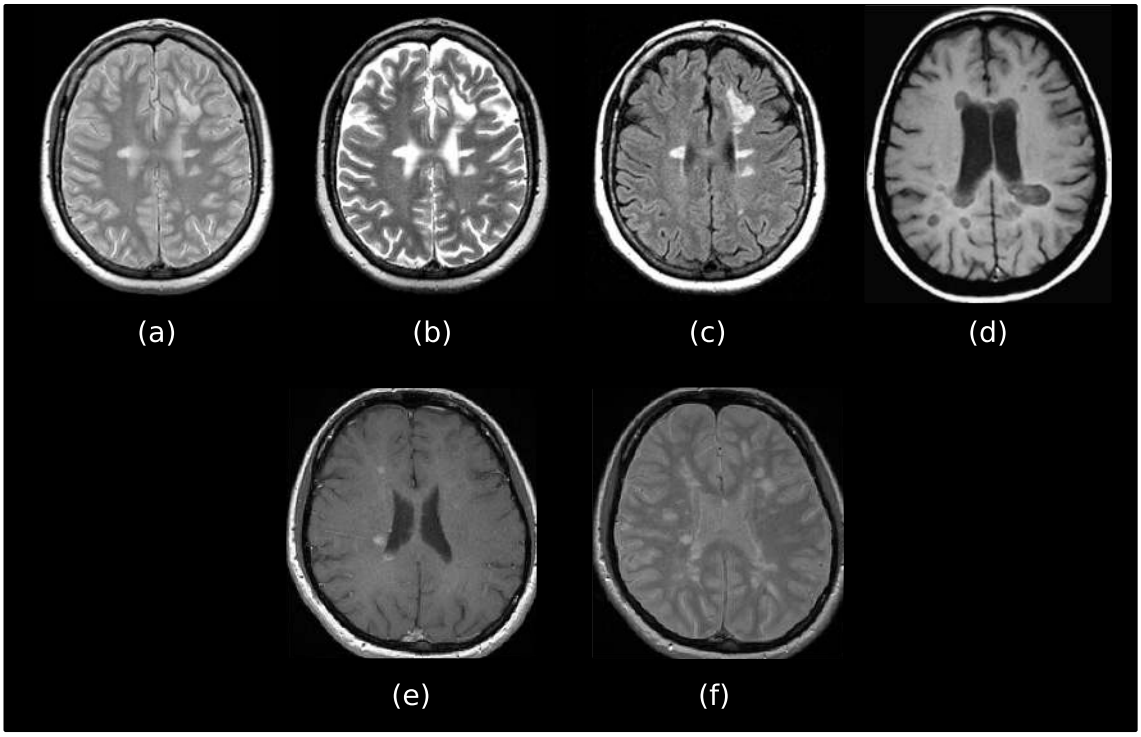


Figure 2.3: Multiple sclerosis lesion types. a), b) and c) are PD-w, T2-w and FLAIR images of a patient with hyperintense lesions, respectively. d) T1-w image of a patient with hypointense lesions (black hole). e) and f) are contrast-enhanced T1-w and PD-w images of a patient depicting enhanced MS lesions in T1-w and their corresponding hyperintense lesions in PD-w, respectively. Images from [43, 44].

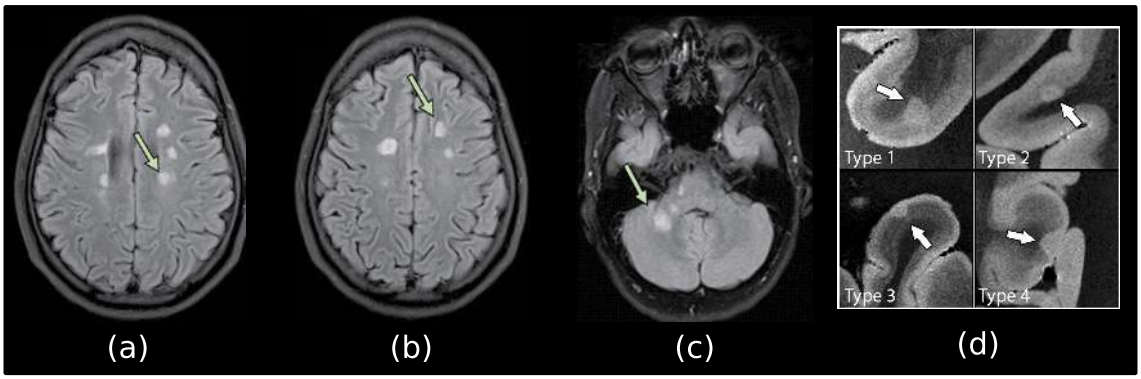


Figure 2.4: Multiple sclerosis lesion location within the brain: a) peri-ventricular, b) juxtacortical, c) infratentorial. (d) cortical lesions which can be classified into leuko-cortical (type 1), intracortical (type 2), subpial (type 3) and lesions that extend to the entire width of the cortex (type 4). Images from [8, 44].

and may serve as surrogate markers of the disability progression in clinical trials.

Enhancing lesions

These lesions appear as bright spots in T1-w images after applying a contrast agent, commonly Gadolinium. Gadolinium-enhanced T1-w imaging (consisting of applying a contrast agent before acquiring the image) is highly sensitive in detecting inflammatory activity. This technique detects disease activity 5 to 10 times more frequently than clinical evaluation of relapses, suggesting that most of these enhancing lesions (EL) are clinically silent. Individual and temporal MRI studies have shown that the formation of new MS plaques is often associated with contrast enhancement, mainly in the acute and relapsing stages of the disease. Approximately 65-80% of contrast enhancing lesions have a corresponding hypointensity in native T1-w images (see Figure 2.3.e-f) [43] and these acute hypointense lesions may become isointense or develop into BL lesions.

2.2 A review of brain MRI analysis in MS

Manual analysis of brain MR scans is, in practice, a highly time-consuming task. It is both challenging and time-consuming because of the large number of MRI slices that compose the three-dimensional information for each patient. Moreover, it is prone to intra-observer variability (the same study analyzed by the same neuroradiologist at different times) and inter-observer variability (the same study analyzed by different neuroradiologists). These conditions have led since the early nineties to the development of a wide number of methods for preprocessing and lesion and tissue segmentation, with the aim of reducing the time needed for manual interaction and the inherent variability of manual annotations [46, 47, 48].

2.2.1 Preprocessing of brain MR images

The automatic analysis of brain MR images is difficult because of variable imaging parameters, overlapping intensities, noise, partial volume, gradients, motion, echoes, blurred edges, normal anatomical variations and susceptibility artifacts [49]. Brain MR images, obtained directly from the scanner, contain the whole head and sometimes the neck. Images may suffer from intensity inhomogeneities or intensity non-uniformity. Also, different MR images may need to be aligned to a common space. So, preprocessing of brain MR images is a key step before starting any processing and analysis of the images by automatic approaches. The following subsection briefly describe the main preprocessing steps for brain MR images.

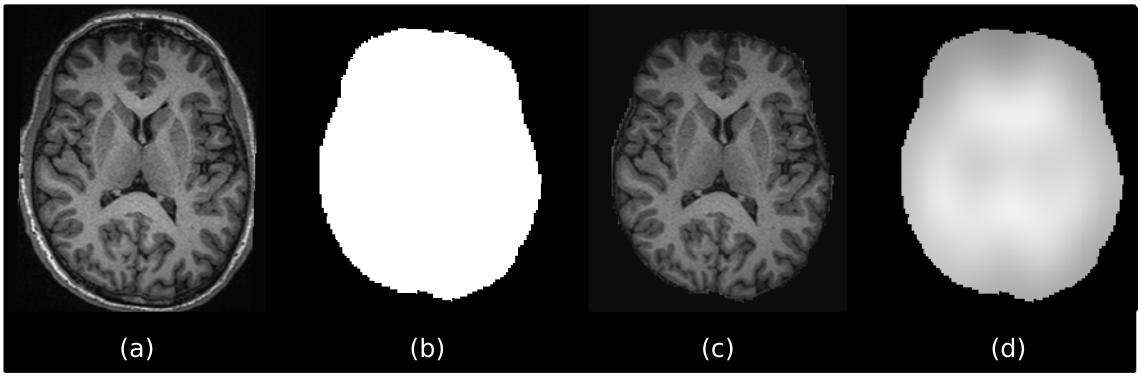


Figure 2.5: MRI preprocessing steps. a) T1-w image sequence. b) Computed brain mask using the ROBEX approach [50] and c) skull stripped T1-w sequence . d) Estimated T1-w bias-field using the N4 method proposed by [51].

Brain extraction

Acquired brain MR volumes incorporate non-brain tissue parts of the head such as eyes, fat, spinal cord or the skull. Skull stripping, also know as whole brain segmentation, is the process of extracting the brain tissue from nonbrain (see Figure 2.5b). This process of removing nonbrain tissue is the first module of most brain MRI studies. Many applications such as brain morphometry, brain volumetry, and cortical surface reconstructions require stripped MR scans [50]. The presence of non-brain regions affects the image histogram distribution and alters the segmentation performance of both tissues and lesions. Among the different methods proposed for skull-stripping [22, 52, 53], methods such as Brain Extraction Tool (BET) [54] and Brain Surface Extractor (BSE) [55] are being replaced by more modern methods such as ROBEX [50] and BEaST [56].

Bias field correction

Intensity inhomogeneity, also know as intensity nonuniformity or bias field, is an adverse phenomenon that appears in images obtained by different imaging modalities, not only in MRI but in microscopy, computer tomography, and ultrasound as well. Intensity inhomogeneity in MRI arises from the imperfections of the image acquisition process. It defines itself as a smooth intensity variation across the image (see Figure 2.5-d). Because of this phenomenon, the intensity of the same tissue varies with the location of the tissue within the image. Although intensity inhomogeneity is usually hardly noticeable to a human observer, many medical image analysis methods, such as segmentation and registration, are highly sensitive to the spurious variations of image intensities [57]. Note that image intensity correction process will be performed over the brain mask obtained in the skull stripping process. Among the available strategies [58, 59], the N3 [60] and N4 [51] methods are currently the most widely used tools used for bias field correction.

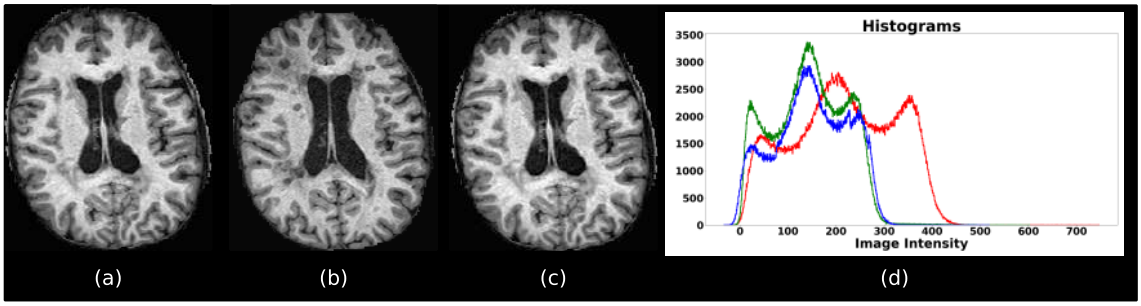


Figure 2.6: Histogram matching example. a) N4 normalized baseline T1-w image (target image), b) N4 normalized follow-up T1-w image (source image), c) histogram matched target image onto the source image, and d) histogram of (a) in red, (b) in green, and (c) in blue.

Histogram matching

In MRI, there is no a standard and quantifiable interpretation of image intensities. MR images taken for the same patient on the same scanner at different times may appear different from each other due to a variety of scanner-dependent variations and, therefore, the absolute intensity values do not have a fixed meaning [61]. Histogram matching aims at bringing together the intensity distribution of two images at a specified number of sample values. The work of Nyúl et al. [61] is currently the most widely used tools for intensity normalization. The main idea of the method is to deform the image histograms so that they match a mean histogram determined through training. The actual matching is based on certain landmarks identified on the histograms. In case of matching two images like matching the baseline image to the follow-up image in case of longitudinal study, training is not needed. The landmarks of the target image are matched to the corresponding landmarks of the source image (see Figure 2.6).

Registration

Once the brain has been extracted in both volumes, the bias field has been corrected, and the images have been normalized by histogram matching, they are ready for the registration process. The registration process consists of aligning two objects that are in different spaces. Registration is a key step in many automatic brain MRI applications. It is a fundamental step for both intra- and inter- subject analysis. Intra-subject registration is used to align different sequences from the same subject (also known as co-registration process) [22, 62], while the inter-subject registration is used when the source and target images usually belong to the same sequence from different subjects or when registering a subject image to a template (atlas registration). The registration methods follow the same strategy. Basically, they deform a source image to match a target image as much as possible by an optimization process of some energy function and a transformation model.

Registration is an important step in both cross-sectional and longitudinal studies. For instance, in cross-sectional studies, when the brain of a new patient with some unknown symptoms is compared with a healthy subject, i.e. a healthy brain without anatomical malformations, they also have to be aligned, while in longitudinal studies, to quantify the evolution of a disease, the patient must undergo follow-up scans at regular intervals, and the position of the head inside the scanner can be different every time so the different scans must be aligned in order to be comparable. Two brains are perfectly aligned when the corresponding voxel in both scans have the same physical spatial localization.

Several surveys and reviews [63, 64, 65] have compared the different registration techniques, but they are mainly based on two steps:

- **Rigid and affine registration:** is well-suited when there is no big difference between the two images. In rigid registration, the matching of two images is performed by finding the rotation and translation (6 degree-of-freedom (DOF)), while in affine registration, a 12 DOF includes shape recovering (scaling and shearing), that optimize some mutual function of the images.
- **Nonrigid registration:** allows the deformation of each pixel locally depending on their local similarity and position. These algorithms may need a regularization term in order to control the deformations since they can adopt undesirable effects. Methods can be classified into classical optimization approaches or learning-based approaches. Deformation models differentiate between these methods, elastic or hyperelastic models [66], viscous fluid [67, 68] and Demons (optical flow) [69, 70], more suitable for large deformations, and free form deformation (B-splines) methods [71], a smooth and continuous deformation controlled by a mesh of control points.

Lesion filling

In MS, when hypointense WM lesions are not included in the segmentation model, they have to be preprocessed before tissue segmentation in order to reduce the effects of WM lesions on the segmentation. Historically, WM lesions have been masked out of the T1-w before segmentation, and their volume added to the WM afterwards [72]. Although this method effectively reduces the error in tissue volume, it has been shown in several studies that this approach is not optimal [73, 74], because on images with high lesion load, the lack of lesion voxels may be modifying the original WM tissue distribution of the image, introducing significant differences in tissue segmentation.

In this respect, several strategies have proposed inpainting lesions on the T1-w with signal intensities of the normal appearing WM before tissue segmentation [73, 74, 75, 76], a process known in the literature as lesion filling (see Figure 2.7 for an example). However, most of the available lesion filling methods require manual delineations of lesions, which may be a tedious, challenging and time-consuming task

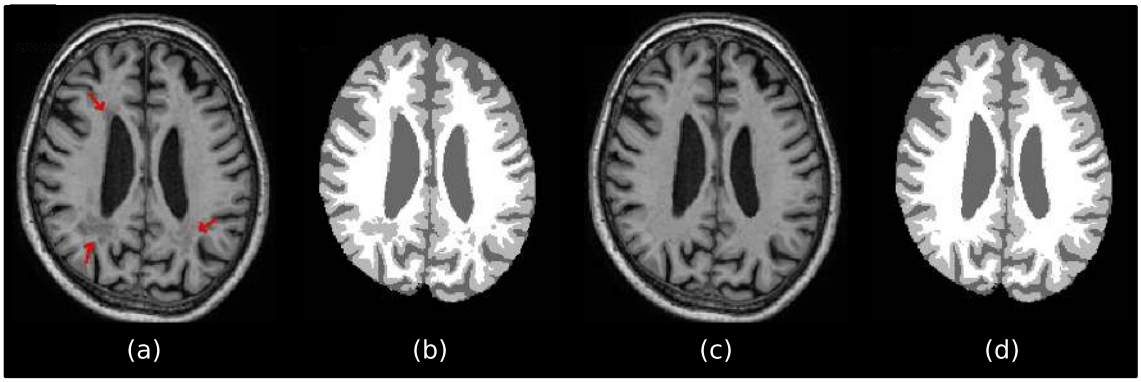


Figure 2.7: Lesion filling example on a slice of a T1-w scan. a) T1-w image sequence containing WM lesions (depicted by red arrows). b) Segmented T1-w sequence containing lesions. GM is depicted in light gray color, WM in white color and CSF in dark gray color. c) T1-w sequence after lesion filling. d) Segmented lesion filled T1-w sequence.

depending on the characteristics of the image [31]. When available, lesion filling has demonstrated a significant reduction not only in the associated errors of WM lesions in tissue volume measurements [77], but also in image registration [35, 76, 78] and cortical thickness measurements [75].

2.2.2 Brain tissue segmentation in MS

Brain tissue segmentation is the process of partitioning the brain into its three main tissues WM, GM, and CSF (see Figure 2.8). It is considered as an active research topic in medical image analysis as it provides doctors with meaningful quantitative information, such as tissue volume and shape measurements [40]. This information is widely used to diagnose brain pathologies and evaluate progression through regular MRI analysis over time [11, 79]. It is also important for neuroscientific studies, such as cortical surface extraction [80, 81], atrophy and volume measurements [23, 82], brain extraction [22, 83, 84], MS lesion segmentation [24, 27, 85], etc. It has been proved that there is a correlation between brain tissue atrophy measurements and MS disability status [86, 87].

A wide number of brain tissue segmentation methods have been proposed so far, usually on T1-w sequences, as this modality has clear difference in the intensity distributions of these three tissues. The well-known Markov random field is the basis of the FAST [88], which is part of the FMRI Software Library. SPM5/8/12 are three of the available versions of the SPM toolbox. This toolbox includes several image processing methods, one of which is the tissue segmentation based on a Gaussian Mixture Model, atlas registration and a bias field correction performed iteratively [89]. Most of the unsupervised automatic tissue segmentation methods in the current state of the art rely only on the signal intensity in T1-w sequences [90, 91, 92]. In

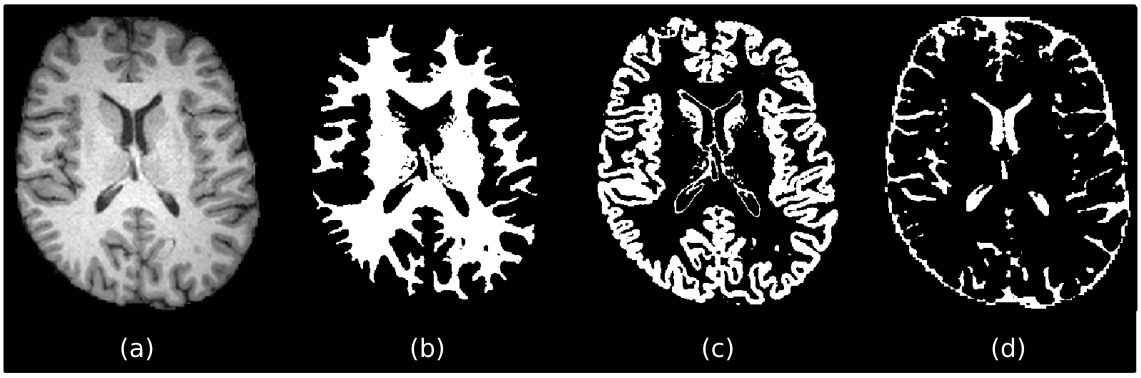


Figure 2.8: MRI brain tissue segmentation example using FAST [88]. a) T1-w image. b), c) and d) are WM, GM and CSF tissue segmentations, respectively.

contrast, supervised learning approaches also combine T1-w sequences with other modalities such as T2-w and PD-w [93, 94, 95]. Many supervised machine learning methods have been proposed after the CNN era [96, 97, 98].

The brain tissue segmentation is also an important step when detecting neurological lesions, usually present in WM but also seen in GM. Therefore, the tissue information is commonly used for lesion detection, although, at the same time, the lesions may affect the tissue segmentation accuracy. Most of these brain tissue segmentation methods are not designed to deal explicitly with MS lesions, which can reduce their accuracy when applied to MS patient images [38, 73, 74, 99]. A commonly used technique to overcome this issue consists of inpainting the lesions (see section 2.2.1) on the T1-w sequence with signal intensities of the NAWM before segmentation, achieving a significant reduction in the associated errors of WM lesions in tissue volume measurements [100].

2.2.3 MS lesion segmentation

Automatic segmentation of MS lesions in brain MRI has been widely investigated in recent years with the goal of helping MS diagnosis and patient follow-up. These plaques of demyelination are typically observed in MRI with different contrasts depending on the image sequence. cMRI described in 2.1.1 are highly sensitive in detecting MS plaques and can provide quantitative assessment of inflammatory activity and lesion load. They are commonly seen as hyperintense lesions in T2-w, PD-w and FLAIR and usually appear as dark areas in T1-w images (see Figure 2.2). Both acute and chronic MS plaques appear as focal high-signal intensity areas on T2-w sequences, reflecting their increased tissue water content. The increase in the signal indicates edema, inflammation, demyelination, reactive gliosis and/or axonal loss in proportions that differ from lesion to lesion. They are typically discrete and focal at the early stages of the disease, but become confluent as the disease progresses [101].

MS lesions are located in characteristic regions of the brain (see Figure 2.4), that include the periventricular, cortical or juxtacortical, and infratentorial regions. Cortical lesions, at the same time, can be classified according to their location within the GM as leuko-cortical (involving the deeper layers of the gray matter as well as the adjacent white matter at the gray/white matter junction), intra-cortical (small demyelinated lesions often centered around blood vessels and confined within the cortex), subpial (extending from the pial surface into the cortex) and lesions that extend to the entire width of the cortex.

As seen in section 1.3, McDonald criteria aims to use MR images in order to provide evidence of lesion dissemination in space and time, conditions that have to be fulfilled for a definite MS diagnosis. So, an automatic system to detect and segment MS lesions would help in the clinical practice to diagnose, as well as to evaluate a patient's follow-up and the effect of drug therapy. In what follows, a review of cross-sectional MS lesion segmentation (MS lesion segmentation in a single time point) and longitudinal MS lesion segmentation (MS lesion segmentation between successive time points) is presented.

Cross-sectional MS lesion segmentation

A wide number of automated WM lesion segmentation techniques have been proposed over the last few years. The voxel intensity is the most common feature used for lesion segmentation [102]. Analyzing the literature, one may distinguish between single-channel or multi-channel approaches, i.e. approaches that use only one MR image or those that combine several images, the later being the most widely used in the literature [101]. Single-channel approaches are mainly used to segment the brain tissues. For instance, T1-w images are widely used for this purpose, since they show the best contrast between the three main brain tissues: WM, GM and CSF. This initial tissue segmentation may then be used to help obtain the final lesion segmentation, and T2-w, PD-w, and FLAIR are the classical images for detecting MS lesions. The work of Khayati et al. [103] is an example of the single-channel approach in which the MS lesions are segmented using just the FLAIR sequence. Spatial information features are also used in some approaches. These features can be included using Markov Random Fields [104, 105], Fuzzy Connectedness [106] or probabilistic atlas [107].

Based on a review proposed by Lladó et al. [101], methods can be classified into either supervised and unsupervised segmentation strategies (see Figure 2.9). In supervised approaches, MS lesion segmentation is based on using some kind of a priori information or knowledge. These methods can be subclassified into either atlas-based methods or manual segmentation-based methods. In atlas-based methods [108, 109, 110, 111, 112, 113], the priori information may come from both statistical and topological atlases. A statistical atlas provides the prior probability of each voxel to belong to a particular tissue class while a topological atlas is usually used to preserve topology and to lower the influence of competing intensity clusters in regions that are spatially disconnected. As a drawback, these approaches rely on building

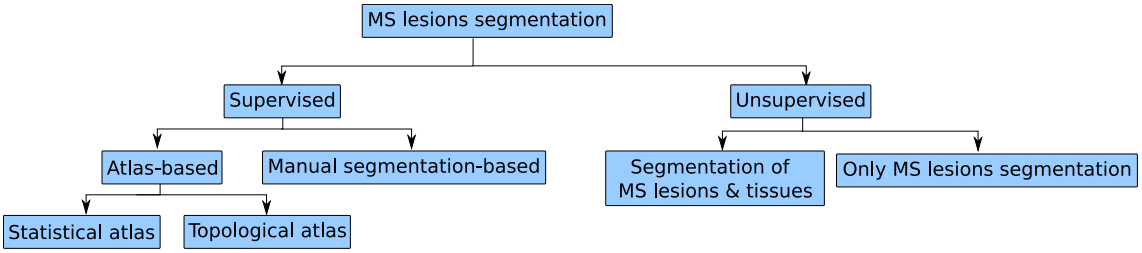


Figure 2.9: Classification of cross-sectional MS lesion analysis based on a review proposed by Lladó et al. [101]

an atlas, which is not a simple task. In addition, they also introduce the registration problem into the MS lesion segmentation. Note that this registration step is even more difficult when dealing with cases with severe atrophy, large numbers of lesions, etc. In manual segmentation-based methods[114, 115, 116, 117, 118, 119, 120, 121], manually-segmented images annotated by neuroradiologists are used to segment the MS lesions. These methods use mainly the image intensities of different MR images to train a classifier for the segmentation purpose. Note that unlike atlas-based approaches these approaches do not need any registration process between the analyzed images and the atlas. However, some of these methods include the use of registration algorithms that focus on the intra-sequence and inter-sequence preprocessing registration steps.

In the unsupervised strategies, where no prior knowledge is used, methods can be subclassified into either methods that segment brain tissue to help lesion segmentation or methods that use only the lesion properties for segmentation. In former approaches [122, 123, 124, 125], there are methods that either segment the tissue first and then the MS lesions, or segment the tissue and the lesions at the same time. In the later approaches [27, 126, 127, 128, 129, 130, 131], the methods directly segment the lesions according to their properties, without providing tissue segmentation. Segmenting the tissue help neuroradiologists to evaluate the GM tissue volumetry and monitor the progression of cerebral atrophy.

Recently, the literature offers several methods based on CNNs for MS lesion segmentation in cross-sectional images [132, 133, 134, 135, 136, 137]. Interestingly, excellent results have been reported in the last few years within this topic, with methods achieving segmentations that are close to human expert inter-rater variability. For instance, Valverde et al. [28] proposed an automated method for WM lesion segmentation of MS patient images. Within this MS lesion segmentation framework, a cascade of two identical CNNs is optimized, where the first network is trained to be more sensitive to revealing possible candidate lesion voxels, while the second network is trained to reduce the number of false positive outcomes. Their method was the best ranked approach on the MICCAI2008 challenge, outperforming the rest of 60 participant methods when using all the available input modalities (T1-w, T2-w and FLAIR), while still in the top-rank (3rd position) when using only T1-w and FLAIR modalities. For a complete description of the details and motivations for the

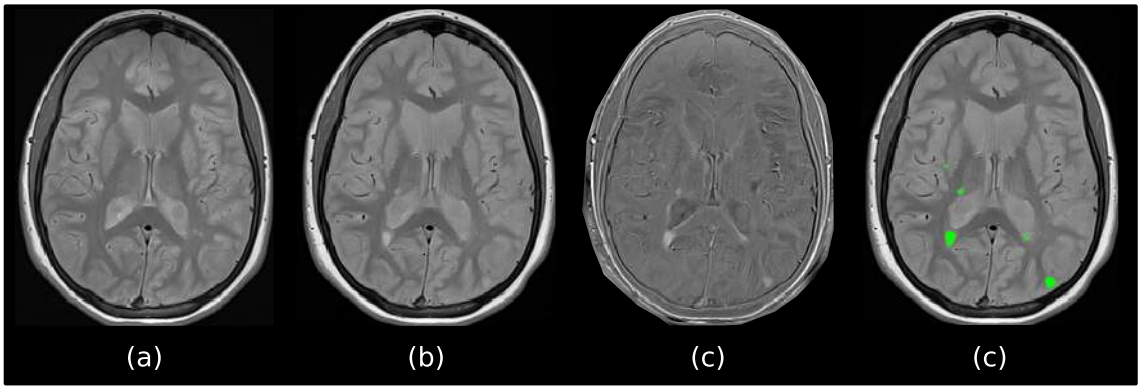


Figure 2.10: An example of MS lesion longitudinal analysis. a) baseline PD-w, b) follow-up PD-w, c) the subtraction and d) the manual lesion annotations of the slices performed by an expert radiologist overlaid in green on the follow-up image.

proposed architecture, please refer to the original publication.

Longitudinal MS lesion segmentation

Follow-up brain MRI is required in patients who have not been diagnosed yet as MS patients but they show clinical and radiological findings suggestive to MS [138]. 3-6 months were suggested to be the optimal interval between the baseline and the follow-up scan. A third scan can be acquired 6-12 months later if no new lesions are seen the first follow-up scan [138, 139]. Different methodologies and approaches have been proposed for getting MS biomarkers from individual patients by combining clinical and MRI criteria evaluated after 6 or 12 months from therapy start [140, 141, 142, 143, 144, 145]. However, the detection of this disease activity is performed visually by comparing the follow-up and baseline scans. Due to the presence of small lesions, misregistration, and high inter-/intra-observer variability, it is difficult to visually detect active T2-w lesions in patients with MS [146]. Automatic methods can overcome these issues by eliminating stable lesions and also highlighting evolving T2-w lesions [147, 148]. Figure 2.10 shows an example of MS lesion longitudinal analysis of a patient's brain taken with a year's difference, together with the manual annotations made by an expert.

Based on an another review proposed by Lladó et al. [31], methods can be classified into either lesion detection approaches or change detection approaches (see Figure 2.11). In the lesion detection approaches, both static and dynamic MS lesions on a single-time MR volume of a patient are detected. These segmentation-based methods, which can be supervised or unsupervised, rely on the intensity homogeneities of the tissues and typically apply data mining techniques (clustering, classification) to distinguish lesions from normal tissues. In longitudinal analysis, lesion quantification approach is subsequently needed to compute the volumetric changes of each segmented lesion between two time points for the MS lesion evolution. Recently, Schmidt et al. [149] proposed an automated algorithm for segmentation of WM

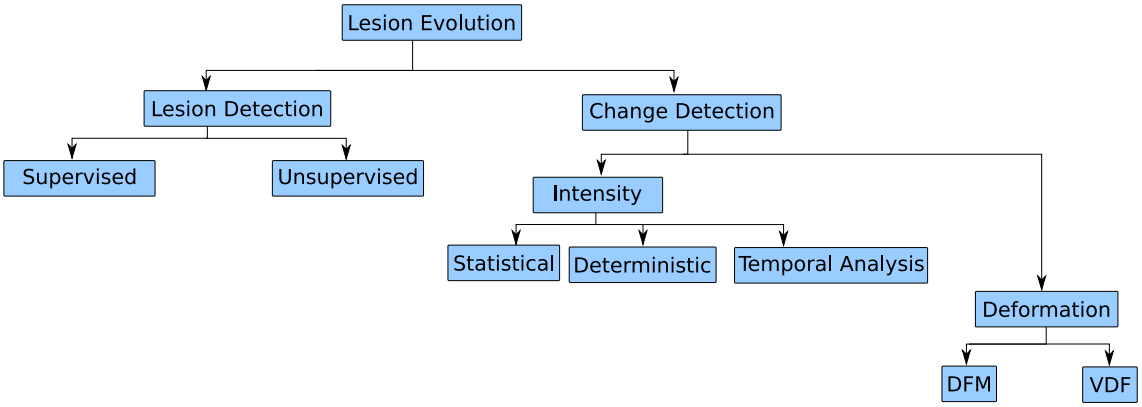


Figure 2.11: Classification of longitudinal MS lesion analysis methods based on a review proposed by Lladó et al. [31].

lesion (WML) changes by extending their earlier work on cross-sectional WML segmentation [85]. Their algorithm requires three-dimensional gradient echo T1-w and FLAIR images at 3T as well as available cross-sectional lesion segmentations of both time points. Preprocessing steps include lesion filling and intra-subject registration. For segmentation of lesion changes, initial lesion maps of different time points are fused; herein changes in intensity are analyzed at the voxel level. Significance of lesion change is estimated by comparison with the difference distribution of FLAIR intensities within normal appearing white matter.

In the change detection approaches, the differences between successive MRI controls at both 2D and 3D image levels are analyzed not at a single time point. An MS lesion is generally seen as the combination of two different effects, tissue transformation and tissue deformation [150]. Tissue transformation refers to the intensity change in the tissue of the lesion, while tissue deformation refers to the modification of its surrounding tissue, due to lesion expansion or contraction. These methods can be subclassified into either intensity-based approaches or deformation-based approaches.

In the intensity-based approaches, voxel-wise comparisons are performed between successive scans. Moraal et al. [151] mentioned that subtraction imaging allowed the direct quantification of positive and negative disease activity. They also mentioned that 3D subtraction imaging increased the detection of active MS lesions in various parts of the brain compared with 2D subtraction imaging [147]. Elliott et al. [152] presented a framework for automated detection of new MS lesions using a two-stage classifier that first performed a joint Bayesian classification of tissue classes at each voxel of the baseline and follow-up images using intensities and subtraction values, and then a lesion-level classification was performed using a random forest classifier. Ganiler et al. [32] extended the pipelines of Moraal et al. [147] and Elliott et al. [152] by adding multi-channel information and several additional steps such as constraining the region of interest to the WM and using simple postprocessing steps based on the baseline and follow-up image intensities. Supervised learning is a machine learning task which consists in predicting a function from labeled training

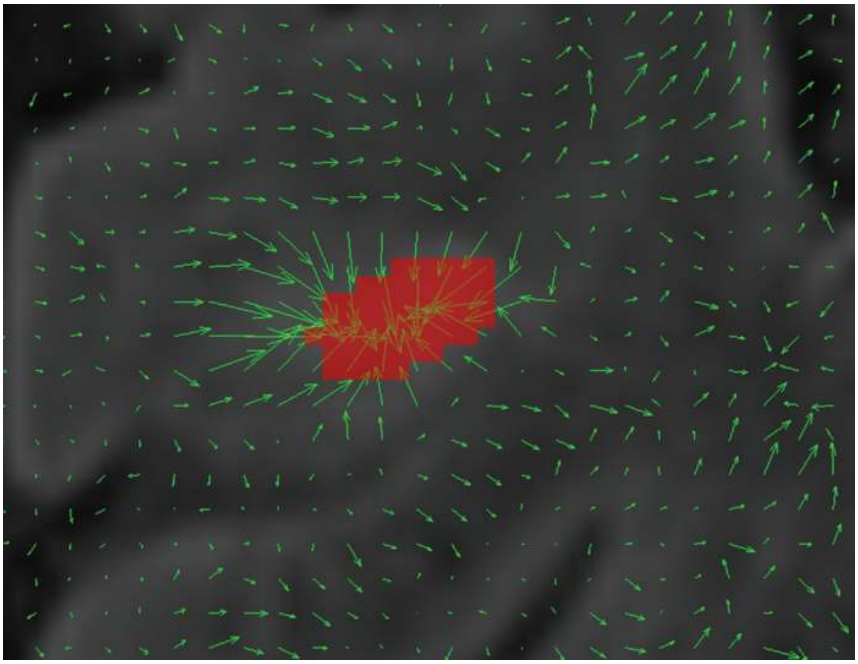


Figure 2.12: An example of the DF inside a new lesion. All arrows point to the lesion center [33].

data [153]. Different algorithms can be used to learn a mapping function from input feature vectors to the desired output values [154, 155]. Sweeney et al. [156] proposed the automated SuBLIME method for segmenting lesion incidence between two MRI studies based on a supervised LR model trained only using features from the follow-up study and the subtraction between timepoints.

In the deformation-based approaches, the new T2-w lesions detection is performed by analyzing the DFs obtained by nonrigid registration between successive scans. Nonrigid registration provides a discrete local displacement field that defines the deformation occurring between two images (see Figure 2.12). Thirion and Calmon [150] used the DF to detect evolving lesions in longitudinal MRI. They defined several DF operators to automatically detect regions that present changes. They proposed to use both the divergence and the norm of the displacement vector fields in order to be sensitive to deformation and intensity change. Therefore, high values of the norm indicated large deformation areas, while high divergence indicated evolving lesions, where the sign of the divergence operator showed whether the lesion was growing or shrinking. Rey et al. [157] improved the approach of Thirion and Calmon [150] by using the Jacobian operator to determine local volume changes instead of using the divergence and norm of the vector fields. Furthermore, they used multiresolution levels to avoid the influence of the motion in the center of a lesion by the vectors in the boundary. Using the Jacobian operator, it is possible to distinguish the lesion's evolution. As it is commonly accepted, the authors stated that a Jacobian operator larger than 1 indicates a local expansion, while smaller values indicate local shrinking.

New lesion detection approaches have been also proposed combining information from different sources. For instance, Fartaria et al. [158] proposed a strategy for longitudinal analysis of MS lesions based on a combination of segmentation-based and intensity-based approaches to assess the performance of the partial-volume aware lesion segmentation tool and to propose a method for the generation of a lesion progression map between two time points. Moreover, several methods have been proposed as a combinations of intensity-based and deformation-based approaches. Cabezas et al. [33] improved the subtraction pipeline proposed by Ganiler et al. [32] by combining subtraction and DF operators to decrease the number of false positive lesions detected by the subtraction pipeline. In their work, an automated threshold was computed for each subtraction image (PD-w, T2-w, and FLAIR) and applied separately to obtain 3 initial lesion masks. The thresholds were computed as the mean of the subtraction image within the WM plus 5 standard deviations to guarantee that only hyperintense regions were detected and to maintain a large number of true-positives (TPs). Lesions smaller than 3 voxels were excluded to reduce the effects of noise. The intersection of the 3 masks (PD-w, T2-w, and FLAIR) was used to differentiate between errors and true lesions in each mask. Finally, two different postprocessing approaches were used independently to refine the initial generated lesion mask. The first one was based on intensity by applying different intensity-based rules to the baseline and follow-up images while the second was based on DFs in which Divergence, Jacobian, and Concentricity were used to accept or reject the candidate lesions [33].

2.3 Machine learning for medical image analysis

2.3.1 What is machine learning?

Machine learning (ML) is an exciting field of research in computer science and engineering. It is considered a branch of artificial intelligence (AI) [159]. ML contains a set of methods, which enable a machine to learn meaningful patterns from data directly with minimal human interaction. More recently, machines have shown that they are capable of learning and even mastering tasks that were thought to be too complex for machines, showing that machine learning algorithms are potentially useful components of computer-aided diagnosis and decision support systems. Computers seem to be able to recognize patterns that are beyond human perception which has led to increased interest in the field of ML, and specifically, how it might be applied to medical images [159].

As shown in Figure 2.13, the most common machine learning techniques are as follows:

- **Supervised learning:** a supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. This is the

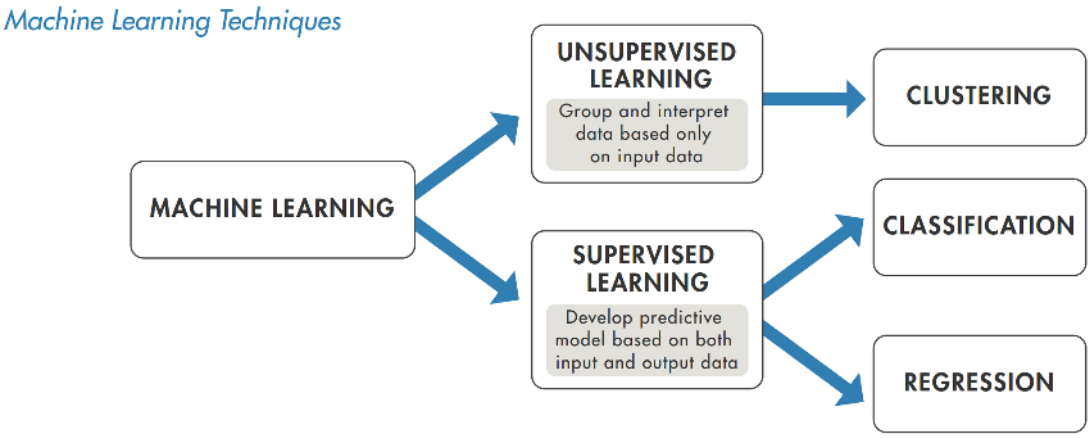


Figure 2.13: ML Techniques. Image extracted from https://www.mathworks.com/help/stats/machine-learning-in-matlab.html.

most common scenario associated with classification, regression, and ranking problems. Supervised learning uses classification and regression techniques to develop predictive models. The classification techniques predict discrete responses while the regression techniques predict continuous responses. Examples of supervised learning algorithms include support vector machine [160], decision tree [161], linear regression [162], logistic regression [163, 164], naive Bayes [163, 165], k-nearest neighbor [166], random forest [167], AdaBoost, and neural network methods [168].

- **Unsupervised learning:** an unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning technique is clustering which analyzes data to find hidden patterns or groupings. The algorithm system determines how many groups there are and how to separate them. Examples of unsupervised learning algorithm systems include K-means [169], mean shift [169, 170], affinity propagation [171], hierarchical clustering [171, 172], DBSCAN (density-based spatial clustering of applications with noise) [173], Gaussian mixture modeling [173, 174], Markov random fields [175], and fuzzy C-means systems [176].

2.3.2 Neural networks

The neural network (or artificial neural network or ANN) is derived from the biological concept of neurons which are the basic computational unit of the brain. Figure 2.14 shows a cartoon drawing of a biological neuron and a common mathematical model. Each neuron receives input signals from its dendrites and produces output signals along its (single) axon. The axon eventually branches out and connects via synapses to dendrites of other neurons [177]. In the computational model of a

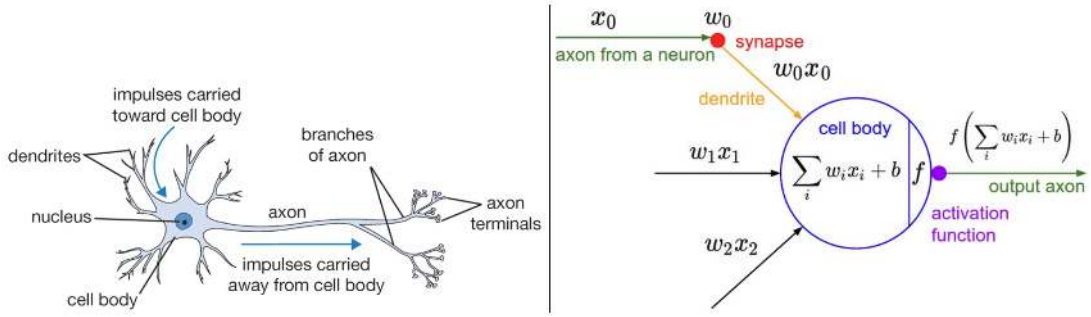


Figure 2.14: Biological neuron (left) and its mathematical model (right). A real brain neuron receive signals from other neuron through dendrites. The neuron sends signals at spikes of electrical activity through a long thin stand known as an axon and an axon splits this signals through synapse and send it to the other neurons. The mathematical model of a neuron is nothing more than a set of inputs, a set of weights, and an activation function. The neuron translates these inputs into a single output, which can then be picked up as input for another layer of neurons later on. Image extracted from http://cs231n.github.io/neural-networks-1/.

neuron (the artificial neuron), the signals that travel along the axons (e.g. x_0) interact multiplicatively (e.g. w_0x_0) with the dendrites of the other neuron based on the synaptic strength at that synapse (e.g. w_0). The idea is that the synaptic strengths (the weights w) are learnable and control the strength of influence of one neuron on another. In the basic model, the dendrites carry the signal to the cell body where they all get summed. If the final sum is above a certain threshold, the neuron can fire, sending a spike along its axon. In the computational model, the firing rate of the neuron is modeled with an activation function f. The activation function represents a linear combination of the input x to the neuron and the parameters w, followed by an element-wise nonlinearity $\sigma()$, referred to as a transfer function:

$$a = \sigma(w^T * x + b).$$

Typical transfer functions for traditional neural networks are the sigmoid and hyperbolic tangent function.

ANN consists of large number of simple processing elements that are interconnected in an acyclic graph with each [178, 179]. ANN models are often organized into distinct layers of neurons. For regular neural networks, the most common layer type is the fully-connected layer in which neurons between two adjacent layers are fully pairwise connected, but neurons within a single layer share no connections. Figure 2.15 shows two examples of ANN topologies that use a stack of fully-connected layers. The multi-layered perceptrons (MLP), the most well-known of the traditional neural networks, have several layers of these transformations. In MLP, there is an input layer, output layer, and some layers in between. Layers in between the input and output are often referred to as hidden layers. Moreover,

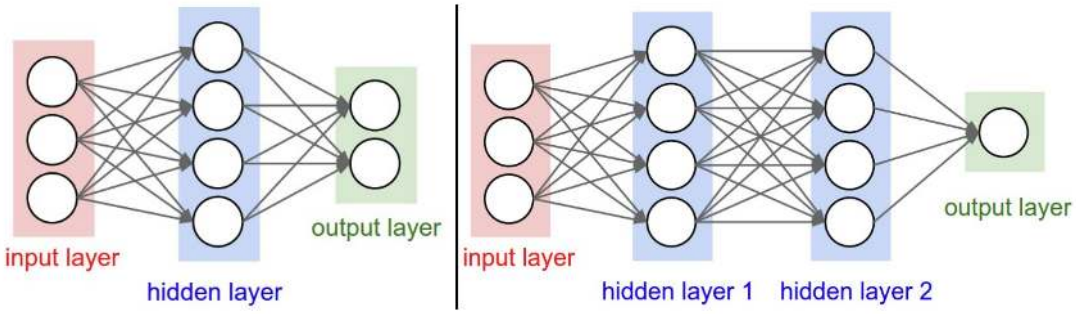


Figure 2.15: A 2-layer Neural Network (one hidden layer of 4 neurons (or units) and one output layer with 2 neurons), and three inputs (left). A 3-layer neural network with three inputs, two hidden layers of 4 neurons each and one output layer (right). Notice that in both cases there are connections between neurons across layers, but not within a layer. Image extracted from http://cs231n.github.io/neural-networks-1/.

the neurons of the final layer of the network do not have often activation function. This is because the last output layer is usually taken to represent the class scores (e.g. in classification). So, a distribution over classes is generated by feeding the activations in the final layer through a softmax function and the network is trained using maximum likelihood. When a neural network contains multiple hidden layers it is typically considered a deep neural network, hence the term deep learning.

2.3.3 Deep learning

As computational power is getting improved and enormous amounts of data is becoming available, deep learning, also known as deep neural network learning has become the default machine-learning technique because it can learn much more sophisticated patterns than conventional machine-learning techniques. Conventional machine-learning techniques were limited in their ability to process data in their raw form. To construct a ML system, a feature extractor should be designed and engineered carefully so that it can transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input [180].

Unlike conventional machine-learning techniques, deep learning methods simplify the feature extraction process and they could be applied to raw data directly (see Figure 2.16). This is very important for the field of medical image analysis since it allows more researchers to exploit new ideas easier and faster. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress. Deep learning has been shown to produce competitive results in medical applications such as cancer cell classification, lesion detection, organ segmentation or image enhancement [181, 182]. However, there are

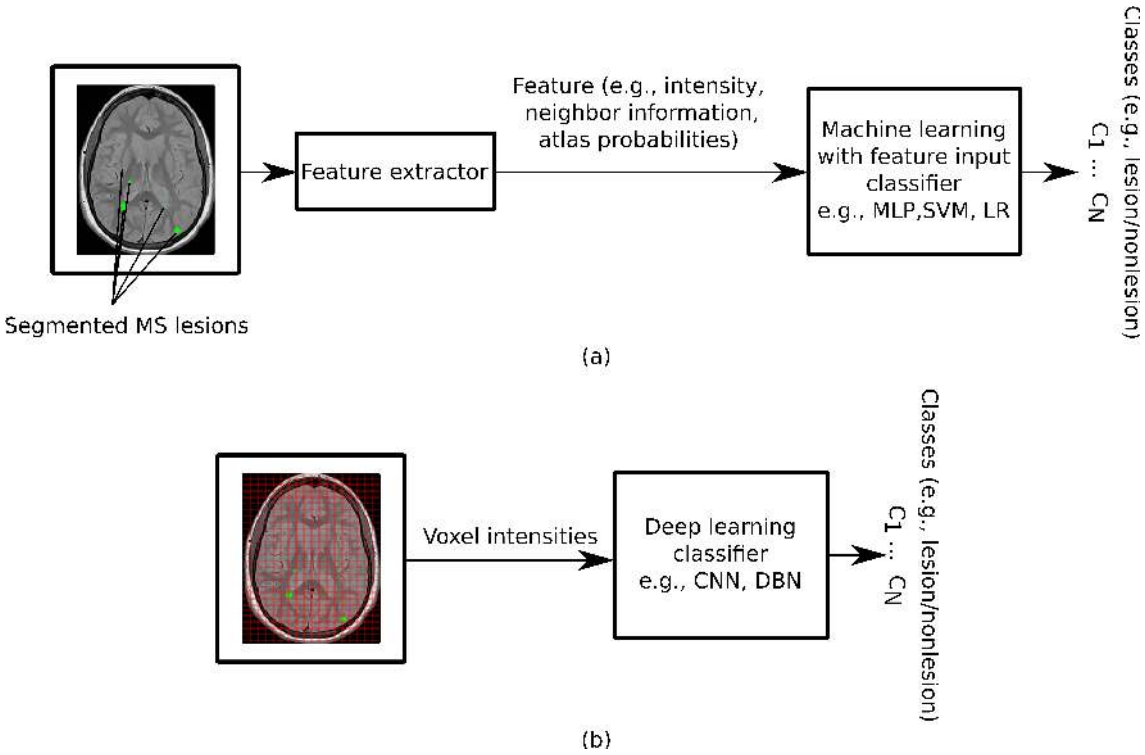


Figure 2.16: ML in medical image analysis before and after deep learning. (a) ML before deep learning: standard ML for classifying lesions. Features are extracted from a segmented lesion in an image. Those features are entered as input to an ML model with feature input (classifier) such as an MLP and a support vector machine (SVM). (b) ML after deep learning: voxel values from an image are directly entered as input to an ML model such as a CNN and a deep belief net (DBN).

some drawbacks of deep learning such as the difficulties to obtain enough data and the necessity to have data from several sites (i.e., from different scanners/protocols in case of medical imaging), otherwise the system will not be robust enough and highly dependent on the training data.

Early neural networks were typically only a few (< 5) layers deep, mainly because the computing power was not sufficient for more layers and owing to challenges in updating the weights properly. Deep learning refers to the use of neural networks with many layers. The parallel computing power of graphics processing units (GPU) such as those built by NVidia Corporation enabled this kind of deeper network to come to the world. Some deep learning algorithm tools are deep neural networks, stacked auto encoders, deep Boltzmann machines, and CNNs. We will focus on CNNs because these are most commonly applied to medical image analysis.

2.3.4 Convolutional neural networks (CNNs/ConvNets)

CNNs are designed to process data that come in the form of multiple arrays, for example a color image composed of three 2D arrays containing pixel intensities in

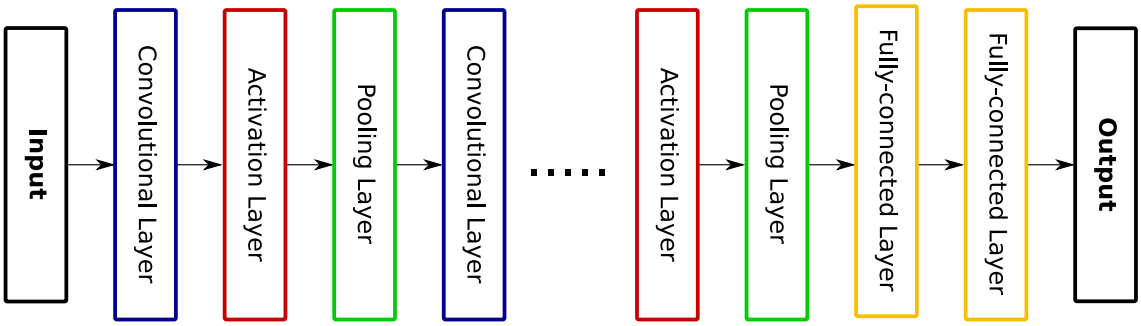


Figure 2.17: General CNN architecture.

the three color channels. There are many data modalities in the form of multiple arrays. For instance, 1D for signals and sequences, including language; 2D for images or audio spectrograms; and 3D medical images. There are four key ideas behind CNNs that take advantage of the properties of natural signals: local connections, shared weights, pooling and the use of many layers [180].

CNNs are similar to regular neural networks. The difference is that CNNs assume that the inputs have a geometric relationship like the rows and columns of images. The input layer of a CNN has neurons arranged to produce a convolution of a small image (i.e., kernel) with the image. This kernel is then moved across the image, and its output at each location as it moves across the input image creates an output value [159]. An important benefit of CNN deep learning algorithms, as compared with traditional machine learning methods, is that there is no need to compute features as a first step. The CNN effectively finds the important features as a part of its search process. As a result, the bias of testing only those features that a human believes to be important is eliminated. The task of computing many features and then selecting those that seem to be the most important also is eliminated [159, 183].

Architecture Overview

Figure 2.17 depicts the general architecture of CNNs. The three main types of layers to build CNNs architectures are as follows:

- **Convolutional Layer (CONV layer).** It is the core building block of a CNN. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field, but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product between the entries of the filter and the input, producing a 2-dimensional activation map of that filter. As a result, the network learns filters that activate when detect some specific type of feature at some spatial position in the input. For instance, a CONV layer with 12 filters will produce a separate 2-dimensional activation map for each filter. These activation maps will be stacked along the depth dimension and produce the output volume.

- **Activation Layer.** Every activation function (or nonlinearity) takes a single number and performs a certain fixed mathematical operation on it. There are several activation functions like Sigmoid, Tanh, (rectified linear unit) ReLU. For instance, ReLU effectively removes negative values from an activation map by setting them to zero [184]. It increases the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolution layer. ReLU is often preferred to other functions because it trains the neural network several times faster without a significant penalty to generalization accuracy [185].

- **Pooling Layer.** It is an important layer in CNNs. It is a non-trainable layer which takes groups of inputs and applies a simple function to each of these groups independently as a form of nonlinear down-sampling. The most common functions used are the max function and the mean function. When the max function is used, the pooling layer will take groups of inputs and output the maximum value for each one of these groups. In the same way, if the mean function is used, the pooling layer computes the mean (average) for each group of inputs. It has different usages like spatial size reduction or overfitting control.

- **Fully-connected Layer (FC layer).** Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. It can be computed with a matrix multiplication followed by a bias offset.

More changes can be done to this architecture like the number of convolutional layers and the size and number of kernels in each layer, stacking of different convolutional layers, different activation functions, and the number of the FC layers.

There are two key differences between MLPs and CNNs. First, in CNNs weights in the network are shared in such a way that the network performs convolution operations on images. This way, the model does not need to learn separate detectors for the same object occurring at different positions in an image, making the network equivariant with respect to translations of the input. It also drastically reduces the amount of parameters (i.e. the number of weights no longer depends on the size of the input image) that need to be learned. The second key difference between CNNs and MLPs, is the typical incorporation of pooling layers in CNNs, where pixel values of neighborhoods are aggregated using a permutation invariant function, typically the max or mean operation. This can induce a certain amount of translation invariance and increase the receptive field of subsequent convolutional layers. At the end of the convolutional stream of the network, FC layers (i.e. regular neural network layers) are usually added, where weights are no longer shared. Similar to MLPs, a distribution over classes is generated by feeding the activations in the final layer through a softmax function and the network is trained using maximum likelihood [181].

There are several CNNs architectures that are famous and common in the field of

computer vision. In 1998, LeCun et al. [186] developed LeNet which was relatively shallow, consisting of two CONV layers (see Figure 2.18.a). It was used to read zip codes, digits, etc. In 2012, Krizhevsky et al. [185] developed AlexNet which was the first work that popularized CNNs in computer vision. It consisted of five CONV layers (see Figure 2.18.b). AlexNet used rectified linear units instead of the hyperbolic tangent as activation function, which are now the most common choice in CNNs [181]. After 2012 many novel deeper architectures were developed. They were based on stacking smaller kernels, instead of using a single layer of kernels with a large receptive field, a similar function can be represented with less parameters. Examples of these architectures are VGG16, VGG19 [187], GoogLeNet [188], and ResNet [189].

Fortunately, the convolution and dot product are both linear operators and thus inner products can be written as convolutions and vice versa. The only difference between FC and CONV layers is that the neurons in the CONV layer are connected only to a local region in the input, and that many of the neurons in a CONV volume share parameters. By rewriting the fully connected layers as convolutions, the CNN can take input images larger than it was trained on and produce a likelihood map, rather than an output for a single pixel. The resulting network is called fully convolutional network (FCNN) [190]. FCNN can then be applied to an entire input image or volume in an efficient fashion (see Figure 2.18.f). Ronneberger et al. [191] proposed the U-net architecture, comprising a regular FCNN followed by an upsampling part where up-convolutions are used to increase the image size, coined contractive and expansive paths. The authors combined it with the so called skip-connections to directly connect opposing contracting and expanding convolutional layers (see Figure 2.18.g).

CNNs can simply be used to classify each pixel in the image individually, by presenting it with patches extracted around the particular pixel [192]. The training data in terms of patches is much larger than the number of training images. There are drawbacks of this approach. First, it is quite slow because the network must be run separately for each patch, and there is a lot of redundancy due to overlapping patches. Secondly, there is a trade-off between localization accuracy and the use of context. Larger patches require more max-pooling layers that reduce the localization accuracy, while small patches allow the network to see only little context [191].

2.3.5 Deep learning: hardware and software

As mentioned before, the widespread availability of GPU and GPU-computing libraries (CUDA, OpenCL) are behind the popularity of deep learning methods. GPUs are highly parallel computing engines, which have an order of magnitude more execution threads than central processing units (CPUs). With current hardware, deep learning on GPUs is typically 10–30 times faster than on CPUs [181].

Next to hardware, the other driving force behind the popularity of deep learning methods is the wide availability of open-source software packages. These libraries

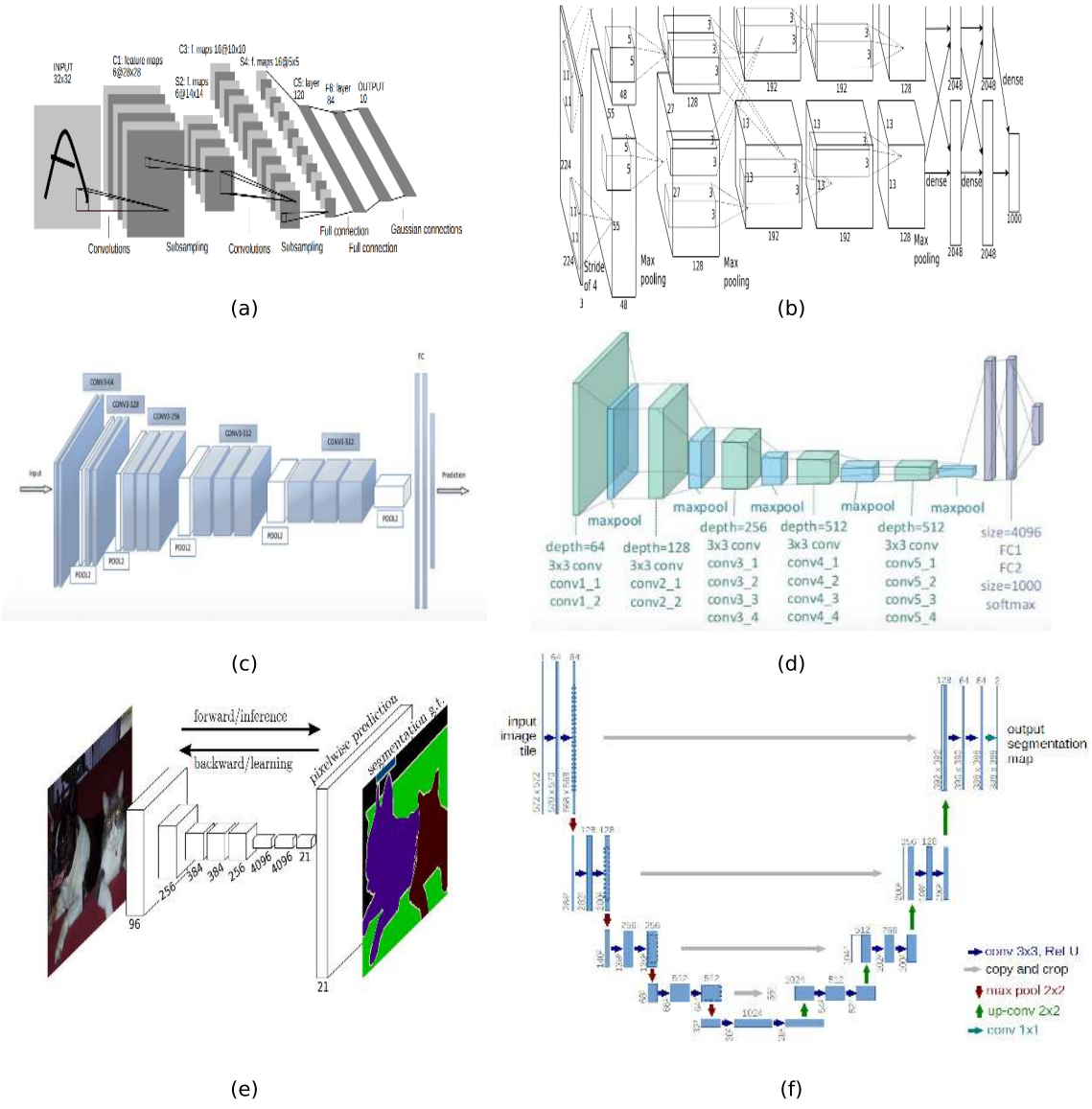


Figure 2.18: Examples of CNN architectures. a) LeNet [186]. b) AlexNet [185]. c) and d) VGG16 and VGG19, respectively [187]. e) FCNN [190]. f) U-net [191].

provide efficient GPU implementations of important operations in neural networks, such as convolutions; allowing the user to implement ideas at a high level rather than worrying about efficient implementations. At the time of writing, the most popular packages were:

- **Caffe [193].** Provides C++ and Python interfaces, developed by graduate students at UC Berkeley.
- **Torch [194].** Provides a Lua interface and is used by, among others, Facebook AI research.
- **PyTorch [195].** A machine learning library for Python, based on the Torch library. It is primarily developed by Facebook's artificial-intelligence research group.
- **Theano [196].** Provides a Python interface, developed by MILA lab in Montreal.
- **Tensorflow [197].** Provides C++ and Python and interfaces, developed by Google and is used by Google research [1].

There are third-party packages written on top of one or more of these frameworks, such as Lasagne [2] or Keras [3].

2.3.6 Deep learning in medical image analysis

Deep neural networks are now the state-of-the-art machine learning methods across a variety of areas, from image analysis to natural language processing, and widely deployed in academia and industry. These developments have a huge potential for medical imaging technology, medical data analysis, medical diagnostics and healthcare in general. Deep learning is providing exciting solutions for medical image analysis problems and is seen as a key technology for future applications.

Figure 2.19 depicts some medical imaging applications in which deep learning has achieved state-of-the-art results. From top-left to bottom-right: mammographic mass classification (Kooi et al. [198]), segmentation of lesions in the brain (top ranking in brain tumor segmentation challenges (BRATS), ischemic stroke lesion segmentation challenge (ISLES) and MR brain image segmentation challenge (MRBrains) challenges), image from Ghafoorian et al. [199], leak detection in airway tree segmentation (Charbonnier et al. [200]), diabetic retinopathy classification (Kaggle Diabetic Retinopathy challenge 2015, image from van Grinsven et al. [201]), prostate segmentation (top rank in PROMISE12 challenge), nodule classification (top ranking in LUNA16 challenge), breast cancer metastases detection in lymph nodes

[1] https://www.tensorflow.org/
[2] https://github.com/Lasagne/Lasagne
[3] https://keras.io/

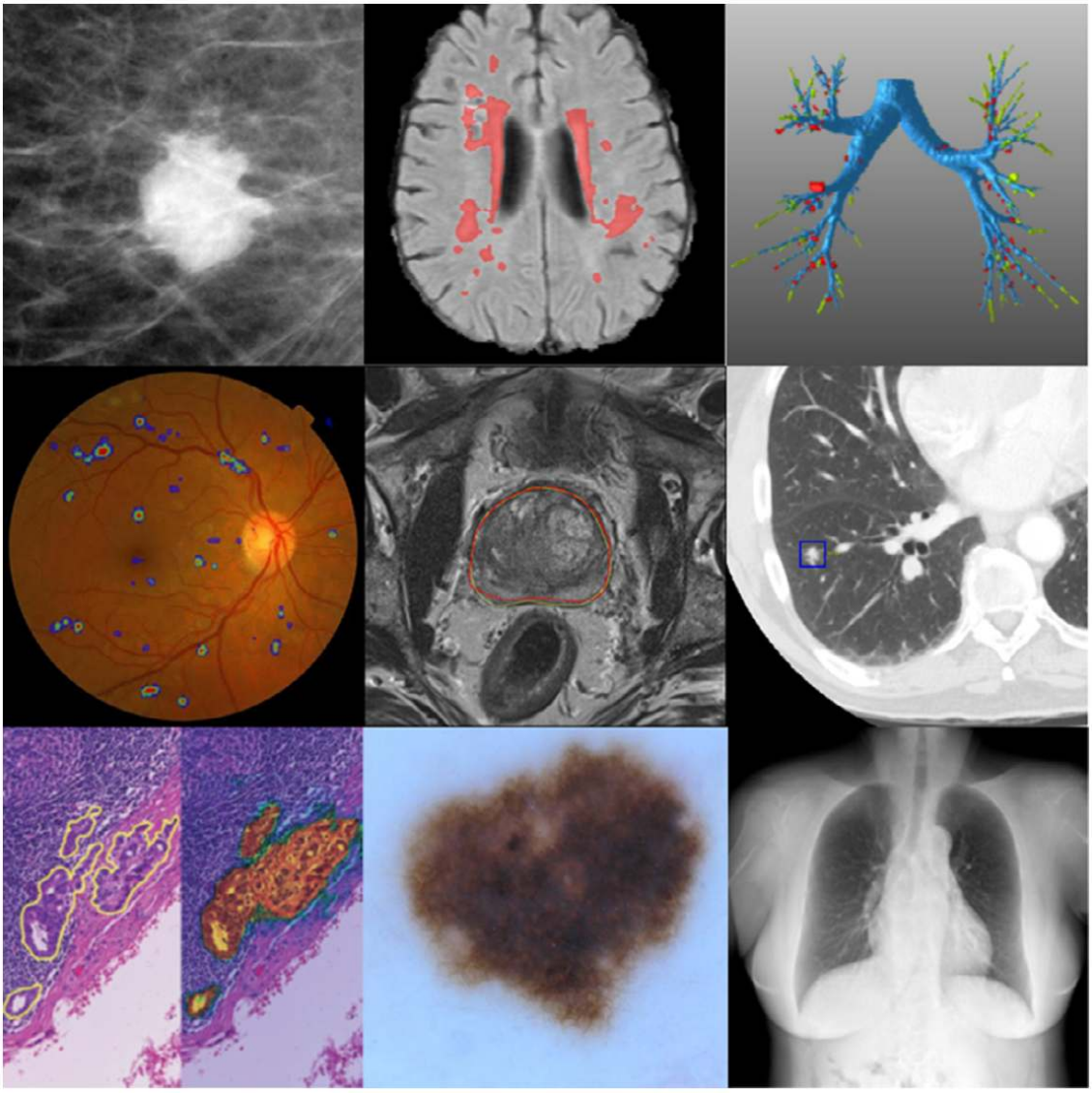


Figure 2.19: Visual example of some deep learning uses in medical image analysis [181].

(top ranking and human expert performance in CAMELYON16), human expert performance in skin lesion classification (Esteva et al. [202]), and state-of-the-art bone suppression in x-rays, image from Yang et al. [203]).

2.3.7 Deep learning applications for brain image analysis

In brain image analysis, deep neural networks (DNNs) have been used in many different application domains [204]. Nowadays, deep CNN architectures are widely used in brain MRI. In brain preprocessing tasks, it has been applied by Kleesiek et al. [205] for brain extraction, Nie et al. [206] for image construction and Yang et al. [207] for image registration. In brain segmentation tasks, it has been applied by Guo et al. [208] for hippocampus segmentation, Zhang et al. [96] for tissue segmentation, Brosch et al. [209] for lesion segmentation and Kamnitsas et al. [210] for tumor segmentation. Moreover, DNNs recently been applied for disorder classification by Suk and Shen [211], Sarraf and Tofighi [212], and Shi et al. [213]. Even though brain images are 3D volumes in all surveyed studies, some methods work in 2D, analyzing the 3D volumes slice-by-slice. This may be because of the reduced computational requirements or the thick slices relative to in-plane resolution in some data sets. More recent publications had also employed 3D networks [181].

Regarding brain image analysis challenges, DNNs have completely taken over many of them. From 2014-2018 BRATS, the 2015 longitudinal multiple sclerosis lesion segmentation challenge, from 2013-2018 ISLES, and the 2013 MRBrains, the top ranking methods were DL-based ones.

2.4 Discussion

As seen from this chapter, MRI is an essential tool for the diagnosis and evaluation of MS. Lesion detection approaches are required to detect static lesions and for diagnostic purposes, while either quantification of detected lesions or change detection algorithms are needed to follow up MS patients. In this latter case, deformation field-based algorithms have allowed the mass effect of the lesions to be considered. From the reviewed literature, it is clear the need to improve the performance of such methods, trying to make them more robust and accurate. We would like to point out the importance of using prior knowledge to guide the lesion detection and segmentation. Supervised approaches that rely on similar segmented cases usually outperform unsupervised strategies. We believe that basic supervised ML approaches based on DF provide a good starting point for new MS lesions detection in longitudinal analysis. We will study the effect of different DF operators combined with voxel intensities and subtraction images to improve lesion detection on MS patients. As DL techniques are now the state-of-the-art machine learning methods in medical image analysis, we will also explore and investigate DL techniques for new lesion detection. To the best of our knowledge, these is no longitudinal approach based on CNN that deals with lesion changes in brain MRIs. Other longitudinal

approaches based on CNNs have been presented before [134], but those methods independently provide a cross-sectional segmentation of lesions at each time point using longitudinal information. In the following two chapters, we propose two novel supervised new T2-w MS lesion detection approaches. One is based on a classical LR model combining intensity and DF features, and the other one is based on the use of a deep learning (CNN) strategy.

CHAPTER 3

A LOGISTIC REGRESSION MODEL FOR NEW T2-W LESION DETECTION IN MULTIPLE SCLEROSIS

3.1 Overview

As described in chapter 2, MRI allows to demonstrate with high specificity and sensitivity the dissemination of WM lesions in space and time, a key factor in recent diagnostic criteria [12]. Furthermore, quantification of new T2-w lesions is a high-impact prognostic factor to predict evolution to MS or risk of disability accumulation over time [214].

Analyzing the state-of-the art on MS lesion detection approaches in section 2.2.3, we concluded that the effect of a lesion does not always appear as an intensity change on the tissue where it is located (the so-called tissue transformation), but can also influence the appearance of surrounding tissues (known as the mass effect). Observing the lesion evolution without change in intensity but with displacement on the surrounding tissues (deformation) is also important. So, both tissue transformation (changes in intensity) and tissue deformation generally occur. Following this fact, we propose here to merge intensity- and deformation-based approaches in an automated multi-channel supervised voxel-wise LR classification. In contrast with the previous supervised approaches like the one of Sweeney et al. [156] that uses only intensity features, our model will use features not only from the baseline, follow-up, and subtraction images but also from the DF operators obtained from the nonrigid registration between time-points scans.

3.2 Methods

Figure 3.1 depicts the whole pipeline proposed for the detection of new T2-w lesions. For each modality (T1-w, T2-w, PD-w, and FLAIR), an affine transformation from baseline to follow-up is computed and the images are subtracted. Also, the images are nonrigidly registered to get a deformation field. Afterwards, the baseline and follow-up intensities, the subtraction values, and the DF features are used to train a voxel-wise LR classifier. In the post-processing, the probabilistic maps are thresholded to obtain a binary segmentation. In what follows, each step is explained in more details.

3.2.1 Registration and subtraction

For each patient, T1-w and FLAIR images from the same study are registered to the PD-w image by using a 3D multi-stage multi-resolution registration approach. First, a 3D rigid registration with only one resolution level is performed. Then, a 3D affine registration is performed with 3 levels of resolution. Both registration methods are done using ITK v4 framework [215]. The Mattes Mutual Information cost function is minimized by Regular Step Gradient Descent Optimization, and resampling is performed by B-spline interpolation.

To perform the image subtraction, the baseline images are warped to the follow-up space. The same 3D multi-stage multi-resolution registration approach described above is used. The affine transformation is computed between both PD-w images and then applied to the other images (using B-spline interpolation) to compute the subtraction. To avoid interpolation more than one, baseline T1-w and FLAIR are resampled using the combined affine transformation.

Since multi-channel data increases the probability of lesion activity detection [216], the four images (T1-w, T2-w, PD-w, and FLAIR) are subtracted after the affine registration. As stated in Díez et al. [35], the rigid and affine registration methods are not sensible to the presence of lesions, and only deformation models can show the effect of new lesions as a distortion around those regions. DF can be obtained by using a nonrigid registration technique. In this study, we apply the multi-resolution Demons registration approach from ITK v.4 initialized with the previous affine transformations [217]. This algorithm can produce large localized deformations and has been widely used in brain MRI.

To be able to incorporate the DF information as features, we compute the following three DF operators at each voxel [33]:

- **Jacobian** [157]: This operator is widely used in continuum mechanics [218] and it represents the local volume variation.
- **Divergence** [150]: This DF operator represent the volume density of the outward flux of a vector field from an indefinitely small volume around a given point.

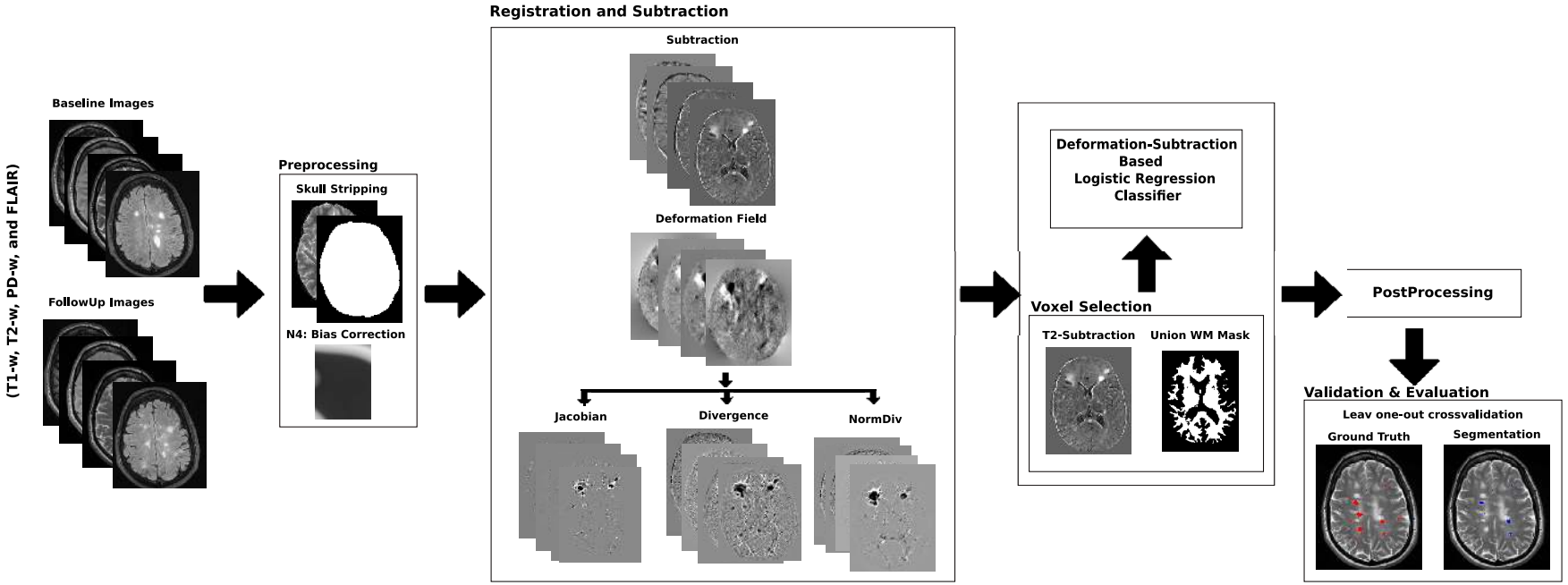


Figure 3.1: Scheme of the new T2-w MS lesion detection pipeline. The preprocessing in both baseline and follow-up for every modality (T1-w, T2-w, PD-w, and FLAIR) consists of ROBEX skull stripping, N4 bias field correction, and Nyúl histogram matching. For each modality, an affine transformation from baseline to follow-up is computed and the images are subtracted. Also, the images are nonrigidly registered to get a deformation field. Afterwards, the baseline and follow-up intensities, the subtraction values, and the DF features are used to train a LR classifier. In the post-processing, the probabilistic maps are thresholded to obtain a binary segmentation where all lesions smaller than three voxels are removed.

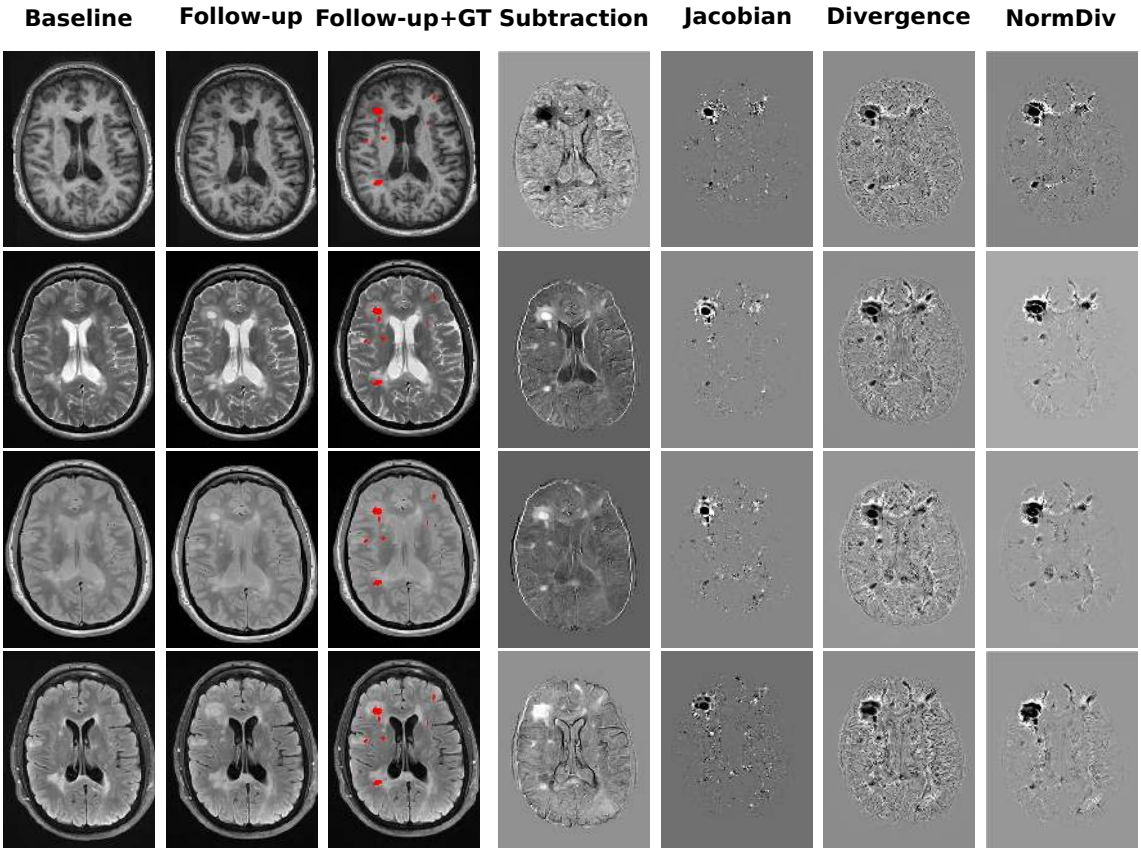


Figure 3.2: Relationship among baseline, follow-up, GT, subtraction image and the DF operators (Jacobian, Divergence, and NormDiv) of the four modalities. From top to bottom, each row represents T1-w, T2-w, PD-w, and FLAIR respectively. All the images are both from the same patient and slice. The GT is overlaid in red in the third column.

- **NormDiv**: This operator is the multiplication of the divergence and the norm of the DF. As successfully tested by Thirion and Calmon [150], this operator helps in the detection of active lesions.

Figure 3.2 shows a slice example of the baseline, follow-up, and subtraction image, and the DF operators (Jacobian, Divergence, and NormDiv) with the Ground Truth (GT) overlaid in red.

3.2.2 Deformation-subtraction based LR model

Our model uses a voxel-wise LR classifier like the work of Sweeney et al. [156] to predict the lesion probability of each voxel using the baseline and follow-up intensities, subtraction values, and the DF operators for T1-w, T2-w, PD-w, and FLAIR images. To train the model, we perform a voxel selection step in where candidate voxels that are likely to be part of a new lesion are selected to decrease the number

of training samples. As new lesions appear hyperintense in T2-w subtraction images, we only train the LR model with those candidate voxels. Some regions may have a high intensity in the subtraction images as a result of noise, inhomogeneity, registration errors, or small anatomic differences. To avoid that, the T2-w subtraction images are smoothed with a Gaussian kernel and only voxels with a subtraction value larger than the mean are included as candidates. As the aim of the study is to detect new T2-w lesions inside WM, a WM mask is used to limit the region of interest. This WM mask is computed with an automated atlas-based multi-channel tissue-segmentation algorithm [24] applied to the baseline and follow-up images before registration. This algorithm uses an expectation maximization algorithm to maximize the log-likelihood between the real MRI data and a Gaussian model of four classes: the pure tissue classes (WM, GM, and CSF) and a partial volume class (GM/CSF). For pure tissue classes, prior probabilities are provided by an atlas, while for the partial volume class, a weighted atlas of CSF and GM is used. Afterwards, lesions are segmented by applying a threshold on the FLAIR image. For each time point, we combine the WM mask and the lesion mask to obtain both a baseline and a follow-up mask. Even though new and enlarging lesions may be misclassified in the follow-up WM mask, these voxels should appear as normal WM in the baseline image. After registering the baseline WM mask to the follow-up space, a final WM mask is obtained as the union of the baseline and follow-up WM masks in the follow-up space. After the voxel selection step, a LR model is fitted over these candidate voxels.

3.3 Experimental setup

3.3.1 Datasets

VH dataset: The database used in this chapter consists of images from 60 different patients with a CIS or early relapsing MS who underwent brain MRI in the Vall d'Hebron Hospital's center for monitoring disease evolution and treatment response. Each patient underwent brain MRI within the first 3 months after the onset of symptoms (baseline) and at 12 months' follow-up after the onset. Thirty-six of the patients (13 women and 23 men; 35.4 ± 7.1 years of age) confirmed MS with new T2-w lesions, while 24 patients did not present new T2-w lesions. The baseline and follow-up scans for all patients were obtained in the same 3T magnet (Tim Trio; Siemens, Erlangen, Germany) with a 12-channel phased array head coil. The MRI protocol included the following sequences: 1) transverse proton density (PD)- and T2-w fast spin-echo (TR = 3080 ms, TE = $21 - 91$ ms, voxel size $= 0.78 \times 0.78 \times 3.0$ mm^3), 2) transverse fast FLAIR (TR = 9000 ms, TE = 87 ms, TI = 2500 ms, flip angle = 120°, voxel size = $0.49 \times 0.49 \times 3.0$ mm^3), and 3) sagittal T1- weighted 3D magnetization-prepared rapid acquisition of gradient echo (TR = 2300 ms, TE = 2.98 ms, TI = 900 ms, voxel size = $1.0 \times 1.0 \times 1.2$ mm^3). The Vall d'Hebron Hospital's ethics committee approved the study, and written informed

consent was signed by the participating patients.

Only new T2-w lesions that were visually detected on the follow-up scan were annotated on the PD-w images and semiautomatically delineated using Jim 5.0 software[1]. First, an expert neuroradiologist detected changes visually by using baseline and follow-up scans, and then a trained technician delineated them semiautomatically by using the subtraction image. The raters always annotated the complete new lesion and only the new part of the lesion in the case of large lesion growth. The dataset used in our study contained only two growing lesions, and the remaining were new lesions. Finally, the expert neuroradiologist confirmed the final segmentation. This analysis was used as the reference standard for comparison. The 36 patients with new T2-w lesions exhibited a total of 191 lesions. The lesions were distributed as 15.15% small (3-10 voxels), 53.53% medium (11-50 voxels), and 31.31% large (more than 50 voxels).

Preprocessing: We followed the main preprocessing steps described in 2.2.1. For each patient, the same preprocessing steps were performed on both baseline and follow-up images. First, a brain mask was identified and delineated on the PD-w image using the ROBEX Tool[2] [50]. Second, the four images underwent a bias field correction step using the N4 algorithm from the ITK library[3] with the standard parameters for a maximum of 400 iterations [219]. The T1-w and FLAIR images were linearly registered to the PD-w using Nifty Reg tools[4] [220, 221]. Finally, the baseline and the follow-up intensity values were normalized per modality and per patient (i.e., between the baseline and the follow-up scans, and not across the entire dataset) using a histogram matching approach based on Nyúl et al. [61] [5].

3.3.2 Evaluation

We evaluated the proposed framework in two scenarios. Firstly, we analyzed the accuracy of the detection using a leave-one-out cross-validation strategy with the 36 patients with new MS lesions. This strategy was applied per patient on our 36 images from the MS patient dataset. From all these images, the candidate voxels were around four million, including only 13 thousand voxels per lesions while the rest were negative samples. The classifier was trained using 35 patients and tested with the remaining one. This process was repeated until all patient images were used as a test image. Secondly, we analyzed the specificity of the method with the 24 patients with no new T2-w lesions. To do this, we performed a new training using all the 36 images with new MS lesions. We compared also the obtained results with those of recent state-of-the-art approaches [32, 33, 156].

Standard measures such as the true positive fraction (TPF), the false positive

[1]http://www.xinapse.com/home.php
[2]https://www.nitrc.org/projects/robex
[3]https://itk.org/Doxygen/html/classitk_1_1N4BiasFieldCorrectionImageFilter.html
[4]https://sourceforge.net/projects/niftyreg/
[5]https://itk.org/Doxygen/html/classitk_1_1HistogramMatchingImageFilter.html

fraction (FPF), and the Dice similarity coefficient (DSC), which were computed as follows, were used for the evaluation:

$$TPF = \frac{TP}{TP + FN}$$

$$FPF = \frac{FP}{FP + TP}$$

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where TP, FN, and FP are the number of true positives, false negatives, and false positives, respectively. In terms of detection, a lesion was considered as a TP if there was at least one voxel overlapping. In terms of segmentation, only the voxel-wise DSC was computed.

To depict the impact of both the deformation field operators and the baseline intensities features in the detection and segmentation of new T2-w lesions, we analyzed the following models:

- **LR-DF** (Logistic Regression with DF): This is our main model which uses the four image intensities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up, the subtraction values, and the DF operators (Jacobian, Divergence, and NormDiv) per voxel.
- **LR-NDF** (Logistic Regression without DF): This model incorporates the four image intensities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up and the subtraction values per voxel but DF are not used. This model is used for comparison with **LR-DF** to highlight the impact of the DF operators.
- **LR-DFNB** (Logistic Regression with DF without Baseline): This model uses the four image intensities (T1-w, T2-w, PD-w, and FLAIR) in only the follow-up images, the subtraction values, and the DF operators (Jacobian, Divergence, and NormDiv) per voxel. This model is used for comparison with **LR-DF** to highlight the impact of the baseline intensities.
- **LR-NDFNB** (Logistic Regression without DF without Baseline): This model uses the four image intensities (T1-w, T2-w, PD-w, and FLAIR) in only follow-up images and the subtraction values per voxel. This model is used for comparison with **LR-NDF** to highlight the impact of the baseline intensities in the absence of DF operators. This model corresponds to our implementation of the approach proposed by Sweeney et al. [156].

Moreover, similarly to the works of Ganiler et al. [32] and Cabezas et al. [33], we studied the performance of the model according to the different lesion sizes. We

analyzed the same categories, where lesions of $[3-10]$ voxels were considered small, lesions of $[11-50]$ voxels were considered medium, and lesions of $+50$ voxels were considered large. This division is useful to analyze the effect of the deformation fields on different lesion sizes.

3.3.3 Postprocessing

After training the model, we create 3D maps of the estimated lesion probability at each voxel. As done by Sweeney et al. [156], we smooth these maps with Gaussian kernels mainly to decrease noise and to remove some small FP lesions. The smoothed probabilistic maps are thresholded to get the final binary lesion segmentation. The threshold is empirically selected as the best trade-off between sensitivity (i.e. TPF) and specificity (i.e. $1-FPF$). Specifically, the best threshold is obtained by the higher F-score value of both measures, which gives the harmonic mean of the measures and is computed as:

$$\text{F-score} = 2\,\frac{TPF * (1-FPF)}{TPF + (1-FPF)}$$

A more detailed description is provided in the results section, showing how this selection is done and the effect of using different probability thresholds. Moreover, all lesions with size lower than 3 voxels are removed from the generated masks.

3.3.4 Statistical analysis

The statistical significance of the performance between proposed methods was computed by running a series of permutation tests between the DSCs (Segmentation) and DSCd (Detection) obtained by each method [39, 222]. Permutation tests select random subsets of independent subjects of the dataset, and for each pair of methods, perform all possible permutations of their values in the corresponding subset, counting the number of times that the differences of one method are significant with respect to the other with ($p \leq 0.05$). After repeating this process over a number of iterations S, the mean and standard deviation (μ_0, σ_0) of the fraction of times when each method produced significant p-values is calculated over all the iterations. With this approach, methods with higher means indicate a higher significance of their reported values. The methods were then ranked into three different levels according to the difference between the mean score of the best method $\mu_0 \pm \sigma_0$ and the distance with respect to the mean scores of the rest of the methods. Hence, Rank 1 contained methods with mean scores of $(\mu_0 - \sigma_0,\ \mu_0]$, Rank 2 contained those with mean scores of $(\mu_0 - 2\sigma_0,\ \mu_0 - \sigma_0]$ and Rank 3 those in the interval $(\mu_0 - 3\sigma_0,\ \mu_0 - 2\sigma_0]$. For all the tests, we set the number of comparisons between each pair of methods to S = 1000.

Table 3.1: Lesion detection results: Comparison between the different models evaluated. Results stand for mean detection TPF, FPF, DSC_d and mean segmentation DSC_s.

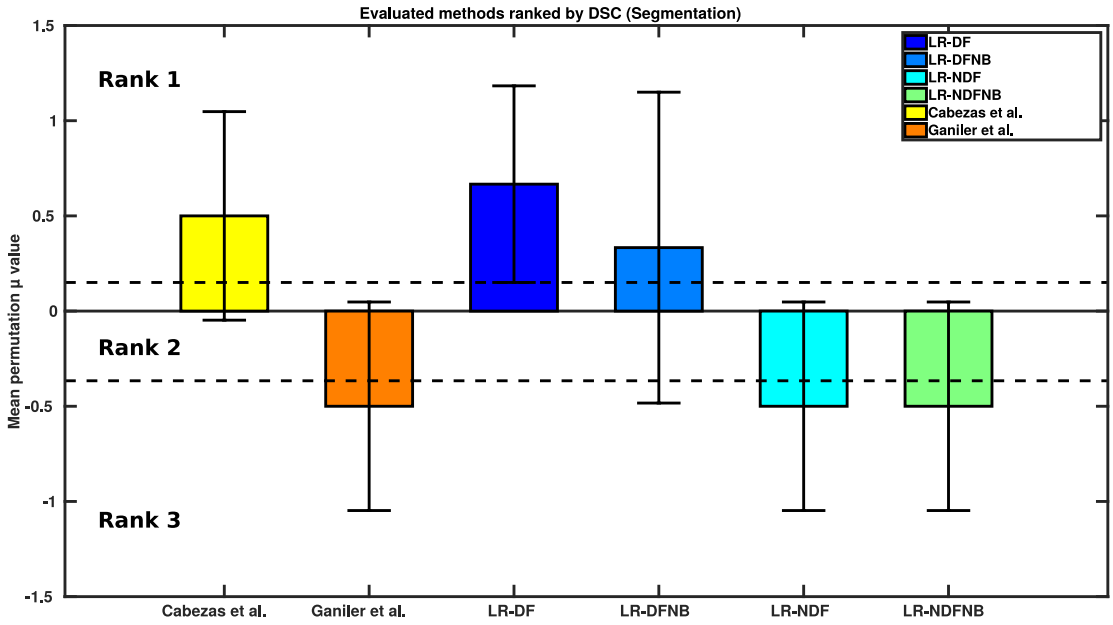
Method	TPF	FPF	DSC_d	DSC_s
LR-NDFNB	48.69	16.78	0.54	0.38
LR-NDF	48.46	13.90	0.54	0.39
LR-DFNB	69.88	11.94	0.74	0.52
LR-DF	**74.30**	**11.86**	**0.77**	**0.56**
Ganiler et al. [32]	51.62	35.87	0.46	0.37
Cabezas et al. [33]	70.93	17.80	0.68	0.52

Additionally, the Pearson's correlation coefficient was also used to analyze the linear relationship between manual annotations and the automatic detections obtained with our approach.

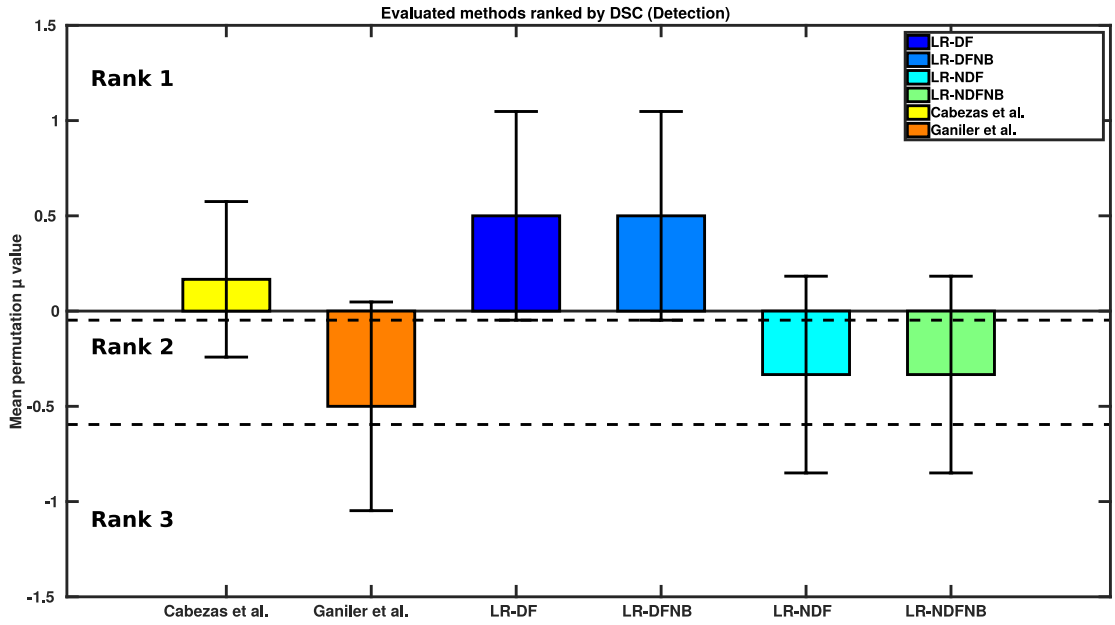
3.4 Results

Table 3.1 summarizes the new T2-w lesion detection and segmentation mean results for our full model (LR-DF), and the three variants with less features (LR-DFNB, LR-NDF, LR-NDFNB). We also included two state-of-the-art approaches for comparison [32, 33] when analyzing the 36 MS patients. Notice that our full model outperformed all the other approaches and had the best values for all the evaluation measures. The lower value of $DSCs$ is because many lesions have a small size which has a high effect on the $DSCs$. Figures 3.3(a) and 3.3(b) visually show the result of the permutation tests for the segmentation and the detection DSC values, respectively. Permutation tests permit to compute the exact P-value, and are not limited by any statistical distribution or minimum number of subjects. Essentially, each method is compared against all others using randomly selected subsets of data using statistical difference-of-mean test that do not require data to follow the normality condition. Notice that the data variability is still present in the fact that mean values obtained by all methods are not too high (best methods obtain $\mu_{Detection} = 0.50$ and $\mu_{Segmentation} = 0.67$). It is, however, possible to see how some methods do better than the other in pairwise comparisons that bear statistical significance. Notice that the methods in rank 1 included only approaches that used DF-based features, whereas non-DF based approaches were placed in ranks 2 and 3. Because ranking between the approaches differed, we can conclude that there is a significant difference in performance when including DFs.

Analyzing the results per patient, we had 12 patients with a TPF of 100% and FPF of 0%, and 5 patients with a TPF of 100% and less than a 33.33% of FPF. The worst cases we had were 3 patients with a TPF lower than 30%. Those patients had mainly small lesions ($[3 - 10]$ voxels) that the pipeline failed to detect.



(a)



(b)

Figure 3.3: Permutation test results for the evaluated methods. Final ranks based on (a) the DSC (Segmentation) and (b) the DSC (Detection).

Table 3.2: Analysis of TPF for different classifiers for different lesion sizes. Lesions between 3 and 10 voxels are considered small; lesions between 11 and 50 voxels, medium; and lesions with 50 voxels, large

Method	3 - 10	11 - 50	+50
LR-NDFNB	11.76	40.84	77.80
LR-NDF	11.76	40.83	77.80
LR-DFNB	28.13	61.52	91.24
LR-DF	34.40	**65.70**	**91.30**
Cabezas et al. [33]	**42.86**	48.57	77.42

Figure 3.4 shows the correlation between the number of new lesions manually annotated and the automatically detected and also the correlation between lesion volume in the GT and the automatically segmented. Significant Pearson's correlation ($R = 0.85$; $Pvalue < 0.00001$; confidence band = 95%) and ($R = 0.87$; $Pvalue < 0.00001$; confidence band = 95%) were found, respectively, between annotations based on visual detection (GT) and our approach (only LR-DF). Regarding the number of the data points used, all the MS patients with lesion progression were used for this correlation (36 data points - 36 patients), but different patients had the same number of GT and automatically detected lesions. Therefore, several points are overlapping in the plot. For example, there are 5, 6, and 4 cases with (2 GT lesions, 2 detected lesions), (1 GT lesion, 1 detected lesion), and (3 GT lesions, 2 detected lesions), respectively. Notice that there are numerous cases in which the number of new lesions per patient is actually very small.

Table 3.2 summarizes the performance of our pipeline according to the different lesion sizes described in Section 3.3.2. The LR-DF model had a better performance than LR-NDFNB and LR-NDF in all lesion size categories, although the results with small lesions had a worse performance when compared with larger lesions. Moreover, LR-DF had also a better performance than Cabezas et al. [33] for medium and large lesion size categories.

The selection of the Gaussian smoothing σ and the threshold value in the postprocessing step was done by maximizing the F-score of TPF and FPF using a leave-one-out cross validation, obtaining the results shown in Figure 3.5. The leave-one-out cross validation was applied per patient on our 36 patients with MS dataset. Notice that increasing σ requires decreasing the threshold value to obtain better results. The highest F-score value was obtained with $\sigma = 0.75$ and threshold = 0.3. Table 3.3 shows how TPF, FPF, DSC (Detection), DSC (Segmentation), and F-score were varying based on changing the threshold on the probability maps smoothed with $\sigma = 0.75$. A higher TPF could be obtained by decreasing the threshold but obtaining a higher FPF. The threshold 0.3 was selected as the best trade-off between TPF and FPF, computed using the F-score value (Figure 3.5). Notice that this thresholding analysis should be also done when using different datasets acquired with different MRI scanners and image protocols to optimize the obtained results. In order to evaluate the effect of postprocessing, we tested also our approach without using it, so

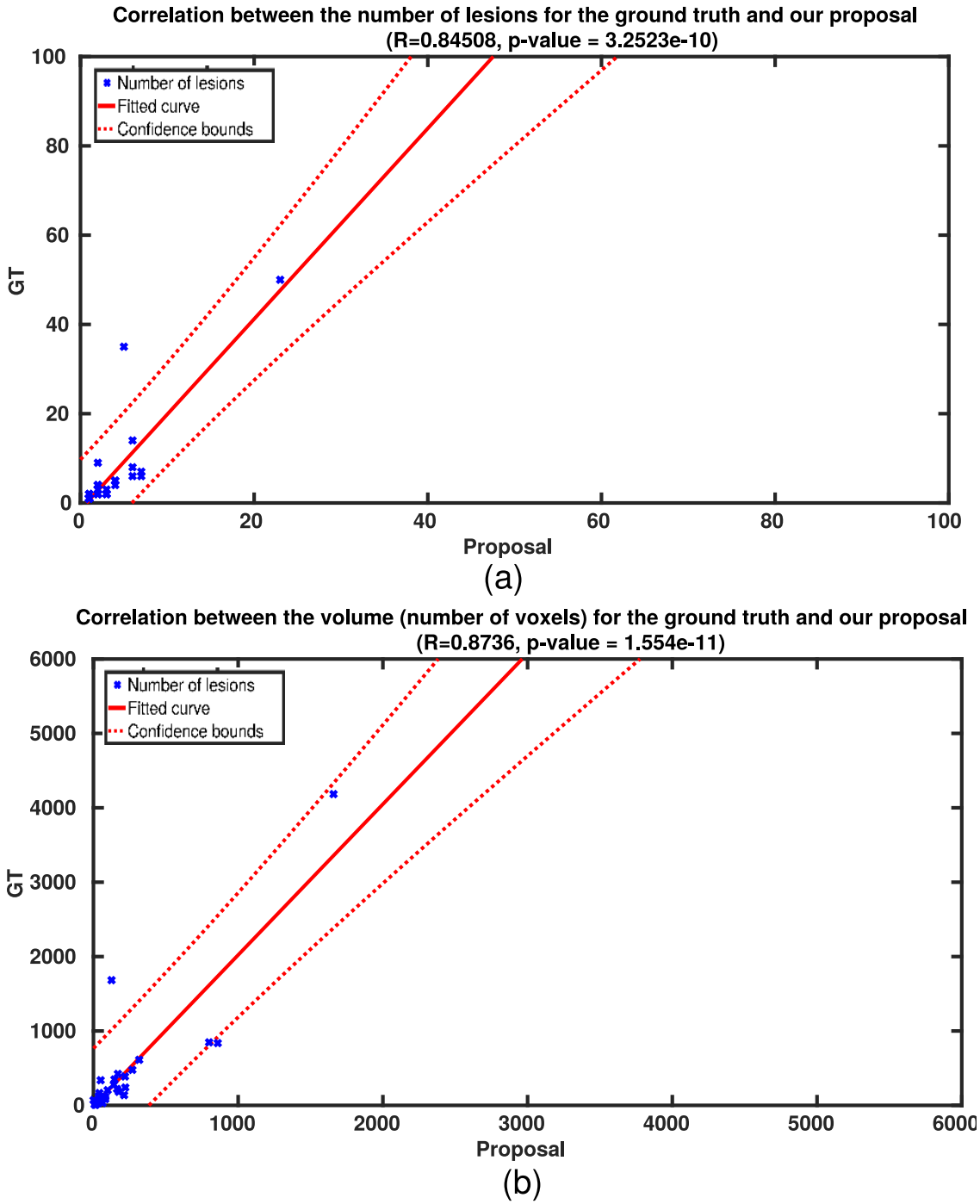


Figure 3.4: Correlation between (a) the number of GT lesions and the number of automatically detected ones using the proposed LR-DF model (Pearson's coefficient $R = 0.85, P_{value} < 3.25e^{-10}$) and (b) the volume (the number of voxels) of GT lesions and the volume of automatically detected ones using the proposed LR-DF model (Pearson's coefficient $R = 0.87, P_{value} < 1.55e^{-11}$). All the MS patients with lesion progression were used for this correlation (36 data points - 36 patients). Notice that different patients have the same combination of number of GT lesions and LR-DF detections. Therefore, several points are overlapping in the plot.

Table 3.3: The effect of varying probability thresholds after smoothing with $\sigma = 0.75$: Results stand for mean detection TPF, FPF, DSC_d, mean segmentation DSC_s, and F-Score. Best values based on F-Score are depicted in bold.

Threshold	TPF	FPF	DSC_d	DSC_s	F-score
0.0	99.26	99.01	0.05	0.007	0.02
0.1	86.84	43.40	0.64	0.49	0.685
0.2	82.53	22.52	0.77	0.57	0.799
0.3	**74.30**	**11.86**	**0.77**	**0.56**	**0.806**
0.4	57.83	6.32	0.65	0.43	0.715
0.5	46.16	6.15	0.54	0.30	0.619
0.6	31.80	6.17	0.40	0.18	0.475
0.7	17.53	3.40	0.24	0.10	0.296
0.8	9.14	0.0	0.12	0.05	0.168
0.9	7.78	0.0	0.09	0.02	0.144
1.0	0.0	0.0	0.0	0.0	0.0

no smoothing was applied and the class with the highest probability was selected (argmax). The results showed better TPF values but with more FPF, especially in those cases with smaller lesions.

Finally, we evaluate the 24 patients with no new T2-w lesions, after training the LR-DF model with all the 36 patients with new T2-w lesions. This allows to clearly state the specificity of our pipeline. Only 5 FP detections where found (in 4 cases) with a total size of 40 voxels.

Figure 3.6 shows a visual example of the performance of our pipeline, where each column corresponds to the baseline T2-w image, follow-up T2-w image, the GT, and the results obtained by LR-DF, LR-NDF, and LR-NDFNB approaches, respectively.

3.5 Discussion

The pipeline proposed in this chapter is fully automated, simple and adjustable to the application in terms of sensitivity and specificity. In order to improve the classifier accuracy, we added DF operators to the approach of Sweeney et al. [156]. As suggested by Cabezas et al. [33], the DF helps to reduce the detection errors due to local inhomogeneities and small changes that affect the accuracy of the subtraction pipelines.

As lesions are clusters of voxels and our approach is a voxel-wise pipeline, spatial information between voxels should also be included in our model. Although the model was not trained with standard spatial features or textures, the neighboring information between voxels was incorporated while smoothing the generated probability maps during the postprocessing step. Moreover, a registration technique that implements a free-form deformation also incorporates this local information into the resulting DF and provides better insight into changes occurring due to development

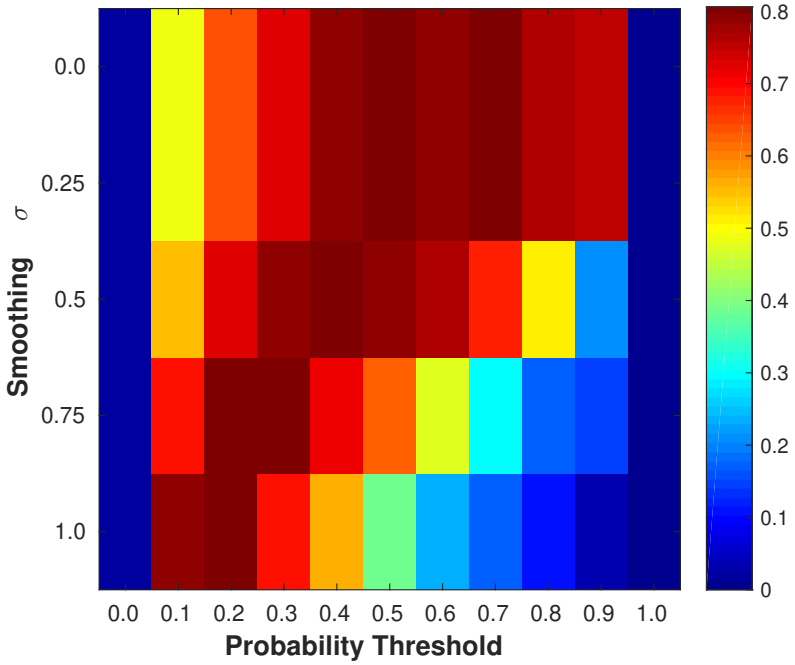


Figure 3.5: Parameter selection. The F-score values of TPF and FPF using leave-one-out cross validation. The maximum F-score was obtained with $\sigma = 0.75$ and $threshold = 0.3$.

of new or enlarging lesions. And, since they are computed using the gradient image of the DF, the DF operators encode spatial relationships too.

In the postprocessing step, we selected the parameters (Gaussian smoothing σ and threshold value) using the maximum F-score value but the pipeline can also be used without any parameters tuning by not smoothing the probability maps and selecting the class with the highest probability (using argmax). In that case, the pipeline had an increase in TPF (80.0%) and also FPF (21.87%) with the same DSC in segmentation and detection compared with our best configuration using postprocessing, mostly due to FPs eliminated by the Gaussian smoothing step in the latter. Because the voxel probabilities are decreased after smoothing, an increase in the smoothing σ value requires a decrease in the threshold value. There is a trade-off between the number of false positives and true positives. The smoothing also eliminates small regions that may be FPs or TPs. For instance, this step had a high impact in reducing the number of false positives in the 24 patients with no new T2-w lesions.

Our results show that the combination of DFs and supervised classification may help to increase the performance when detecting new T2-w lesions. In order to analyze the effect of DFs, we trained a logistic regression classifier with different features. We trained the model with different combinations of the baseline and follow-up intensities, the subtraction values and DF operators. Using only features

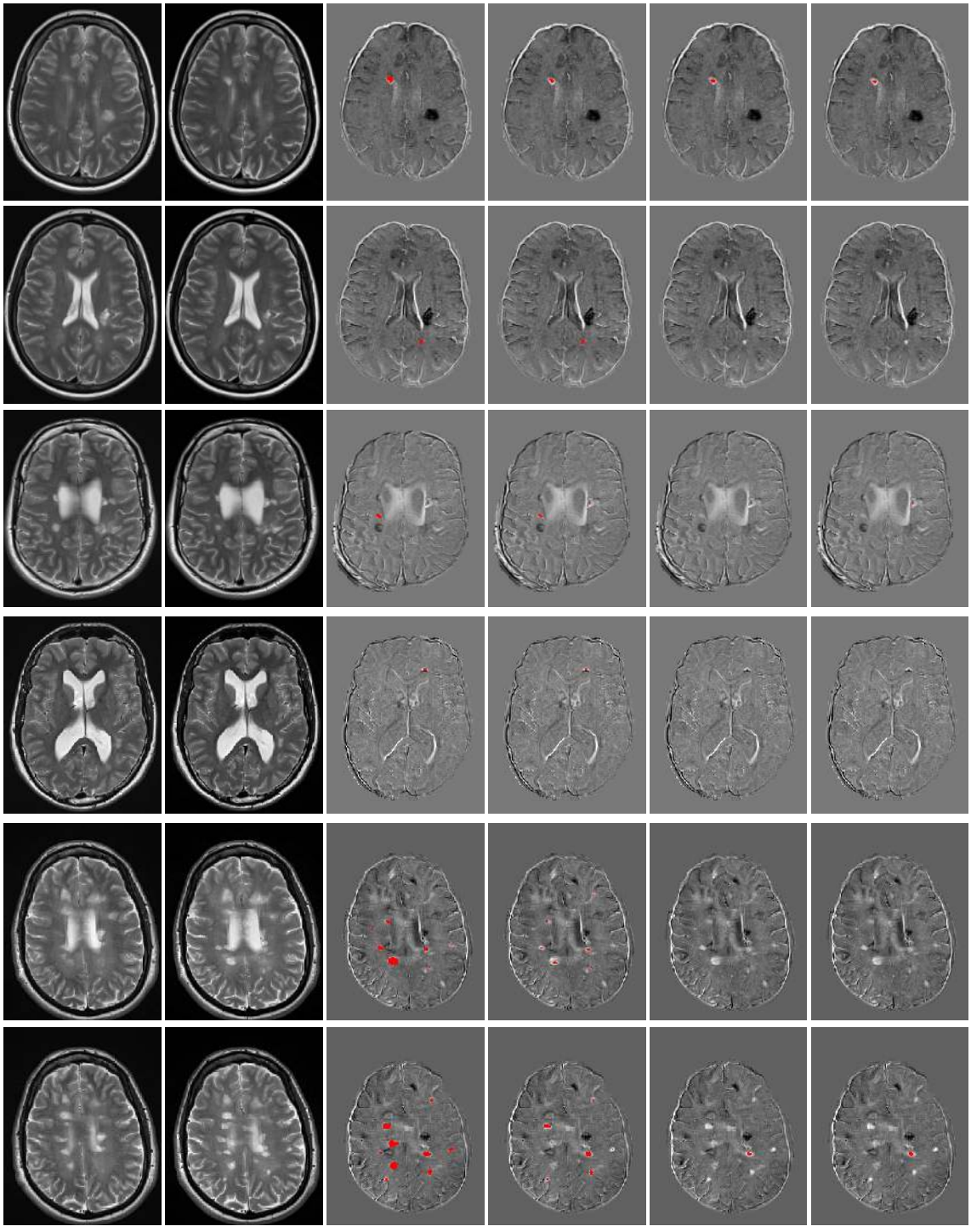


Figure 3.6: Examples of new MS lesions detection in a 12-month longitudinal analysis. Images (a) and (b) show one axial slice of T2-w image at baseline and follow-up, respectively. Image (c) shows the new MS lesions annotations performed by an expert (GT). Images (d), (e) and (f) show the segmentation of LR-DF, LR-NDF, and LR-NDFNB approaches, respectively. Notice that for this axial slice the LR-DF model could detect 7 lesions out of the 9 lesions in the GT. The two missed ones were actually detected in the adjacent slice.

from intensities within a lesion (baseline + follow-up) or subtraction could trigger the detection of new lesions. As mentioned in Table 3.1, the models which do not include DF features (LR-NDFNB and LR-NDF) could detect new lesions with TPF of 48.69% and 48.46% and FPF of 16.78% and 13.90% respectively. As in previous works [33], our results show that the addition of DFs helps to significantly increase the detection of new T2-w lesions while maintaining the number of false positives low. However, our model is capable of improving the results of other unsupervised methods due to the use of a supervised classification model instead of an unsupervised rule-base approach [32, 33]. Furthermore, these improved results are further backed by a strong correlation between the number of automatically detected lesions and the number of visually detected ones. This suggests that our automatic segmentation may help the radiologist to estimate the number of new lesions before annotation.

Given the difficulty to obtain MRI datasets with expert annotations, our evaluation dataset was composed of a single database of 60 MS patient images obtained with the same scanner and protocol. This limits the analysis of our model within a single image domain i.e. the model with the parameter configuration presented in this study evaluated with the image domain of our data set. Further tests and probably a specific parameter adjustment should be performed for optimizing the performance on different data sets acquired with different MRI scanner machines and different image protocols. Moreover, although the available data comprised MS patients with different lesion sizes, the volume of most of the new/enlarging T2-w lesions was relatively low. This can bias the results obtained by our approach, since we noticed that for small lesions, the pipeline had lower accuracy than for larger lesions. As the lesion size increases, the DFs are able to better represent these volume changes. In this regard, one could study the use of different strategies for each lesion size and combine the different outputs (i.e. probability maps) to improve the overall obtained results.

Our pipeline was only tested with the kind of images mentioned in the data section but this does not mean that the approach is limited to them. Further testing with images with different resolution (2D and 3D) and from different scanners and image protocols should be performed. Previous subtraction works such as Ganiler et al. [32] tested their subtraction pipeline with other scanners, image resolutions 2D for instance, and 1.5T and 3T and worked well. Although, one should tune properly the threshold in the postprocessing section for the best performance or use the pipeline without the postprocessing step (argmax).

The proposed approach in this chapter is a conventional machine-learning technique i.e, a feature extraction step is needed to extract important features from input images before the training process. The feature extractor should be designed and engineered carefully. For instance, our proposed method is based on the DF-based features (Jacobian, Divergence, and NormDiv) and the features from the subtraction between the baseline and the follow-up images. As discussed in 2.3.3, deep learning methods simplify the feature extraction process, and could gather unknown patterns to help in the desired task. For instance, as seen in section 2.3.4, CNN-based meth-

ods can be used to perform the segmentation of lesions. In CNNs, there is no need to compute features as a first step since the network effectively finds the important features as a part of its search process and eliminate the bias of testing only those features that we believe to be important. Moreover, the task of computing many features and then selecting those that seem to be the most important. In the next chapter, we will propose a FCNN-based approach for new T2-w MS lesion detection.

CHAPTER 4

A DEEP LEARNING MODEL FOR NEW T2-W LESION DETECTION IN MULTIPLE SCLEROSIS

4.1 Overview

Recently, deep neural networks have attracted substantial interest. CNN have demonstrated groundbreaking performance in brain imaging, especially in tissue segmentation [97, 223] and brain tumor segmentation [210, 224]. In contrast to previously supervised learning methods, CNNs do not require manual feature engineering or prior guidance. Furthermore, the increase in computing power makes them a very interesting alternative for automated lesion segmentation. CNN-based methods have achieved top ranking performance on all of the international MS lesion challenges [225, 226, 227, 228].

As described in section 2.2.3, in the deformation-based approaches, the new T2-w lesion detection is performed by analyzing DF obtained by nonrigid registration between successive scans [33, 150, 157]. Nonrigid registration and the use of DF between time points have been shown to improve the detection of new T2-w MS lesions in longitudinal studies [33, 34]. These DFs can either be obtained using classic nonrigid registration approaches based on optimization or, recently, using also learning-based approaches. In real cases, both tissue transformation (changes in intensity) and tissue deformation generally occur. Hence, the mass effect of the lesion should also be taken into account in order to define a precise lesion evolution. Deformation based approaches are sensitive to these changes in the brain. However, they do not provide information about stable lesions.

Classic registration approaches establish a dense nonlinear correspondence bet-

ween a pair of 3D brain scans. For these approaches, registration is defined as an optimization problem that needs to be solved for each volume pair using a similarity metric while enforcing smoothness constraints on the mapping. Solving this optimization is computationally intensive and therefore, extremely slow in practice [217, 229, 230, 231, 232, 233, 234]. However, different GPU-based accelerated approaches have been proposed to improve the efficiency and speed up the optimization [235, 236, 237].

Common learning-based approaches rely on classification algorithms to register the two scans. These algorithms involve a first stage in which a model is estimated on training data composed of a set of features and their corresponding ground truth and a second stage in which the model is tested on a new dataset to provide the desired results. Classic machine learning methods require hand-crafting feature vectors to extract appearance information [238]. In contrast, CNNs can learn a set of features that are specifically optimized for the current task directly from the image data. Currently, CNNs have demonstrated superior performance in brain imaging specifically for segmenting tissues [97, 239], brain tumors [210, 240, 241] and white matter lesions [39, 242]. In the case of registration approaches, learning-based methods learn a parametrized registration function from a collection of images during training. During testing, a registration field can be quickly computed by directly evaluating the function using the learned parameters. Some proposed methods [243, 244] rely on a precomputed DF as the ground truth, and the others rely only on the images being registered or segmentation masks, without comparing the expected deformation field with a precomputed deformation field [245, 246]. Specifically, Balakrishnan et al. [247] developed a new CNN approach that computes the deformation between two images by training the network using a similarity metric and a regularization term similar to classic registration methods, obtaining comparable results with current state-of-the-art approaches.

In this chapter, we propose an FCNN approach to detect new T2-w lesions in longitudinal brain MR images. The proposed model combines intensity-based and deformation-based features within an end-to-end deep learning approach.

4.2 Methods

4.2.1 Network architecture

Figure 4.1 shows the new T2-w MS lesion segmentation architecture. The proposed network is an FCNN that takes four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up as inputs and outputs the new T2-w lesion segmentation mask. The network consists of two parts. The first part is U-Net blocks that learn the DFs and nonlinearly register the baseline image to the follow-up image for each input modality. The learned DFs and the baseline and follow-up image modalities are then fed to the second part of the network, another U-Net that performs the final detection and segments the new T2-w lesions. The network is

trained end-to-end with gradient descent and simultaneously learns both DF and new T2-w lesion segments.

3D registration architecture: A 3D registration block is built for each input modality following the architecture shown in Figure 4.2(a). This block is inspired by the work of Balakrishnan et al. [247] (VoxelMorph), which is a learning framework for deformable medical image registration. The registration block learns the DF and nonlinearly registers the baseline image to the follow-up image. It is a fully convolutional network that follows a U-shaped architecture [191]. The U-Net architecture consists of four downsample (the contracting path) and upsample steps (the expansive path). The core element (CE) block is a two 3D convolution layer (kernel size = 3 and stride = 1) with K channels. Each convolution is followed by a LeakyReLU layer. The number of channels, K, of CE blocks are (64, 128, 256, and 512) and (512, 256, 128, and 64) for the contracting path and expansive path, respectively. The U-Net's downsampling followed by the upsampling and skip connections allow the network to exploit information at large spatial scales while retaining useful local information. Moreover, as discussed in Drozdzal et al. [248], skip connections facilitate gradient flow during training. The spatial transformation [247, 249] warps the baseline image to the follow-up space using the learned DF and enabling end-to-end training. The LeakyReLU activations are used instead of ReLU so that the learned DFs can have positive/negative values.

3D segmentation architecture: A 3D segmentation block is also used for segmenting the new T2-w lesions. It is a two-branch network where each branch is a U-Net following the architecture shown in Figure 4.2(b). The U-Net architecture is exactly the same as the U-Net used in the registration block but using a ReLU activation layer instead of the LeakyReLU layer. The inputs of the first branch are the four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up, while the second branch input is the four DFs learned from the first registration blocks. The outputs of the two branches are concatenated before the classification step. One UNetCore process the DFs (deformation-based) and another UnetCore process the baseline/follow-up modalities (intensity-based). Note that the model is merging the intensity with the DFs to segment the new lesions.

4.2.2 Loss functions

The loss function used in this work is the summation of two loss functions. One function is an unsupervised loss function that controls the registration part of the network [247]. It consists of two components: a similarity part that penalizes differences in appearance between the moved baseline and follow-up images and a regularization part that enforces a spatially smooth deformation and is often modeled as a linear operator on the spatial gradients of DF as stated in [247]. Therefore, the registration block is trained in an unsupervised manner using the spatial transform block which is used to warp the baseline image to the follow-up space using the learned DF. The block learns the DF by minimizing the mean square error (MSE) between the warped baseline and the follow-up images during training. The other function

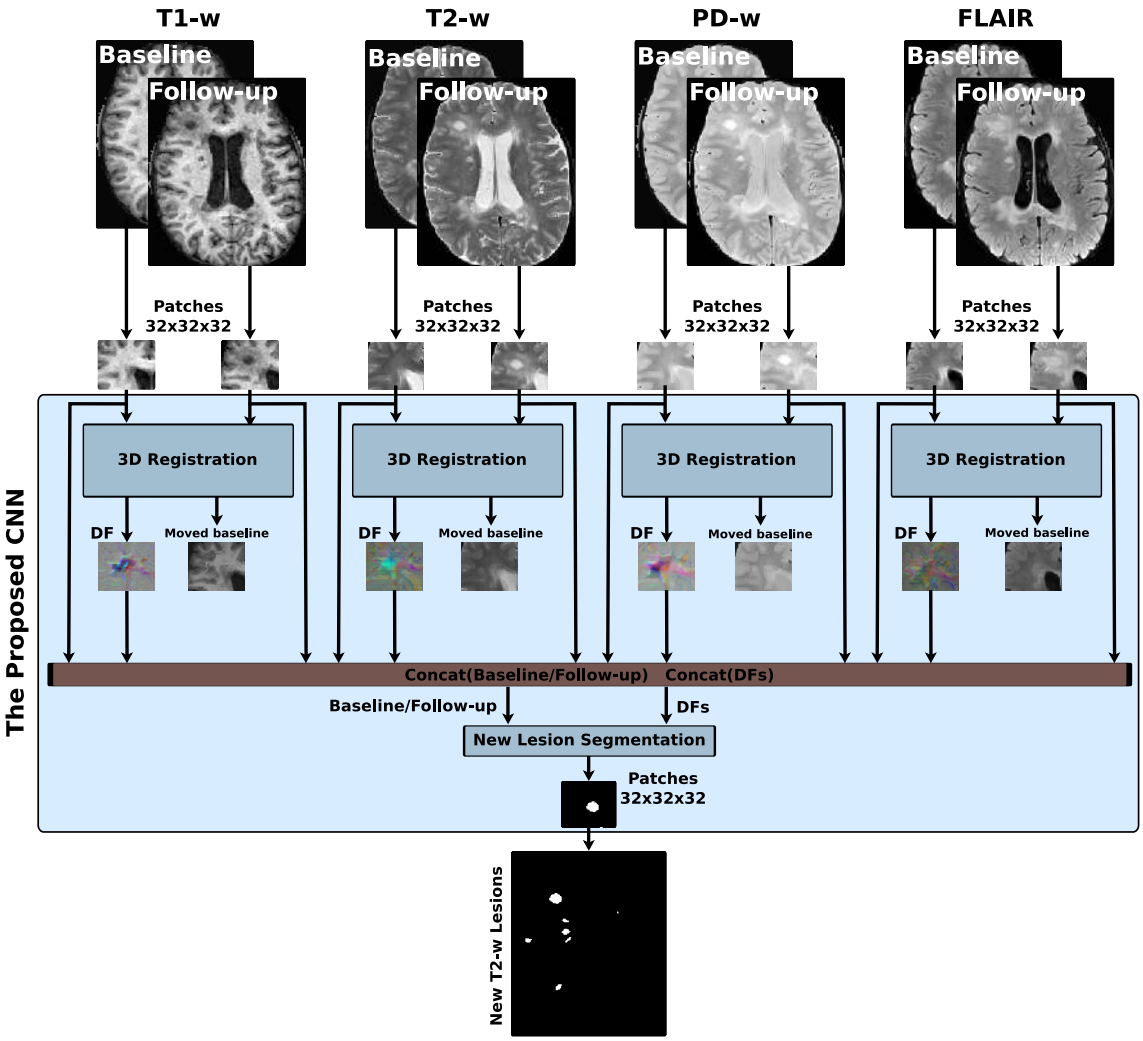
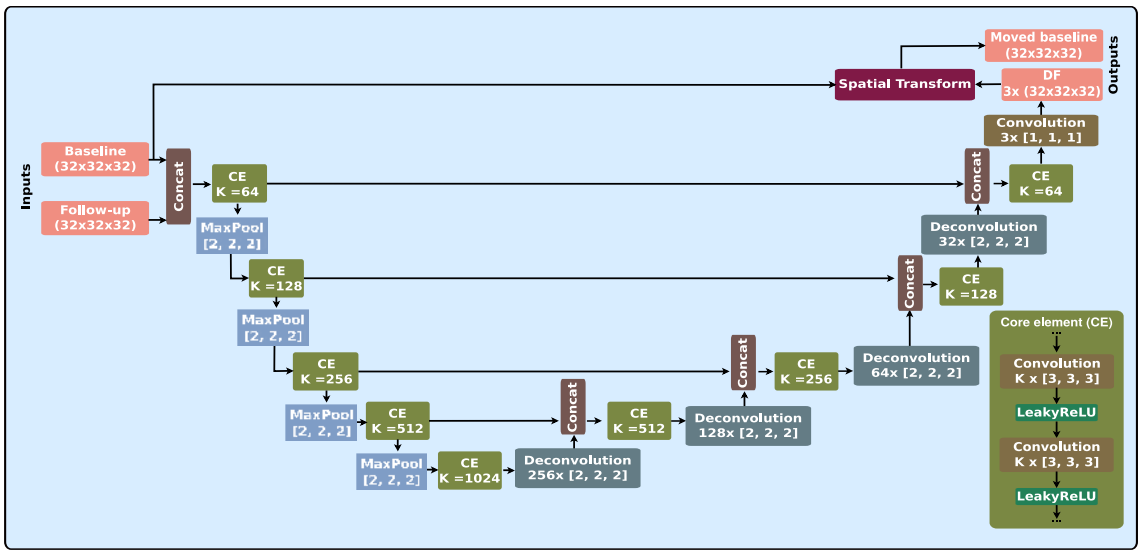
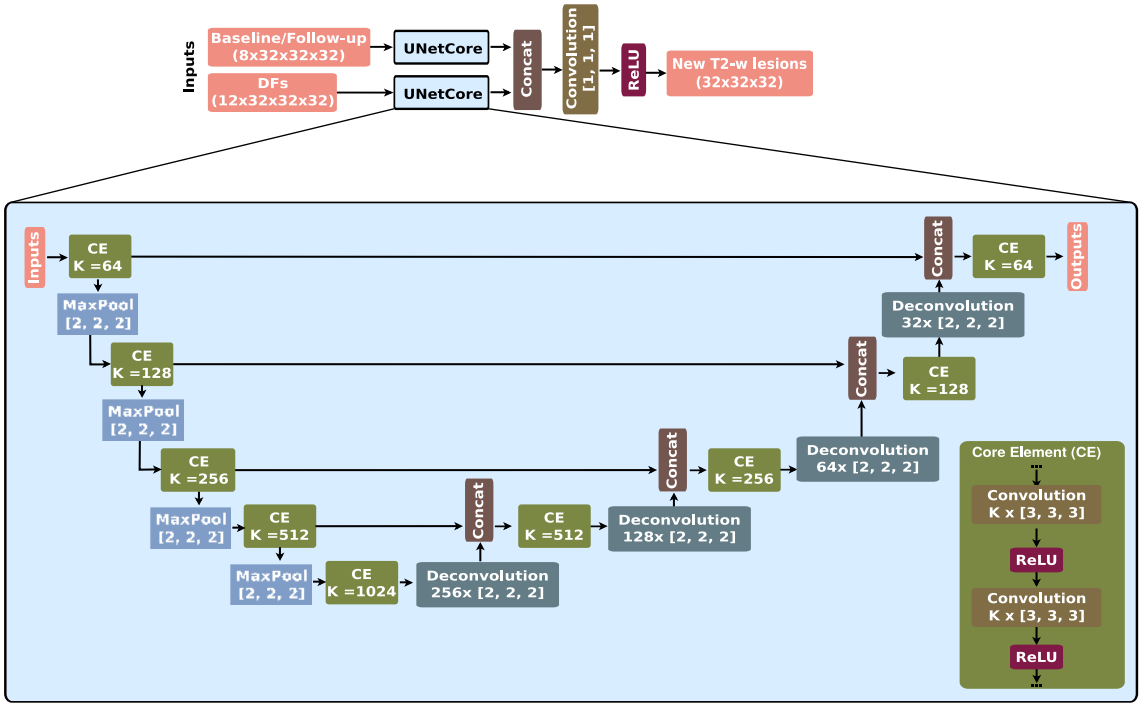


Figure 4.1: Scheme of the new T2-w MS lesion segmentation network. The proposed network consists of four 3D registration blocks and one 3D segmentation block. The inputs are baseline/follow-up images of the T1-w, T2-w, PD-w, and FLAIR. For each input modality, there is a 3D registration block that learns the deformation field (DF) and nonlinearly registers the baseline image to the follow-up image. Afterwards, the learned DFs and the baseline and follow-up images are fed to the segmentation block that performs the final detection and segmentation of the new T2-w lesions. The network is trained end-to-end using a combined loss function.



(a) 3D registration network



(b) 3D segmentation network

Figure 4.2: The 3D registration and segmentation architectures. Each input modality has its 3D registration block (a) that learns the deformation field (DF) and nonlinearly registers the baseline image to the follow-up image. The registration block is a U-Net architecture with four downsample and upsample steps. The spatial transform block is used to warp the baseline image to the follow-up space using the learned DF enabling end-to-end training. The four learned DFs and the baseline and follow-up images are then fed to the segmentation block (b), another U-Net that performs the final detection and segmentation of the new T2-w lesions.

is a supervised loss function $L_{CrossEntropy}$ (CrossEntropy) that controls the segmentation part of the network and penalizes differences between the segmentation and GT. The loss function L_{Total} is as follows:

$$L_{Total} = \underbrace{L_{CrossEntropy}(Seg, GT)}_{Segmentation\ loss\ function} + \underbrace{\sum_{m \in Modalities} \left(\overbrace{\frac{1}{N}\sum_{i=1}^{N}(F_{m_i} - B_m(DF_m)_i)^2}^{Similarity\ part} + \overbrace{\lambda \sum_{p \in DF} \parallel \bigtriangledown DF_m(p) \parallel^2}^{Regularization\ part} \right)}_{Registration\ loss\ function} \tag{4.1}$$

where F_m, $B_m(DF_m)$, and DF_m are follow-up image, baseline image warped by DF (moved baseline), and DF for a modality m, respectively. Seg and GT are the automatic segmentation and the ground truth, respectively. N is the number of voxels in a patch and λ is a regularization parameter.

4.3 Experimental setup

4.3.1 Datasets

Data and preprocessing: The database used in this chapter is the same in-house dataset (VH dataset) used in the evaluation of the method proposed in chapter 3. The database consists of images from 60 different patients with a CIS or early relapsing MS who underwent brain MRI in the Vall d'Hebron Hospital's center for monitoring disease evolution and treatment response. Thirty-six of the patients (13 women and 23 men; 35.4 ± 7.1 years of age) confirmed MS with new T2-w lesions, while 24 patients did not present new T2-w lesions. The dataset was preprocessed the same way as described before. To warp the baseline images to the follow-up space, the baseline PD-w image was linearly registered to the follow-up PD-w image using using Nifty Reg tools[1] [220, 221]. To avoid interpolation more than one, baseline T1-w and FLAIR were warped using the combined affine transformation. See section 3.3.1 for more details.

4.3.2 Training and implementation details

For training the network, 3D 32x32x32 patches with a step size of 16x16x16 were extracted from the baseline and follow-up images of the four input modalities. Zero padding was applied to all the input volumes. This configuration was chosen empirically to give the highest performance of the proposed model. When trying smaller and bigger patch sizes, the performance was not significantly better. Moreover, increasing the patch size was more computationally and memory expensive. Note also

[1] https://sourceforge.net/projects/niftyreg/

that we aimed to learn the registration part from all image locations and not only for those containing new lesions. Therefore, the whole model was trained end-to-end, including the registration and the segmentation part, using a uniform sampling of patches to cover all the image. The extracted patches were divided into training and validation sets (70% for training and 30% for validation). The training set was used to adjust the weights of the neural network, while the validation set was used to measure how well the trained model performed after each epoch. The model was trained using Adam [250] with default parameters and regularization parameter $\lambda = 0.01$ [247]. The extracted patches were passed to the network for training in minibatches of size 4, and the network was set to train for 30 epochs. To prevent overfitting, the training process was automatically terminated when the validation accuracy did not increase after 5 epochs.

The proposed method was implemented in Python[2], using Keras with the TensorFlow backend [251]. All experiments were run on a GNU/Linux machine running Ubuntu 18.04 with 128 GB RAM. The training was carried out on a single TITAN-X GPU (NVIDIA Corp, United States) with 12 GB RAM memory.

4.3.3 Evaluation

We evaluated the proposed framework in different scenarios. First, we analyzed the accuracy of the detection using a leave-one-out cross-validation strategy with the 36 patients with new MS lesions. We chose the leave-one-out cross-validation strategy to be able to perform a quantitative comparison with the results published in [34]. In this evaluation strategy, the proposed network was trained using 35 patients and tested with the remaining patient. This process was repeated until all patient images were used as test images. Moreover, to demonstrate the contribution of simultaneously learning both the DF and the segmentation of new T2-w lesions, the following models were analyzed:

- **SimLearnedDFs**: This is our main model in which the four registration blocks and the segmentation block were trained simultaneously end-to-end using the loss function explained in section 4.2.2. The four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up combined with the learned DFs were fed to the segmentation block as first and second inputs, respectively.

- **SepLearnedDFs**: In this model, the registration blocks and the segmentation blocks were trained separately. The four registration blocks were trained first to obtain the DFs. Then, the four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up combined with the learned DFs were fed to the segmentation block as first and second inputs, respectively. This model was used for comparison with the **SimLearnedDFs** model to highlight

[2]https://www.python.org

the impact of the end-to-end simultaneous training of the DFs and new T2-w lesions.

- **DemonsDFs** (The proposed network using the DFs obtained from Demons [217]): This model did not use the registration blocks of the proposed network shown in Figure 4.1. It used only the segmentation block with four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up as the first input. The second input of the segmentation block was the DFs directly computed by the registering baseline to the follow-up image for every input modality using a multiresolution Demons registration approach from ITK [217]. This model was used for comparison with the **SimLearnedDFs** model to highlight the impact of learned-based DFs with the end-to-end training over the DFs from Demons.

- **NDFs** (The proposed network without DFs): This model did not use the registration blocks of the proposed network shown in Figure 4.1. It used only the segmentation block with only the four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both the baseline and follow-up as input. This model was used for comparison with the other three models to highlight the impact of the addition of the DFs in increasing the detection of new T2-w lesions.

Second, we analyzed the specificity of the method with 24 patients with no new T2-w lesions. Testing the performance on these cases allowed to further study how robust was the proposed method to avoid detecting false positives in patients with inactive disease. To do this, we performed a new training using all the 36 images with new MS lesions. Furthermore, we compared the obtained results with those of recent state-of-the-art approaches [33, 34, 149, 156] [3] applied to the same dataset used in this work (VH dataset). For the work of Schmidt et al. [149], we used their implementation of the longitudinal pipeline[4]. The lesion growth algorithm [85] was used to obtain the initial cross-sectional WML segmentation per time point. The parameter κ was empirically optimized for the current dataset, selecting the value $\kappa = 0.15$. The statistical significance of the performance between proposed model (SimLearnedDFs) and the state-of-the-art approaches [33, 34, 149, 156] was computed by running a series of permutation tests between the DSC (Segmentation) and DSC (Detection) obtained by each method as explained in section 3.3.4. Additionally, the Pearson's correlation coefficient was also used to analyze the linear relationship between manual annotations and the automatic detections obtained with the proposed model (SimLearnedDFs) in terms of number of new lesions and new lesion volume.

Similarly to the evaluation of the LR-based model proposed in the chapter 3, we studied the performance of the model according to the different lesion sizes. We analyzed the same categories, where lesions of $[3-10]$ voxels were considered small, lesions of $[11-50]$ voxels were considered medium, and lesions of $+50$ voxels

[3]Note that Salem et al. [34] is our LR-based model proposed in chapter 3.

[4]https://www.statistical-modeling.de/lst.html

were considered large. This division was useful to analyze the effect of the performance of the proposed model (SimLearnedDFs), the three variants (SepLearnedDFs, DemonsDFs, NDFs), and the state-of-the-art approaches [33, 34, 149, 156] on different lesion sizes.

Moreover, we studied the performance of the proposed model (SimLearnedDFs), the three variants (SepLearnedDFs, DemonsDFs, NDFs), and the state-of-the-art approaches [33, 34, 149, 156] on different brain regions. To the best of our knowledge, there is no current study that states that the location of the lesion is important for the longitudinal assessment of MS, although the lesion location (periventricular, juxtacortical, infratentorial, and deep white matter) is used to prove dissemination in space according to the McDonald criteria [19]. The motivation to perform this study was mainly to analyze the behavior performance of all the approaches on these specific regions. In particular, the analysis of the new MS lesion detection was divided into 4 types (periventricular, juxtacortical, infratentorial, and deep white matter) according to its location in the brain. An atlas with three segmented regions (cortex, ventricles, and (cerebellum and brainstem)) was resampled for each patient space after nonlinearly registering the atlas template to the T1-w image of each patient. After the registration, a new MS lesion was considered periventricular, juxtacortical, or infratentorial if it touched the cortex, ventricles, or cerebellum and brainstem, respectively. Otherwise, it was considered a deep white matter lesion.

Standard measures such as the true positive fraction (TPF), the false positive fraction (FPF), and the Dice similarity coefficient (DSC), which was computed lesion-wise and voxel-wise, were used for the quantitative analysis. In terms of detection, a lesion was considered a true positive if there was at least one overlapping voxel [32, 33, 34]. In terms of segmentation, only the voxel-wise DSC was computed. For all the evaluated pipelines, the automatic segmentation masks were obtained by thresholding the probability maps at 0.5 (using argmax). This thresholding value was not optimized. Since the outputs of the network were two probability maps (for the lesions and background), we used the argmax function which chooses the class with the highest probability. Since we are dealing with a binary problem, the highest probability is always 0.5 or greater. Therefore, using argmax is equivalent to using a threshold of 0.5. All automatic lesions with a size lower than three voxels were removed as done in previous works ([27, 34, 252]). A paired t-test at the 5% level was used to evaluate the significance of the results of the proposed method.

4.4 Results

Table 4.1 summarizes the new T2-w lesion detection and segmentation mean results for our proposed model (SimLearnedDFs) and the three variants (SepLearnedDFs, DemonsDFs, NDFs). We also included four state-of-the-art approaches for comparison [33, 34, 149, 156] when analyzing the 36 MS patients. Notice that our SimLearnedDFs model outperformed the three variants (SepLearnedDFs, DemonsDFs, NDFs) models and had the best values for all the evaluation measures. Addition-

ally, it outperformed all the state-of-the-art approaches in terms of all the evaluation measures (except $DSCs$, it has the same as Salem et al. [34]). Regarding the mean runtime per patient, the SimLearnedDFs model could process a testing case in less than 9 minutes while the other state-of-the-art methods [33, 34], that are based on DF obtained using classic nonrigid registration approach, took on average 18.36 and 18.55 minutes, respectively. Figures 4.3(a) and 4.3(b) visually show the result of the permutation tests for the segmentation and the detection DSC values, respectively. As mentioned in chapter 3, permutation tests permit to compute the exact P-value, and are not limited by any statistical distribution or minimum number of subjects. Essentially, each method is compared against all others using randomly selected subsets of data using statistical difference-of-mean test that do not require data to follow the normality condition. Notice that the data variability is still present in the fact that mean values obtained by all methods are not too high (best methods obtain $\mu_{Detection} = 0.60$ and $\mu_{Segmentation} = 0.40$). It is, however, possible to see how some methods do better than the other in pairwise comparisons that bear statistical significance. Regarding segmentation, the methods in rank 1 included only approaches that used DF-based features, whereas non-DF based approaches were placed in rank 2. Regarding detection, only the SimLearnedDFs and Salem et al. [34] models were in rank 1. Because ranking between the approaches differed, we can conclude that there is a significant difference in performance when including DFs in a supervised way.

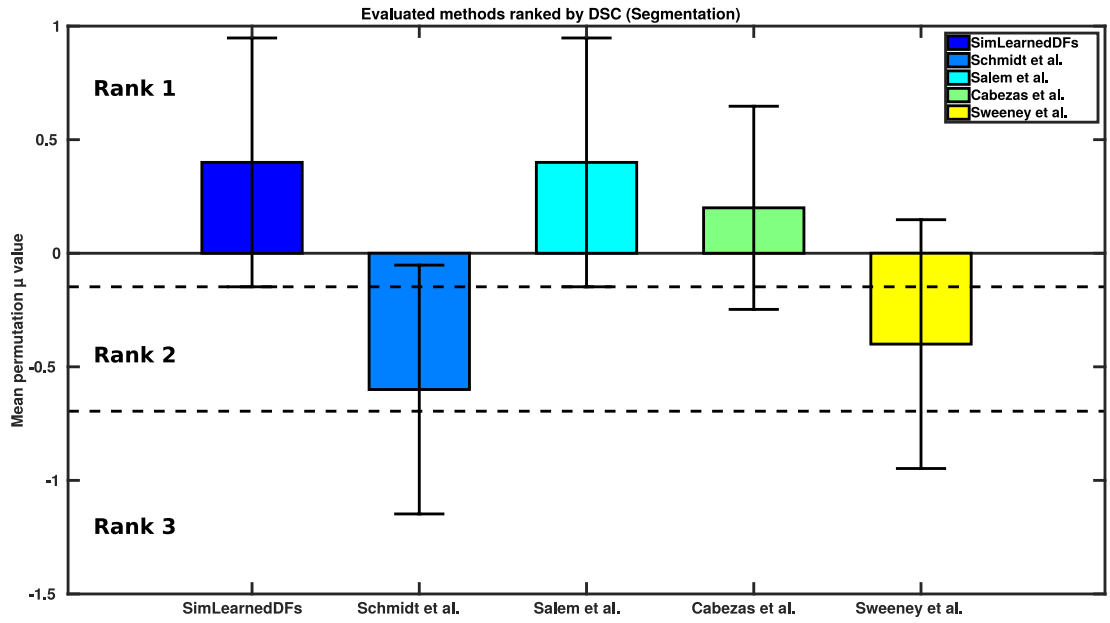
Figure 4.4 shows a visual example of the performance of our SimLearnedDFs model, where each column corresponds to the baseline T2-w image, follow-up T2-w image, GT annotated lesions, and the segmentation of SimLearnedDFs, NDFs, DemonsDFs, and SepLearnedDFs approaches. Figure 4.5 shows the relationship between baseline, follow-up, the learned DF, GT, and the segmentation of the SimLearnedDFs model in the four input modalities.

Regarding false negatives, the SimLearnedDFs model missed about 17% of the total number of lesions being distributed as 48% small lesions, 38% medium lesions, and 14% large lesions. Figure 4.6 shows two examples of false positive detections using the SimLearnedDFs model. Some of the false positives were due to inflammation areas that were not marked as new lesions by the experts and the remaining were mainly due to image artifacts.

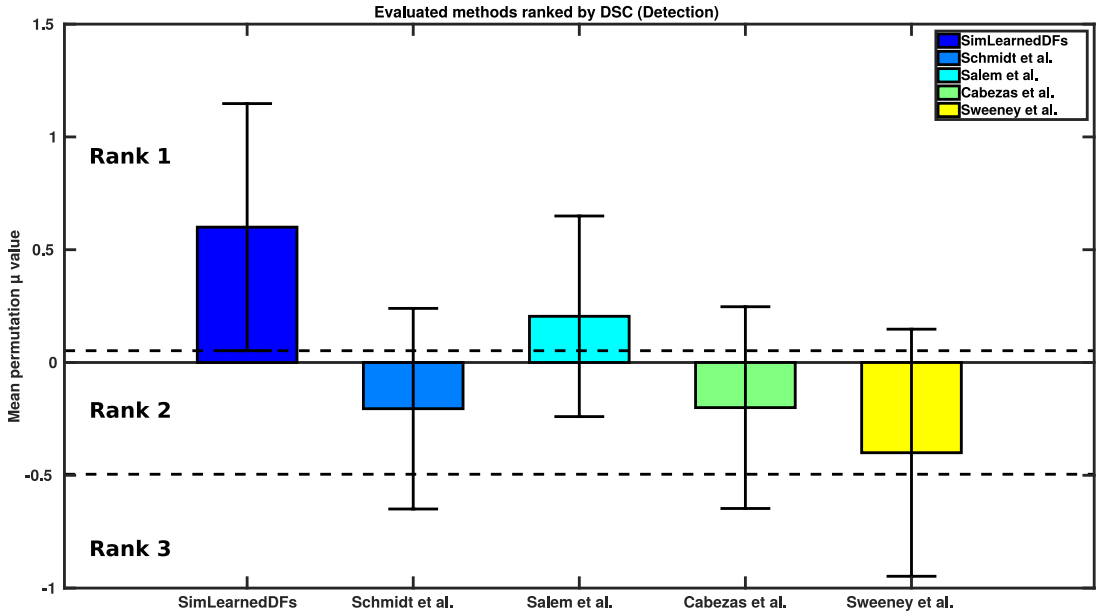
Analyzing the results per patient, Figure 4.7 shows a box plot summarizing the performance of the SimLearnedDFs, the three variants (SepLearnedDFs, DemonsDFs, NDFs), and the state-of-the-art methods on the four metrics used in the evaluation. With this analysis done per patient, we observe that the proposed model (SimLearnedDFs) also provided better sensitivity for the cases that present few new lesions (i.e. 1, 2 or 3). For instance, for the 8 cases containing only one new lesion, our approach obtained a (TPF, FPF)=(100%, 0%), while Sweeney et al. [156], Cabezas et al. [33], Salem et al. [34], and Schmidt et al. [149] models obtained (71.43%, 25%), (85.71%, 0%), (85.71%,14.29%), and (71.43%, 38.33%), respectively.

Table 4.1: Lesion Detection Results: Comparison between the different models evaluated. The results represent the mean detection TPF, FPF, $DSCd$, mean segmentation $DSCs$, and the mean runtime in minutes when analyzing the 36 MS patients using a leave-one-out cross-validation scheme. The automatic segmentation masks were obtained by thresholding the probability maps at 0.5 (using argmax), and all automatic lesions with a size lower than three voxels were removed.

Method	TPF	FPF	$DSCd$	$DSCs$	Runtime (in minutes)
SimLearnedDFs	$\mathbf{83.09 \pm 21.06}$	$\mathbf{9.36 \pm 16.97}$	$\mathbf{0.83 \pm 0.16}$	$\mathbf{0.55 \pm 0.18}$	8.70 ± 0.09
SepLearnedDFs	57.77 ± 34.34	13.67 ± 21.99	0.60 ± 0.31	0.39 ± 0.22	9.08 ± 0.06
DemonsDFs	62.06 ± 32.74	11.98 ± 23.09	0.67 ± 0.29	0.42 ± 0.24	18.10 ± 0.05
NDFs	53.99 ± 38.01	17.20 ± 26.96	0.55 ± 0.35	0.37 ± 0.28	7.58 ± 0.09
Sweeney et al. [156]	59.82 ± 37.59	33.59 ± 33.52	0.57 ± 0.33	0.44 ± 0.26	8.36 ± 0.01
Cabezas et al. [33]	70.93 ± 34.48	17.80 ± 27.96	0.68 ± 0.33	0.52 ± 0.24	18.36 ± 0.02
Salem et al. [34]	80.0 ± 27.77	21.87 ± 26.26	0.76 ± 0.25	0.55 ± 0.22	18.55 ± 0.02
Schmidt et al. [149]	68.66 ± 35.26	31.89 ± 36.10	0.62 ± 0.34	0.40 ± 0.25	7.58 ± 0.03



(a)



(b)

Figure 4.3: Permutation test results for the evaluated methods. Final ranks based on (a) the DSC (Segmentation) and (b) the DSC (Detection).

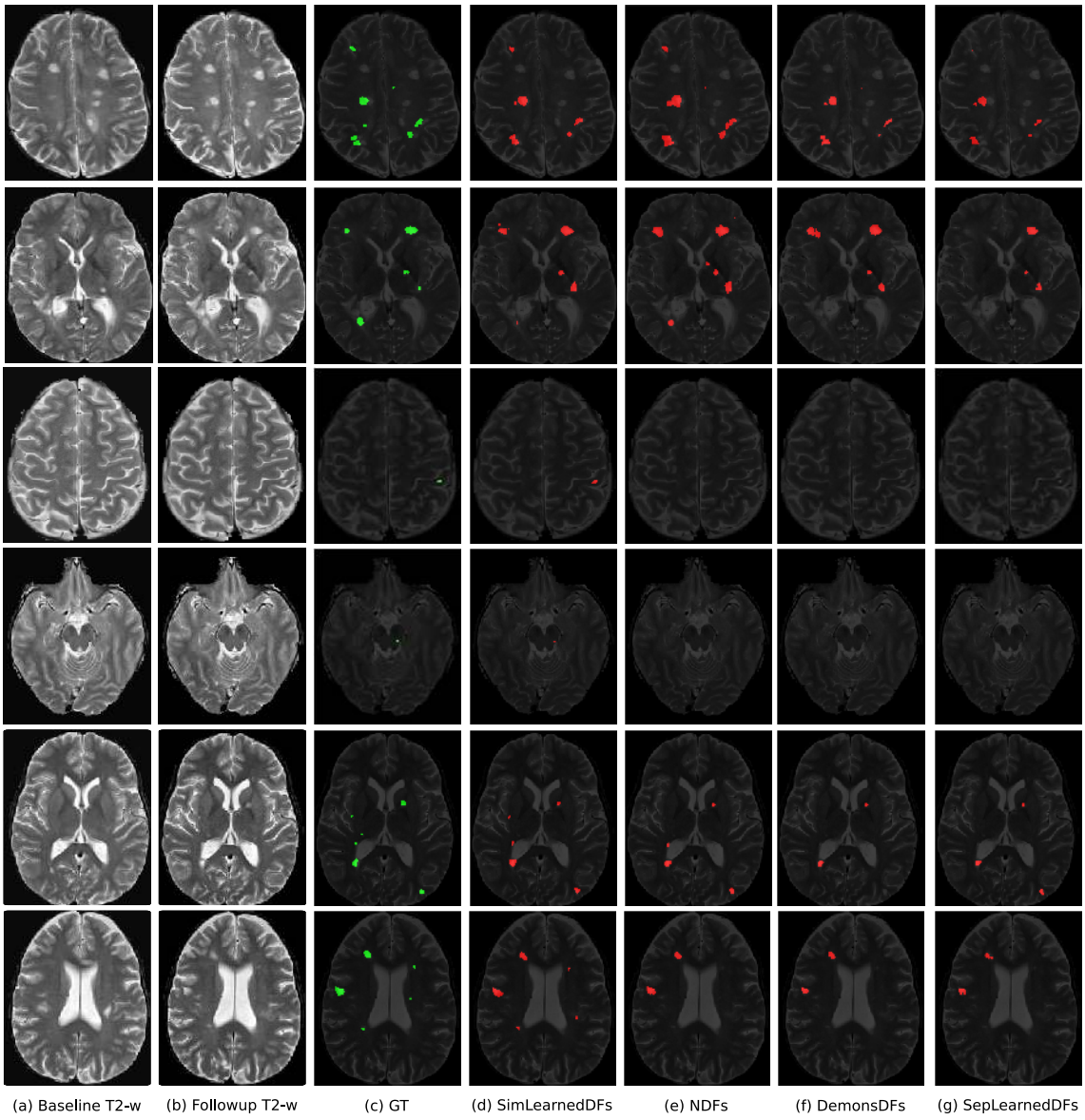


Figure 4.4: Examples of new MS lesion detection in a 12-month longitudinal analysis. (a) and (b) show one axial slice of the T2-w image at baseline and follow-up, respectively. (c) shows the new MS lesions annotations performed by an expert (GT). (d), (e), (f), and (g) show the segmentation of SimLearnedDFs, NDFs, DemonsDFs, and SepLearnedDFs approaches, respectively. The GT and the segmentations are overlaid in green and red, respectively, on the follow-up T2-w image.

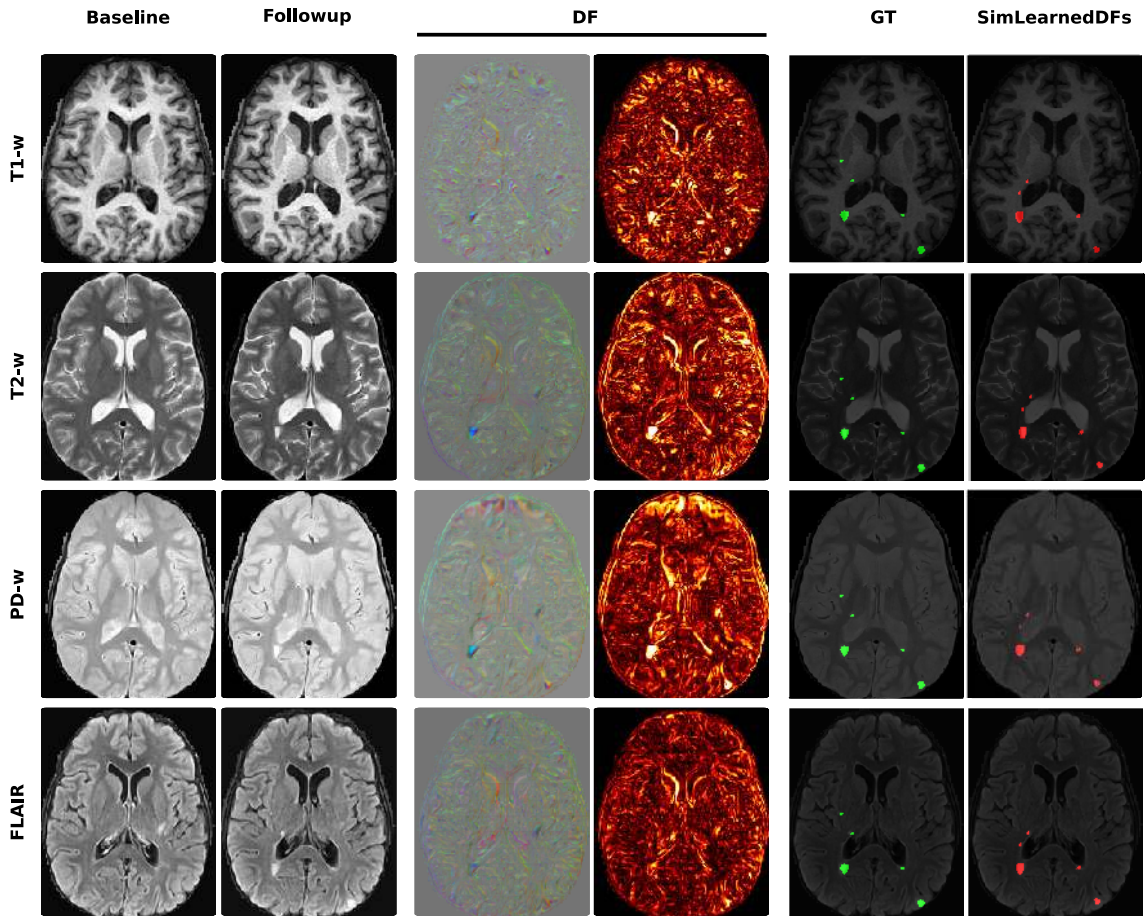


Figure 4.5: Relationship between baseline, follow-up, the learned DFs, GT, and the segmentation of SimLearnedDFs in the four input modalities. All images are from the same patient and the same slice. The DFs are displayed in RGB (third column) and their magnitudes (fourth column) using a hot color map. The GT and the segmentation of SimLearnedDFs are overlaid in green and red, respectively, on the follow-up image.

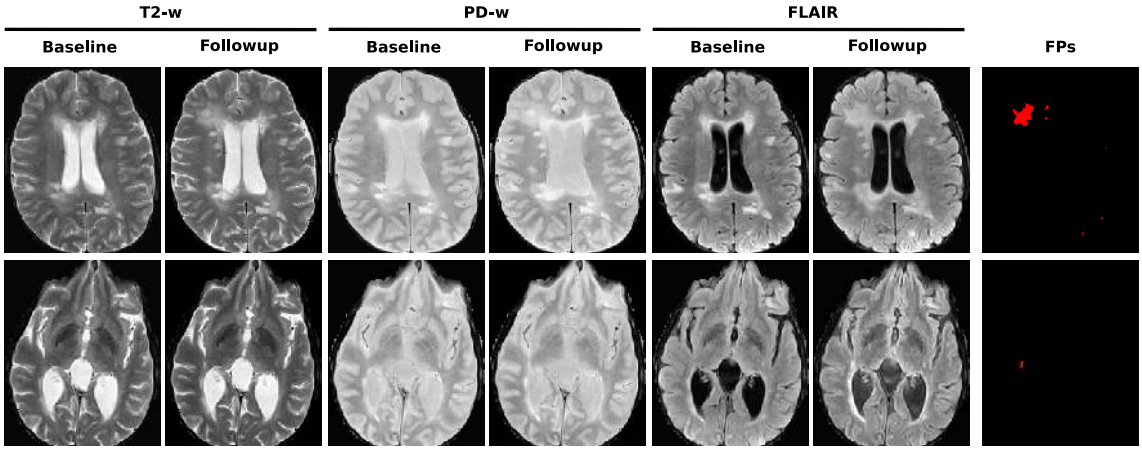


Figure 4.6: False positive detection example. Some false positives (the first row) were due to inflammation areas that were not marked as new lesions by the experts and the others were mainly due to artifacts.

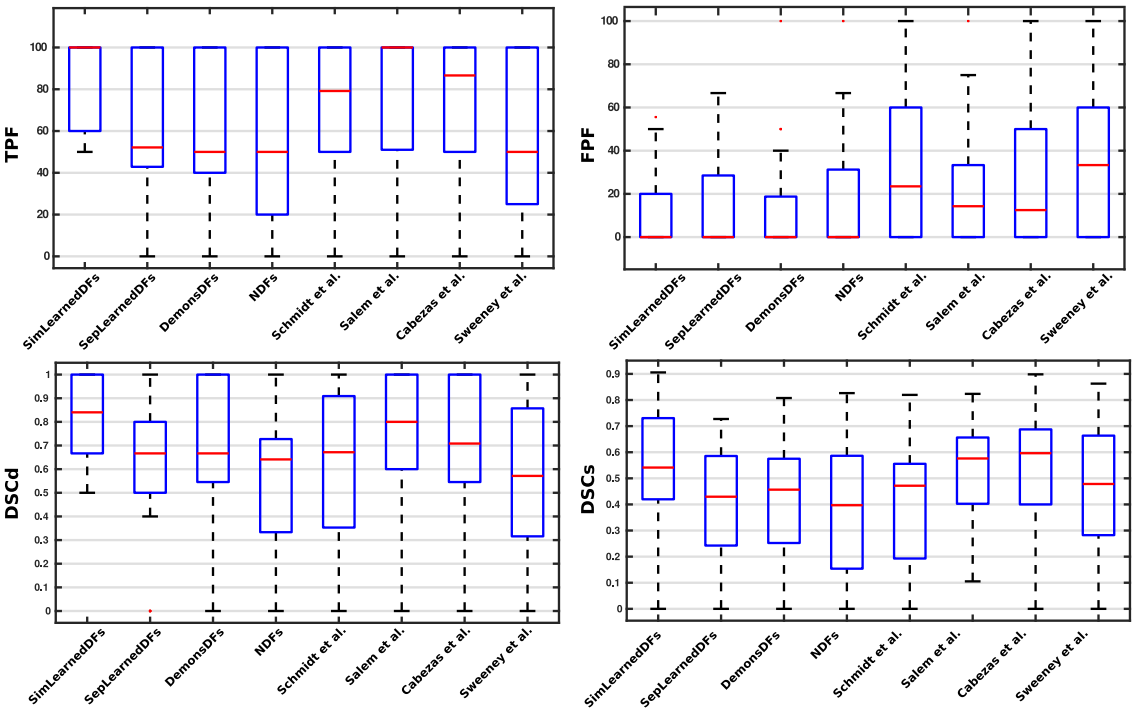


Figure 4.7: Box plot summarizing the performance of the SimLearnedDFs, the three variants (SepLearnedDFs, DemonsDFs, NDFs), and the state-of-the-art methods on the four metrics used in the evaluation.

Table 4.2: Analysis of TPF for different classifiers for different lesion sizes. Lesions between 3 and 10 voxels are considered small; lesions between 11 and 50 voxels, medium; and lesions with 50 voxels, large

Method	$3-10$	$11-50$	$+50$
SimLearnedDFs	39.52	**83.32**	**97.14**
SepLearnedDFs	22.62	49.40	83.14
DemonsDFs	30.0	78.25	90.26
NDFs	14.29	47.71	80.48
Sweeney et al. [156]	16.67	52.06	78.25
Cabezas et al. [33]	**42.86**	48.57	77.42
Salem et al. [34]	34.40	65.70	91.30
Schmidt et al. [149]	13.1	71.92	94.08

Figure 4.8.a shows the correlation between the number of new lesions manually annotated and the automatically detected (Significant Pearson's correlation: $R = 0.97$; $p_{value} = 2.7445e^{-21}$; confidence band = 95%), while Figure 4.8.b shows the correlation between lesion volume in the GT and the automatically segmented (Significant Pearson's correlation: $R = 0.98$; $p_{value} = 5.0233e^{-24}$; confidence band = 95%). Regarding the number of the data points used, all the MS patients with lesion progression were used for this correlation (36 data points - 36 patients). However, several patients had the same number of GT and automatically detected lesions and therefore some points are overlapping in the plot. Notice that there are numerous cases in which the number of new lesions per patient is actually very small.

Table 4.2 summarizes the performance of our pipeline according to the different lesion sizes described in Section 3.3.2. The SimLearnedDFs model had a better performance than the other three variants (SepLearnedDFs, DemonsDFs, and NDFs) in all lesion size categories, although the results with small lesions had a worse performance when compared with larger lesions. Moreover, SimLearnedDFs had also a better performance than the state-of-the-art approaches [33, 34, 149, 156] for medium and large lesion size categories.

Figure 4.9 shows the performance of the new T2-w lesion detection when analyzed according to its location in the brain. Note that here the TPF and FPF were computed per lesion type and not per patient. The proposed model (SimLearnedDFs) appeared to learn well from most of the brain regions, and it had the highest sensitivity everywhere. The dataset had a total of 191 lesions (periventricular = 25, juxtacortical = 34, infratentorial = 12, and deep white matter = 120). Moreover, we evaluated the behavior of the SimLearnedDFs model trained with all 36 patients when tested with the set of 24 patients with no new T2-w lesions. The obtained results showed only 2 cases with one FP detection in each, and these results were better than those obtained with the other approaches.

To analyze the generalization and the performance of the proposed approach when tested in images from a different scanner and image acquisition protocol, we performed a new experiment with data from another collaborating Hospital (Dr.

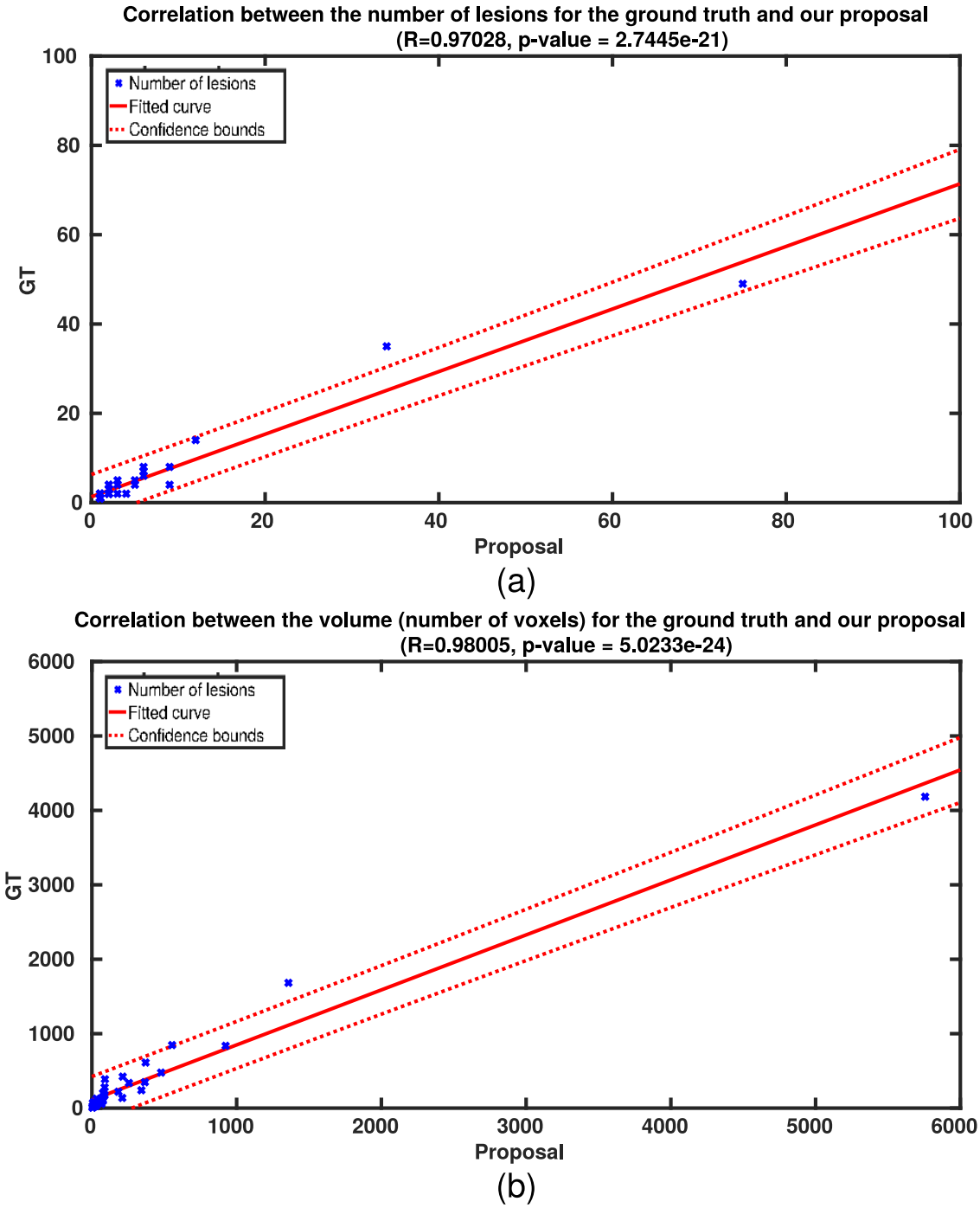


Figure 4.8: Correlation between (a) the number of GT lesions and the number of automatically detected ones using the proposed SimLearnedDFs model (Pearson's coefficient $R = 0.97; p_{value} = 2.7445e^{-21}$) and (b) the volume (the number of voxels) of GT lesions and the volume of automatically detected ones using the proposed SimLearnedDFs model (Pearson's coefficient $R = 0.98; p_{value} = 5.0233e^{-24}$). All the MS patients with lesion progression were used for this correlation (36 data points - 36 patients). Notice that different patients have the same combination of number of GT lesions and the SimLearnedDFs model detections. Therefore, several points are overlapping in the plot.

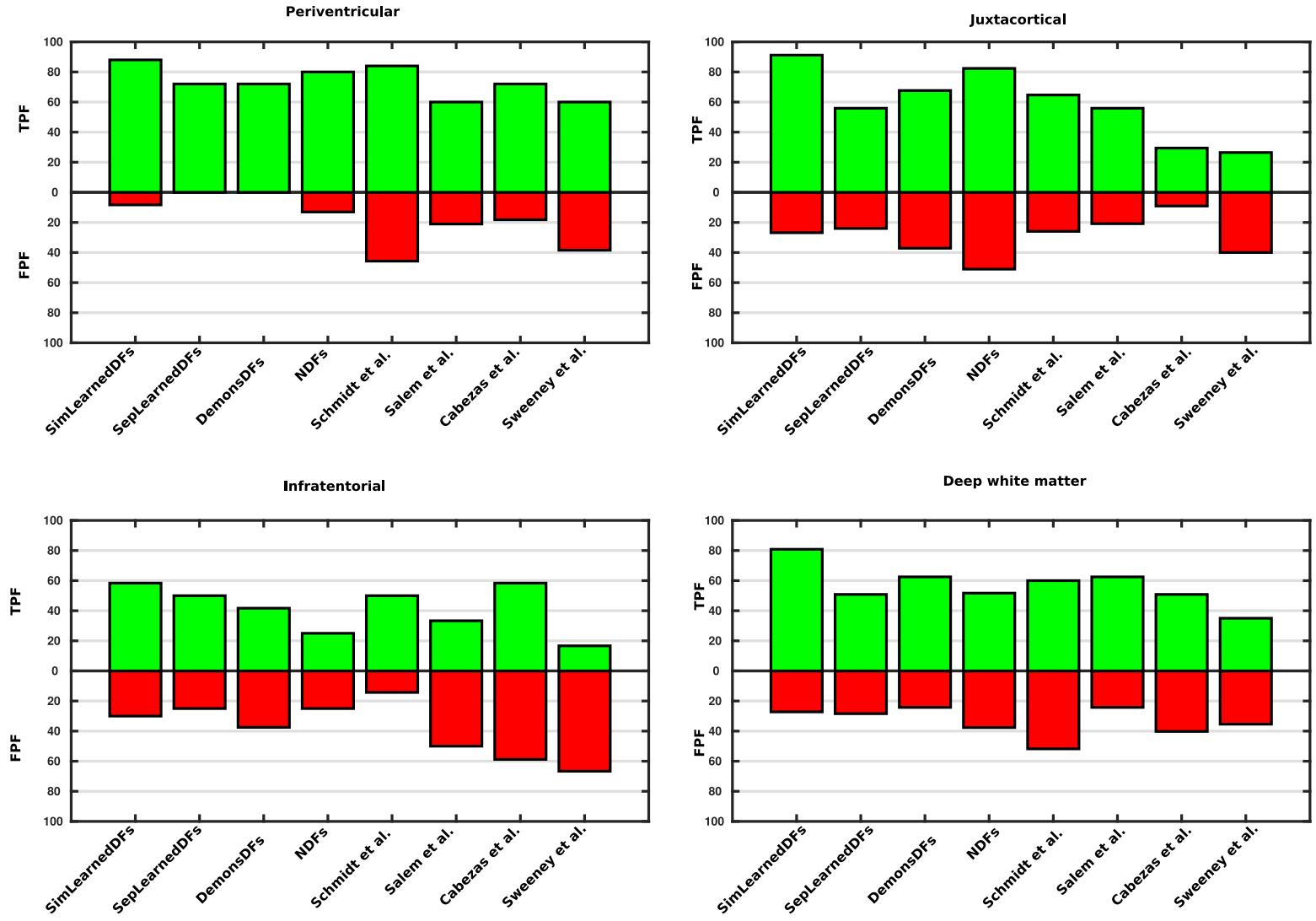


Figure 4.9: Results of the new T2-w lesion detection for the 4 brain regions. The dataset had a total of 191 lesions (periventricular = 25, juxtacortical = 34, infratentorial = 12, and deep white matter = 120). TPF and FPF were computed per lesion type and not per patient.

Josep Trueta Hospital, so we refer to this dataset as Trueta dataset). This dataset consisted of 17 MS patients, 9 of them with new T2-w lesions and 8 with no new T2-w lesions. The baseline and follow-up scans for all patients were obtained in a 1.5T magnet Philips scanner. The MRI protocol included the following sequences: 1) transverse proton density (PD)- and T2-weighted fast spin-echo (voxel size = $1.0 \times 1.0 \times 3.0$ mm^3), 2) transverse fast FLAIR (voxel size = $1.0 \times 1.0 \times 3.0$ mm^3), and 3) sagittal T1- weighted 3D magnetization-prepared rapid acquisition of gradient echo (voxel size = $1.0 \times 1.0 \times 1.0$ mm^3). The dataset was preprocessed in the same way as the VH dataset mentioned in section 3.3.1. The experiment consisted in applying the SimLearnedDFs model and the approach of Salem et al. [34] trained with the 36 cases from the VH dataset and testing them on the unseen Trueta dataset. The obtained results for the 9 cases with new lesions showed that the SimLearnedDFs obtained a TPF of 72.1% and a FPF of 34.97%, while Salem et al. [34] obtained a TPF of 54.81% and a FPF of 62.34%, respectively. Regarding the cases with no new lesions, the SimLearnedDFs model did not find any FP, while Salem et al. [34] obtained at least 1 FP in each case of the 8 cases.

4.5 Discussion

The proposed method is an FCNN for detecting new T2-w lesions in longitudinal brain MRI. The model is trained end-to-end and simultaneously learns both the DFs and the new T2-w lesions. As the DFs are learned inside the network and not computed separately using classic nonrigid registration methods, the execution time of the network on a testing image is reduced compared to the time of the state-of-the-art methods [33, 34]. Moreover, the proposed model is fully automated, simple, and does not require hand-crafting feature vectors to extract appearance information similar to [34] because CNNs learn a set of features that are specifically optimized for the task directly from the image data. The inputs to our model are only the four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up.

To analyze the effect of the end-to-end training, we trained the proposed model (SimLearnedDFs) and the other three variants (SepLearnedDFs, DemonsDFs, and NDFs). As mentioned in Table 4.1, the SimLearnedDFs model detected new lesions with a TPF of 83.09% and an FPF of 9.36%. In terms of TPF, the SimLearnedDFs model was significantly better than all the other methods except Salem et al. [34] method ($p < 0.05$). However, the TPF improved by 3%. In terms of FPF, the SimLearnedDFs model was not significantly better than the SepLearnedDFs (4.31% improvement) and the DemonsDFs (2.62% improvement), but it was significantly better than the other methods ($p < 0.05$). Note that the model trained without any DFs (NDFs) detected new lesions with a TPF of 53.99% and an FPF of 17.20%. This result shows, as previously discussed in [33, 34], that the addition of DFs helps to increase the detection of new T2-w lesions while maintaining a low number of false positives. However, the results also show that training the model end-to-end, simultan-

eously learning both the DFs and the new T2-w lesions (SimLearnedDFs), performs better than learning the DFs separately (SepLearnedDFs) or using DFs computed by classic deformable registration methods such as Demons [217] (DemonsDFs). The increase in performance using simultaneously learning compared to the variants that compute the DF separately could be explained by the use of the combined loss function during the training process. The simultaneously learning model trained the two connected networks (registration and segmentation) end-to-end. That means, in each training epoch, the weights of the registration networks which compute the DFs were updated during the backpropagation to minimize the summation of the cross entropy function (segmentation part) and the similarity function (registration part). These DFs (computed using the updated weights) were then used as inputs together with the intensity images to the segmentation network in the forward pass to compute the new lesion segmentation. So, the DFs were computed in a guided way that improved the new lesion segmentation. Note that the other variants did not include the connection between the registration and segmentation part, so DFs were computed blindly and independently from the segmentation. Moreover, our proposed model (SimLearnedDFs) improved the results of other unsupervised methods due to the use of a supervised classification model instead of an unsupervised rule-based approach [33, 149]. Compared with the state-of-the-art approaches, the proposed model (SimLearnedDFs) had better results than all the state-of-the-art approaches in terms of all the evaluation measures. It also operated orders of magnitude faster than [33, 34] during testing time due to the use of learning-based nonrigid registration. Regarding the analysis of the results when applied to the 24 patients with no new lesions, the proposed model (SimLearnedDFs) had high specificity, with no lesions found in 22 cases (only 2 patients had 1 FP).

Regarding the evaluation according to the lesion location, there was a relevant increase in the performance in the juxtacortical lesions when both the deformation fields and brain lesions were learned jointly. The SimLearnedDFs model had a TPF of 91.18% (31 lesions out of 34) and FPF of 26.19% (11 FPs out of 42 candidates) with DCSd of 0.82 and DCSs of 0.65. The NDFs model also had a high TPF of 82.35% but with a high FPF of 51.02% (DSCd = 0.64 and DSCs = 0.57). In the periventricular region, the lesions were easily observed, which may be explained by the good contrast between ventricular and the new MS lesions. The difference in TPF of all CNN-based methods was not as high. The proposed method (SimLearnedDFs) showed the highest sensitivity while still maintaining some false positives (2 FPs out of 24 candidates, 8.33%) compared to the SepLearnedDFs and DemonsDFsmodels that had no FPs in the periventricular region. Regarding the deep white matter lesions, the SimLearnedDFs model detected the highest number of lesions (97 out of 120, 80.83%), which may be explained by the high number of lesions in this particular region (63% of the total number of lesions). The difference between the three variants in terms of FPs was very low. In contrast, the sensitivity of CNN methods was remarkably lower in the infratentorial region due to a lack of training data (infratentorial lesions were only 6% of the total number of lesions). Furthermore, this may also be one reason for the worse performances of both methods where DF were

learned. In these methods, the learned deformation fields did not efficiently distinguish the complexity of the cerebellum, increasing the number of noninfratentorial lesion activations. The SimLearnedDFs model had only three FPs detected, and these FPs were only detected in one patient. All of the subtraction-based methods like [33, 34, 156] had higher FP lesions in this region, which may be explained by the poor contrast between tissues in the cerebellum region and therefore, a noisy subtraction. We believe that more training data or the use of synthetic MS data like in [30] with more infratentorial lesions, may increase the sensitivity of all CNN-based methods while reducing FP lesions. Schmidt et al. [149] method had high TPF in the periventricular, juxtacortical, and deep white matter regions but also a high FPF. It had (DSCd, DSCs) of (0.68, 0.49), (0.7, 0.51), and (0.54, 0.34) for the periventricular, juxtacortical, and deep white matter regions, respectively. We also observed that in the infratentorial region, it had better performance than the SepLearnedDFs and DemonsDFs models. However, these results should be further analysed with more cases containing periventricular and infratentorial lesions to have a more robust analysis.

We also studied the use of conventional data augmentation methods like geometric transformations such as image translation, rotation, or flip. However, the performance did not increase. One reason might be due to the fact that the generated samples did not represent image appearances in real data, or the generated samples were very similar to the existing images in the training dataset. Working on the development of a framework for generating new longitudinal synthetic MS lesions on patients or healthy MR images, could allow the creation of more data samples for particular lesion locations where few samples are available (i.e, the infratentorial region), helping to improve the trained models.

Regarding the experiment in which the proposed model (SimLearnedDFs) was applied to images from a different hospital, as expected, the TPF and FPF detection values were worse due to the change of domain (change in scanner and MRI protocol). Note however, that the SimLearnedDFs model provided a better generalization than the one not based on deep learning (Salem et al. [34]). Moreover, the obtained results with the SimLearnedDFs model in the Trueta dataset were also better than those of the unsupervised approaches (Cabezas et al. [33], Schmidt et al. [149]), using the parameter configuration optimized for the VH Hospital. The performance without parameter tuning was actually poor, while the optimum configuration provided similar results than those shown on the VH dataset.

In conclusion, the obtained results indicate that the proposed end-to-end training model increases the accuracy of the new T2-w lesion detection. The results also indicate that the DL based model is better than the LR-based model described in chapter 3. One of the drawbacks of DL techniques is the lack of the training data. We believe that having more training data would improve the DL techniques proposed for the MS lesion segmentation and detection in both cross-sectional and longitudinal analysis. In the next chapter, we propose a deep FCNN model for MS lesion synthesis and explain how the synthetic MS lesions can be used as data augmentation for increasing the segmentation and detection accuracy of MS lesions.

CHAPTER 5

MULTIPLE SCLEROSIS LESION SYNTHESIS ON MAGNETIC RESONANCE IMAGING

5.1 Overview

As described in chapters 3 and 4, detecting cross-sectional or longitudinal MS lesions using supervised machine learning algorithms on MR images requires a large number of samples to be annotated by expert radiologists. However, obtaining the annotations of medical images is time consuming. Several attempts have been made to overcome this issue by using data augmentation. One of the most common data augmentation method is to modify the dataset of images using geometric transformations such as image translation, rotation, or flip [185]. However, the generated samples may not represent image appearances in real data, or the generated samples may be very similar to the existing images in the training dataset due to the parameters and image operators used [253]. In contrast, we will propose in this chapter the generation of synthetic MS lesions on patient or healthy MR images as the solution to the lack of expert annotations.

The synthesis of MR images has attracted much interest in several areas of neuroimaging, including how to replace the missing MR modalities with synthetic data [254], to generate a subject-specific pathology-free image that is not present in the input modality [255], to improve image segmentation and registration performance [256] and others. The current state of the art in brain MRI synthesis is the work of Chartsias et al. [257]. The authors proposed an FCNN model for MRI synthesis, which takes different modalities as inputs and outputs synthetic images of the brain in one or more new modalities. This approach could be used for the

synthesis of new lesions. However, there are some limitations that should be considered, such as the ability to control the intensity and the texture inside the lesions and the requirement of ground-truth masks for obtaining the lesion model.

In this chapter, we propose a novel FCNN model for MS lesion synthesis. The model takes as inputs images without MS lesions and outputs synthetic images with MS lesions. The lesion information is encoded as different binary masks passed to the model stacked with the input images. To overcome the limitations of the Chartsias et al. [257] model, we divide the lesions into different regions based on voxel intensities, encoding this information as different binary masks. These binary masks are computed directly by thresholding the hyperintensities in the FLAIR image, so there is no need for the lesions' ground truth. That means the proposed MS lesion synthesis model is trained end-to-end without the need of manual expert MS lesion annotations in the training sets. Therefore, to tackle the lack of available ground-truth data needed for supervised MS lesion detection and segmentation strategies, we use the generated synthetic MS lesion images as data augmentation to improve the lesion detection and segmentation performance. This is done by synthesizing the lesions in new brain images, coming from either healthy subjects or from patients with lesions. We evaluate the proposed model on analyzing the improvement the cross-sectional and longitudinal MS lesion detection and segmentation approaches.

5.2 Methods

5.2.1 Synthetic MS lesion generation pipeline

To learn a model for the generation of synthetic MS lesions, images without lesions (used as inputs to the model) and the correspondent images with lesions (used as outputs to the model) are required. This kind of image set is not easy to obtain. One way to solve this would be using a longitudinal MS dataset; however, MS lesions in the baseline images and new MS lesions on the follow-up images should be annotated. Moreover, the baseline and follow-up images should also be registered. In that way, the model would be trained to generate new lesions in the follow-up scans. Nevertheless, in this scenario, new lesions on the follow-up images may not be sufficient to train the model since the volume of most of the new lesions can be relatively low [34]. Therefore, to overcome the lack of available ground-truth, we use the MS lesion generation pipeline shown in Figure 5.1 which consists of three main stages. First, the creation of an approximate white matter hyperintensity (WMH) mask and several intensity level masks to encode the intensity profile of the WMH voxels. Second, the filling of this WMH mask in the MR images with intensities resembling WM. Finally, the generation of MS lesions using the MS lesion generator network on the filled images. Notice that the proposed MS generator was trained using only a cross-sectional MS dataset. These filled images were considered as images without lesions (used as inputs to the model), while the original images contained MS lesions (used as outputs to the model during the training process).

The following subsections explain the full pipeline in more detail.

WMH mask and intensity level masks

Creating the WMH mask and the intensity level masks is an important step in the proposed MS lesion generator pipeline. The aim is that training the model with intensity level masks instead of MS lesion masks avoids the limitation of having ground-truth. First, the FLAIR image is thresholded to obtain an approximate WMH mask [24]. This mask is used to fill the WMH regions with intensities similar to the ones of the surrounding WM voxels. To learn the model for the generation of WMH voxels and their intensity profile, the range of intensities starting from the initial threshold is divided into different small ranges by increasing the intensity threshold at different steps. These created masks are considered as intensity level masks, which are then used to encode the intensity profile of the WMH voxels. The intensity level masks are stacked with the filled MR images when training the MS generator model. Therefore, the model can be trained with any dataset without requiring manual expert annotations. The approximate WMH mask is computed by FLAIR thresholding. The threshold $T^F_{\gamma_i}$ and intensity level mask IL_i are computed as follows:

$$T^F_\gamma = \mu^F_{GM} + \gamma\sigma^F_{GM} \tag{5.1}$$

$$IL_i = T^F_{\gamma_i} < FLAIR \leq T^F_{\gamma_{i+1}} \tag{5.2}$$

where μ^F_{GM} and σ^F_{GM} are the intensity's distribution parameters of gray matter (GM) tissue on the FLAIR image [24]. A small value of γ must be chosen to obtain an approximate WMH mask so that all the WMH voxels are included in this mask. Different intensity level masks are obtained by increasing the γ value. The higher the value of γ, the more brighter WMH voxels are included in the mask.

In this study, the approximate WMH mask was obtained with $\gamma = 0.5$. This value was found empirically to ensure that all the WMH voxels were included in the WMH mask. Eight intensity level masks with $\gamma = 0.5$, 0.8, 1.1, 1.4, 1.7, 2.1, 2.4, and 2.7 were used to encode the WMH intensity profile. This was a trade-off between the memory required and the minimum number of training samples inside each intensity level mask while training the model. Note that these masks are stacked with each input modality so the higher the number of masks, the higher the memory requirements. Moreover, increasing the number of masks produces a decrease in the number of training voxels per mask. Figure 5.2 describes the creation of the eight intensity level masks (IL_1, IL_2, ..., and IL_8). The $\gamma = 0.5$ WMH mask is used to fill the WMHs in the original image, and the intensity level masks are used to encode the intensity profile in the obtained WMH mask.

Figure 5.1: Scheme of the synthetic MS lesion generation pipeline. $\gamma = 0.5$ WMH mask and the eight intensity level masks were computed by FLAIR thresholding. The $\gamma = 0.5$ WMH mask was used to fill all the input modalities. Afterwards, the eight intensity level masks were stacked to each filled modality to create two 2D inputs with 9 channels each and these were the inputs to the MS lesions generator. For training, the original modalities were used as output. At testing time, if the intensity level masks were passed to the generator network without modification, the output images would be the generated version of the input ones containing all the WMHs found in the input image. Passing modified intensity level masks to the generator network will generate these modifications (i.e., new MS lesions) on the output images.

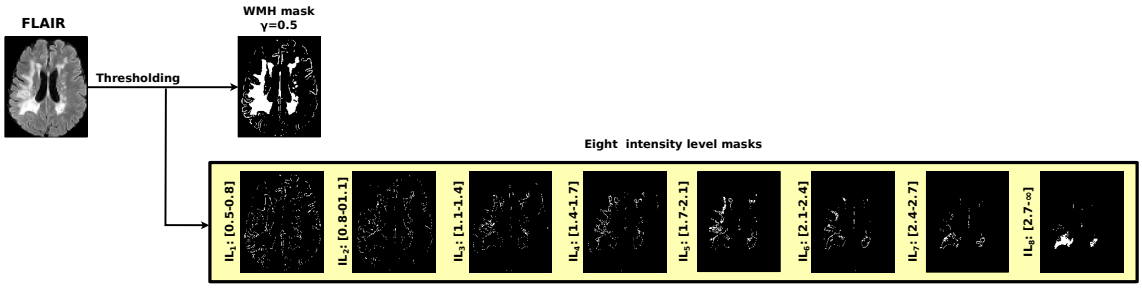


Figure 5.2: The creation of the WMH mask and the eight intensity level masks (IL_1, IL_2, ..., and IL_8) using FLAIR thresholding.

WMH filling

After creating the intensity level masks described in the previous section, the $\gamma = 0.5$ WMH mask regions are filled in the input modalities. As described in section 2.2.1, a local filling method is used here to fill the WMH area with the surrounding WM voxels in all input modalities. First, for each slice in the MR image, the WMHs are split into individual connected regions. Second, each connected region is dilated twice. Each connected region is filled using values normally sampled using the mean and standard deviation of the WM voxels that were laid in the first dilated area. Furthermore, the filled area with its surrounding voxels (voxels in the filled connected region and the two dilated areas) is smoothed using a local Gaussian filter. The second dilation determine the region on which the local Gaussian filter is used to merge the filled region with the surrounding WM areas.

MS lesion generation model

Figure 5.3 shows our MS lesion generator architecture, which is inspired by the work of Chartsias et al. [257]. As shown in Figure 5.3(a), the encoders are used to learn the latent representation for the input modalities, while the decoders are also used to generate the output modalities. Each decoder is used five times (i.e., shared decoder): one to decode each of the four individual latent representations and one to decode the fused latent representation. The fused latent representation is computed by combining the T1-w, T2-w, PD-w, and FLAIR latent representations using a voxel-wise max function (i.e., each voxel of the fused latent representation has exactly the maximum value of the four latent representations). At testing time, we used the synthesis result from the fused latent representation as our output. The model has four 2D input patches with nine channels each (one input patch for each input modality). The eight intensity level masks computed as explained in Section 5.2.1 are stacked with each of the filled input modalities. The first channel is the filled image modality and the other eight channels are the intensity level masks.

Encoder architecture: One independent encoder is built for each input modality following the architecture shown in Figure 5.3(b). The encoders embed input im-

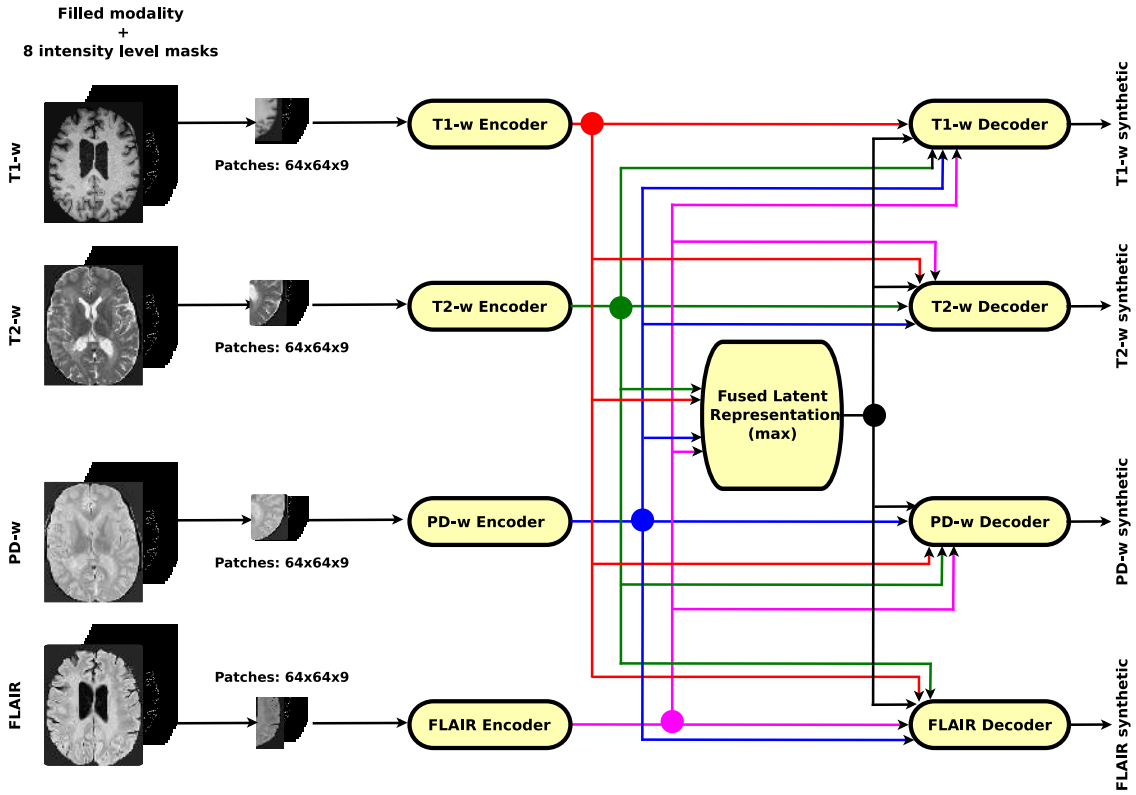
ages into a latent space of 32-channel size. This architecture is inspired by the work of Guerrero et al. [258]. It is a fully convolutional network that follows a U-shaped architecture [191]. The U-Net's downsampling followed by the upsampling and skip connections allow the network to exploit information at large spatial scales, while not losing useful local information. Moreover, as discussed in Drozdzal et al. [248], skip connections facilitate gradient flow during training. Our encoders are shallower than the original U-Net, having three downsample and upsample steps compared to the original four steps. This reduces the training and run times for the model.

Decoder architecture: One decoder is built for each output modality following the architecture shown in Figure 5.3(b). The model is a fully convolutional network to map a multi-channel image-sized latent representation to a single channel image of the required modality with synthetic MS lesions.

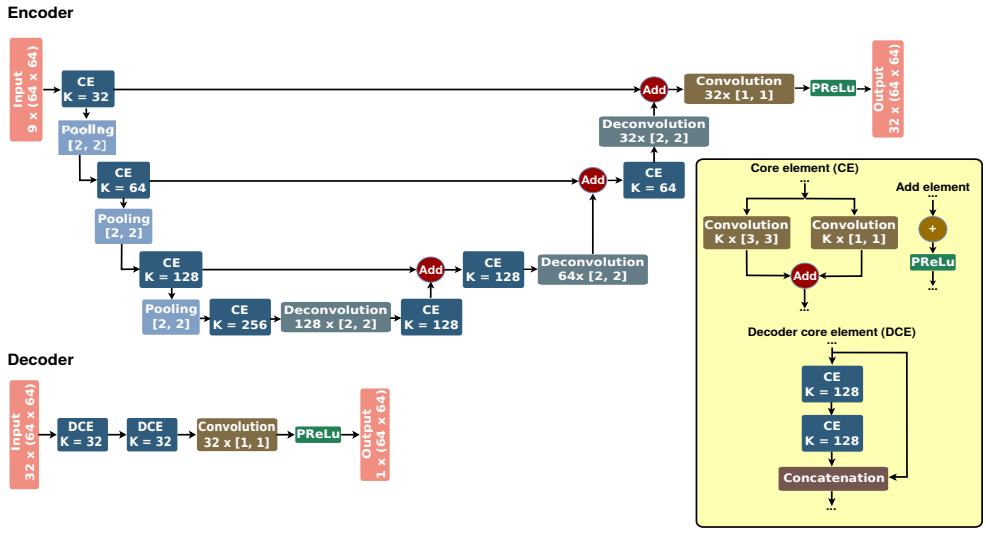
5.2.2 Data augmentation application: Generating new synthetic MS lesions

One of the applications of our synthetic MS lesion pipeline is to generate synthetic MS lesions on patient or healthy images and use these synthetic images as data augmentation to increase the MS lesion segmentation and detection performance. The main idea is to modify the original eight intensity level masks of the target image before passing it through the generator network. At testing time, if the intensity level masks are used without any modification, the output images are a generated synthetic version of the input ones containing all the WMHs found in the input image. Passing modified intensity level masks to the generator network will generate these desired modifications (i.e, new MS lesions) on the output images.

Figure 5.4 depicts how lesion expert annotations for a patient image can be generated on a healthy one through linear and nonlinear registration. After registration, the lesion mask and the eight intensity level masks of the patient subject are resampled to the healthy space. We split the resampled binary lesion mask into individual lesion volumes, in which every single lesion is defined as a spatially disconnected volume. After the lesion separation, the individual lesion volumes are dilated to incorporate the hyperintensities surrounding the lesions that are not annotated as lesion voxels. The intensity level masks of the dilated lesion volumes are copied from the patient resampled masks to the healthy masks. Finally, the healthy images plus their modified intensity level masks are passed through the generator network to add new MS lesions to the synthetic output images. In the same way, new MS lesions can be generated in patient images using patient-to-patient registration. In longitudinal MS analysis, MS lesions are added only to the follow-up scans. So, the new synthetic lesions on the follow-up image can be considered as new MS lesions with respect to the baseline image. The follow-up image with the synthetic lesions together with the untouched baseline are used as data augmentation to increase the longitudinal MS lesion detection performance.



(a) The MS lesion generation model.



(b) Encoder and decoder architectures.

Figure 5.3: MS lesion generator architecture. Each input modality has its own encoder that maps the input image modality to the 32-channel latent space. One decoder is learned for each output modality. The encoder maps the 32-channel latent representations to the outputs of that modality. Each decoder is used five times (i.e., shared decoder): once to decode each of the four individual latent representations and once to decode the fused representation. At testing time, we used the synthesis result from the fused representation as our output.

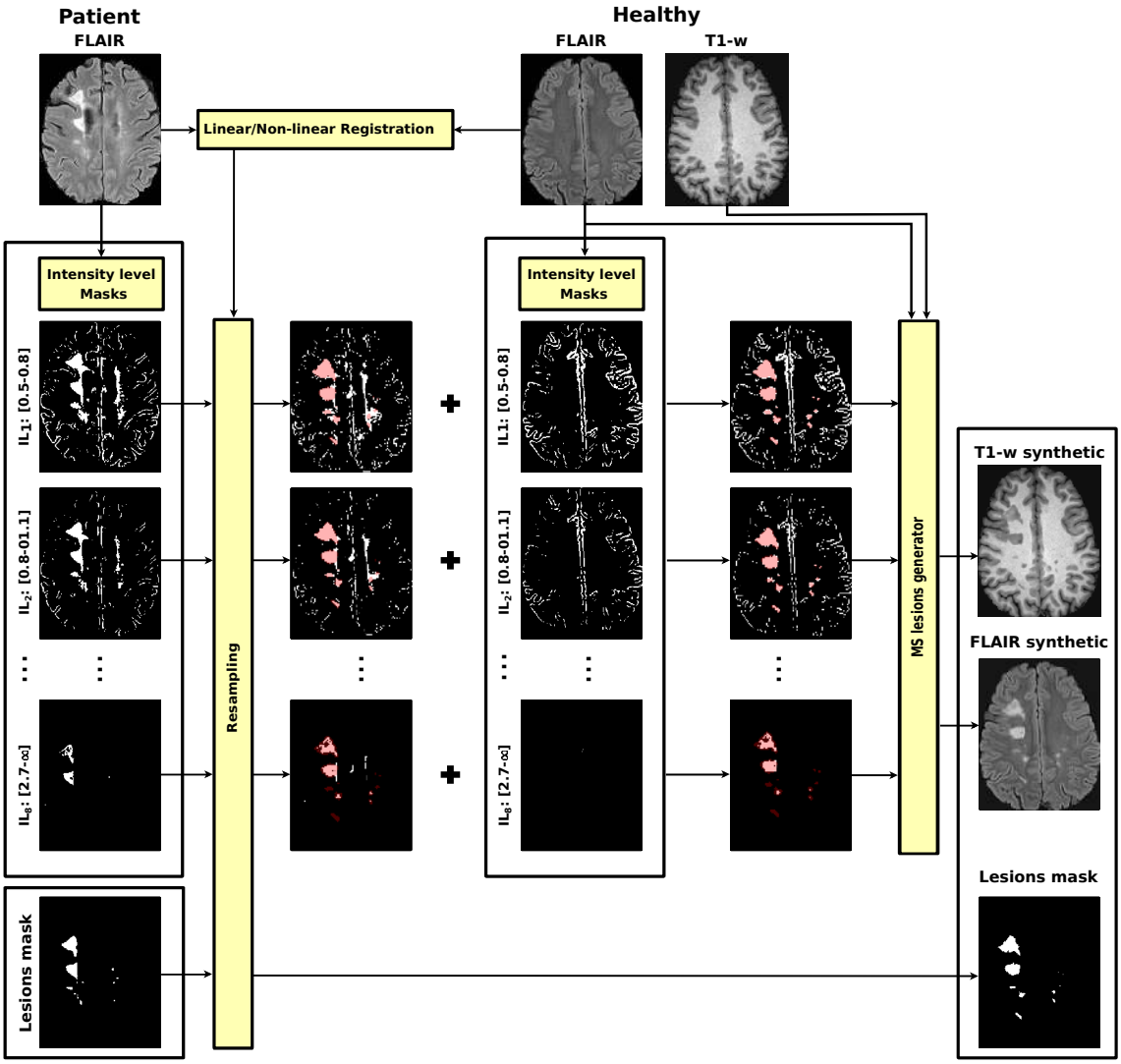


Figure 5.4: Generating MS lesions on healthy subjects using linear/nonlinear registration. After registering the patient FLAIR to the healthy FLAIR, the lesion mask and the intensity level masks of the patient were resampled to the healthy space. The lesions from the patient resampled intensity level masks were copied to the healthy intensity level masks. The healthy images combined with their modified intensity level masks were passed to the MS lesions generated to generate the synthetic MS lesions on the healthy image.

5.2.3 MS lesion segmentation approaches

In the cross-sectional MS lesion analysis, the segmentation framework used for evaluating the proposed MS lesion generator is the state-of-the-art CNN model proposed by Valverde et al. [28]. As described in section 2.2.3, their approach is based on a cascade of two 3D patch-wise CNN and was the best ranked approach on the MICCAI 2008 challenge and MICCAI MSSEG 2016. In the longitudinal MS lesion analysis, our FCNN-based model (SimLearnedDFs) proposed in chapter 4 is used for the evaluation. This DL-based model outperformed the LR-based model proposed in chapter 3 and all the state-of-the-art unsupervised approaches.

5.3 Experimental setup

5.3.1 Datasets

Cross-sectional clinical MS dataset: This dataset consists of 15 healthy subjects and 65 different patients with a CIS or early relapsing MS (Vall d'Hebron Hospital Center, Barcelona, Spain) who underwent brain MRI for monitoring disease evolution and treatment response. Each patient underwent brain MRI within the first 3 months after the onset of symptoms. The scans for all the patients were obtained in the same 3T magnet (Tim Trio; Siemens, Erlangen, Germany) with a 12-channel phased array head coil. The MRI protocol included the following sequences: 1) transverse proton density (PD)- and T2-w fast spin-echo (TR = 3080 ms, TE = $21 - 91$ ms, voxel size = $0.78 \times 0.78 \times 3.0$ mm^3), 2) transverse fast FLAIR (TR = 9000 ms, TE = 87 ms, TI = 2500 ms, flip angle = 120°, voxel size = $0.49 \times 0.49 \times 3.0$ mm^3), and 3) sagittal T1-w 3D magnetization-prepared rapid acquisition of gradient echo (TR = 2300 ms, TE = 2.98 ms, TI = 900 ms, voxel size = $1.0 \times 1.0 \times 1.2$ mm^3). The dataset was preprocessed as follows: for each patient, the T1-w image was linearly registered to the FLAIR using Nifty Reg tools[1] [220, 221]. Afterwards, a brain mask was identified and delineated on the registered T1-w image using the ROBEX Tool[2] [50]. Then, the two images underwent a bias field correction step using the N4 algorithm from the ITK library[3] with the standard parameters for a maximum of 400 iterations [219].

ISBI2015 dataset: This dataset consists of 5 training and 14 testing subjects with 4 or 5 different image time-points per subject from the ISBI2015 MS lesion challenge [227]. Each scan was imaged and preprocessed in the same manner by the own organizers, with data acquired on a 3.0 Tesla MRI scanner (Philips Medical Systems, Best, The Netherlands) with T1-w MPRAGE, T2-w, PD and FLAIR sequences. For more information about the image protocol and preprocessing de-

[1] https://sourceforge.net/projects/niftyreg/
[2] https://www.nitrc.org/projects/robex
[3] https://itk.org/Doxygen/html/classitk_1_1N4BiasFieldCorrectionImageFilter.html

tails, refer to the challenge organizers website[4]. On the challenge competition, each subject image was evaluated independently, which led to a final training set and a testing set composed of 21 and 61 images, respectively. Additionally, manual delineations of MS lesions performed by two experts were included for each of the 21 training images.

MICCAI2016 dataset: This dataset consists of 15 training scans acquired in three different scanner vendors: 5 scans (Philips Ingenia 3T), 5 scans (Siemens Aera 1.5T) and 5 scans (Siemens Verio 3T) from the MICCAI 2016 MS lesion segmentation challenge [259]. For each subject, 3D T1-w MPRAGE, 3D FLAIR, 3D T1-w gadolinium enhanced and 2D T2-w/PD-w images are provided. Please refer to the original publication for more details for the exact details of the acquisition parameters and image resolutions [259]. Manual lesion annotations for each training subject are provided as a consensus mask among 7 different human raters. Preprocessed images are already provided. The preprocessing pipeline consisted of a denoising step with the NL-means algorithm [260] and a rigid registration [261] of all of the modalities against the FLAIR image. Then, each of the modalities are skull-stripped using the volBrain platform [262] and bias corrected using the N4 algorithm [51]. Finally, all the training images were also interpolated to ($= 1.0 \times 1.0 \times 1.0$ mm^3)) using the FSL-FLIRT utility [263].

Longitudinal clinical MS dataset: This database is the same in-house dataset (VH dataset) used in the evaluation of our longitudinal approaches proposed in chapters 3 and 4. It consists of images from 60 different patients with a CIS or early relapsing MS. 36 of the patients confirmed MS with new T2-w lesions, while 24 patients did not present new T2-w lesions. The dataset was preprocessed the same way as described before. ROBEX Tool was used to get the brain mask. The four images underwent a bias field correction step using the N4. Finally, Nyúl et al. [61] histogram matching approach was used for intensity normalization. See section 3.3.1 for more details.

For all datasets, brain tissue volume was computed using the FAST segmentation method [88]. As explained in Section 5.2.1, the WMH mask and the eight intensity level masks were computed by FLAIR thresholding and all the used modalities were filled using the $\gamma = 0.5$ WMH mask.

5.3.2 MS lesion generator training and implementation

The proposed generator was evaluated to improve the performance of the cross-sectional and longitudinal MS lesions detection and segmentation approaches. In the cross-sectional evaluation, the proposed pipeline was used to generate MS lesions on T1-w and FLAIR images using only two encoders and two decoders, while in the longitudinal evaluation, it was extended to generate MS lesions on T1-w, T2-w, PD-w, and FLAIR images through the addition of two other encoder/decoder. To perform our experimental tests, we trained the lesion generator models into two

[4]http://iacl.ece.jhu.edu/index.php/MSChallenge/data

different scenarios, one being the cross-sectional MS clinical dataset and the other one the ISBI2015 dataset (see Table 5.1 for the images used for training). For training the generation network, 2D 64x64 patches with step size of 32x32 were extracted from the original images, the filled images, and the eight intensity level masks. The extracted patches were split into training and validation sets (70% for training and 30% for validation). The training set was used to adjust the weights of the neural network, while the validation set was used to measure how well the trained model was performing after each epoch. The extracted patches were passed to the network for training in mini batches of size 32 and the network was set to train for 200 epochs. To prevent overfitting, the training process was automatically terminated when the validation accuracy did not increase after 15 epochs. Regarding the MS lesion segmentation framework, the CNN training and inference procedures were identical to those proposed by Valverde et al. [28].

The proposed method has been implemented in Python, using Keras with the TensorFlow backend [251]. All experiments have been run on a GNU/Linux machine box running Ubuntu 18.04, with 128 GB RAM memory. The model training was carried out on a single TITAN-X GPU (NVIDIA Corp, United States) with 12 GB RAM memory. To promote the reproducibility and usability of our research, the proposed MS lesion generation pipeline is currently available for downloading at our research website[5].

5.3.3 Evaluation metrics

To evaluate the performance of the proposed MS lesion generator, we computed the similarity between the original and the synthetic images using the following similarity metrics:

- Mean Square Error (MSE):

$$MSE(G, R) = \frac{1}{N}\sum_{i=1}^{N}(G_i - R_i)^2$$

  where G and R are the intensities of the generated and the real images, respectively, and N is the number of voxels in the R image.

- Structural Similarity Index (SSIM):

$$SSIM(G, R) = \frac{(2\mu_G\mu_R + c_1)(2\sigma_{GR} + c_2)}{(\mu_G^2 + \mu_R^2 + c_1)(\sigma_G^2 + \sigma_R^2 + c_2)}$$

  where (μ_G, σ_G^2) and (μ_R, σ_R^2) are the intensity's (mean, variance) of the generated and the real images, respectively, and σ_{GR} is the covariance between

[5]https://github.com/NIC-VICOROB/MS_Lesions_Generator

Table 5.1: Datasets. Total number of images, images used for training and testing the MS lesion generator, and images used for training and testing the MS lesion segmentation model for the cross-sectional clinical MS and ISBI2015 datasets.

Datasets	Total number of images	MS lesion generator	MS lesion segmentation
Cross-sectional clinical MS dataset	- 65 patient images: Group A: 36 images Group B: 29 images	Training: Group A: 36 images Testing: Group B: 29 images	Group B is split into: Training: VHtrain: 15 images Testing: VHtest: 14 images
	VHhealthy:15 healthy images		
ISBI2015 dataset	ISBItrain: 21 patient images	Training: ISBItest	Training: ISBItrain
	ISBItest: 61 patient images	Testing: ISBItrain	Testing: ISBItest
Longitudinal clinical MS dataset	LongNewLesions: 36 patient images LongNoNewLesions: 24 patient images		Training: LongTrain Testing: LongTest

them, c_1 and c_2 are two constants to stabilize the division with weak denominator. SSIM actually measures the perceptual difference between two similar images.

On the other hand, the quantitative evaluation of the proposed MS lesion generator was performed by segmenting both the original and synthetic images individually using the same MS lesion segmentation framework and comparing the difference between the segmentation results. As explained before, the segmentation framework used to evaluate the proposed MS lesion generator is the MS lesion segmentation method proposed by Valverde et al. [28], although the proposed data augmentation strategy could be applied to any approach. The evaluation of the resulting segmentations against the available lesion annotations was carried out using standard evaluation metrics such as DSC, sensitivity, and precision. A paired t-test at the 5% level was used to evaluate the significance of the data augmentation results. Significant results are shown in bold in all tables. For the longitudinal evaluation, the same standard measures such as TPF, FPF, and the Dice were used for the quantitative analysis as described in section 4.3.3. The automatic segmentation masks were obtained by thresholding the probability maps with 0.5 (using argmax), and all automatic lesions with a size of 3 voxels or less were removed.

5.4 Cross-sectional: Experiments and results

5.4.1 MS lesion synthesis

In these experiments, qualitative and quantitative evaluations were undertaken by measuring the similarities between the real and the synthetic images in terms of MSE and SSIM metrics and in terms of cross-sectional MS lesion detection and segmentation using a state-of-the-art MS lesion segmentation method [28] and the evaluation metrics described in section 5.3.3 (see Table 5.1 for the images used).

Evaluation

Cross-sectional clinical MS dataset: Both VHtrain and VHtest sets were generated using the proposed MS generator yielding VHtrainGen and VHtestGen, respectively. The evaluation of the proposed MS generator on this dataset was performed by measuring the MSE and SSIM metrics between the real and the synthetic images (using Group B images, see Table 5.1) and by training and testing the MS lesion segmentation model [28] as follows: 1) training with the VHtrain set and testing on the VHtest set; 2) training with the VHtrainGen set and testing on the VHtestGen set; 3) training with the VHtrainGen set and testing on the VHtest set; and 4) training with the VHtrain set and testing on the VHtestGen set.

ISBI2015 dataset: The ISBItrain set was generated using the proposed MS generator yielding ISBItrainGen. Note that the evaluation of the ISBI 2015 chal-

lenge is performed blind by submitting the segmentation masks of the 61 testing cases to the challenge website evaluation platform[6]. The evaluation of the proposed MS generator on this dataset was performed by measuring the MSE and SSIM metrics between the real and the synthetic images (using ISBItrain set, see Table 5.1). The performance of the two MS lesion segmentation models, one trained with the ISBItrain set and the other trained with the ISBItrainGen set, was evaluated by submitting to the challenge's evaluation platform, and comparing the accuracy between them.

MS lesion generation on healthy subjects: To evaluate the generation of MS lesions on healthy subjects by using registration, the MS lesions of the VHtrain dataset were generated on the VHhealthy images using linear and nonlinear registration as described in section 5.2.2. We refer to them as VHGenLinear and VHGenNonlinear, respectively. The evaluation of the proposed MS generator on these datasets was performed by training 3 MS lesion segmentation models using the VHGenLinear, the VHGenNonlinear, and (VHGenLinear + VHGenNonlinear) and testing on the VHtest set.

Results

Table 5.2 summarizes the MSE and SSIM between the real and synthetic images of the cross-sectional clinical MS and ISBI2015 datasets. Furthermore, the MSE and SSIM of $\gamma = 0.5$ WMH mask voxels are reported. The MSE and SSIM between the non-background voxels are better than $\gamma = 0.5$ WMH mask voxels. Also, we can see that the synthetic FLAIR images are close to the real ones than T1-w images inside $\gamma = 0.5$ WMH mask voxels for the two datasets. Figure 5.5 and 5.6 show the qualitative assessment of the proposed MS lesion generator of the cross-sectional clinical MS and ISBI2015 datasets, respectively. Figure 5.7 shows the qualitative assessment of the proposed MS lesion generator of the synthetic MS lesions generated on healthy subjects using linear/nonlinear registration. The slices are also displayed using jet color maps to show the similarity of intensities inside the original and the synthetic lesions. One of the advantages of using the intensity level masks described in section 5.2.1 is the appearance of the gradients inside the lesions of the synthetic images. Table 5.3 summarizes the MS lesion detection and segmentation results, showing the obtained mean values when training with the original and synthetic images of the clinical MS and ISBI2015 datasets. The performance was very similar in terms of the three evaluation metrics when training with real or synthetic images and testing on the real images for the two datasets. The mean results when training with the synthetic MS lesions generated on healthy images using the clinical MS dataset lesion set are shown in Table 5.4.

[6]https://smart-stats-tools.org/node/26

Table 5.2: Similarity results. MSE and SSIM between the original and synthetic images of the cross-sectional clinical MS (Group B set) and ISBI2015 (ISBItrain set) datasets for nonbackground and $\gamma = 0.5$ WMH mask. The reported values are the mean ± standard deviation.

Cross-sectional clinical MS dataset (Group B set)

	Non-background voxels		$\gamma = 0.5$ WMH mask voxels	
	MSE	SSIM	MSE	SSIM
T1-w	0.03 ± 0.01	0.96 ± 0.01	0.07 ± 0.03	0.93 ± 0.03
FLAIR	0.02 ± 0.01	0.98 ± 0.02	0.03 ± 0.07	0.98 ± 0.01

ISBI2015 dataset (ISBItrain images)

	Non-background voxels		$\gamma = 0.5$ WMH mask voxels	
	MSE	SSIM	MSE	SSIM
T1-w	0.03 ± 0.01	0.97 ± 0.01	0.13 ± 0.05	0.94 ± 0.03
FLAIR	0.01 ± 0.01	0.98 ± 0.01	0.01 ± 0.01	0.99 ± 0.01

Table 5.3: Lesion segmentation and detection results. Comparison between the training using original images and synthetic images on Cross-sectional clinical MS and ISBI2015 datasets. For each coefficient (*DSC*, *sensitivity*, and *precision*), the reported values are the mean ± standard deviation. For the ISBI2015 dataset, the reported values are extracted from the challenge results board.

Cross-sectional clinical MS dataset

Train/Test	*DSC*	*sensitivity*	*precision*
VHtrain/VHtest	0.70 ± 0.16	0.69 ± 0.13	0.73 ± 0.15
VHtrainGen/VHtest	0.68 ± 0.16	0.72 ± 0.14	0.71 ± 0.13
VHtrainGen/VHtestGen	0.67 ± 0.17	0.65 ± 0.14	0.70 ± 0.17
VHtrain/VHtestGen	0.68 ± 0.15	0.66 ± 0.15	0.70 ± 0.16

ISBI2015 dataset

Train/Test	*DSC*	*sensitivity*	*precision*
ISBItrain/ISBItest	0.64 ± 0.12	0.57 ± 0.16	0.79 ± 0.12
ISBItrainGen/ISBItest	0.64 ± 0.13	0.56 ± 0.17	0.80 ± 0.14

Table 5.4: Cross-sectional clinical MS dataset results of training using synthetic images generated on healthy subjects as described in section 5.2.2. For each coefficient (*DSC*, *sensitivity*, and *precision*), the reported values are the mean ± standard deviation.

Train/Test	*DSC*	*sensitivity*	*precision*
VHGenLinear/VHtest	0.63 ± 0.21	0.63 ± 0.17	0.63 ± 0.16
VHGenNonlinear/VHtest	0.63 ± 0.20	0.62 ± 0.14	0.62 ± 0.16
Both/VHtest	0.65 ± 0.20	0.64 ± 0.14	0.64 ± 0.17

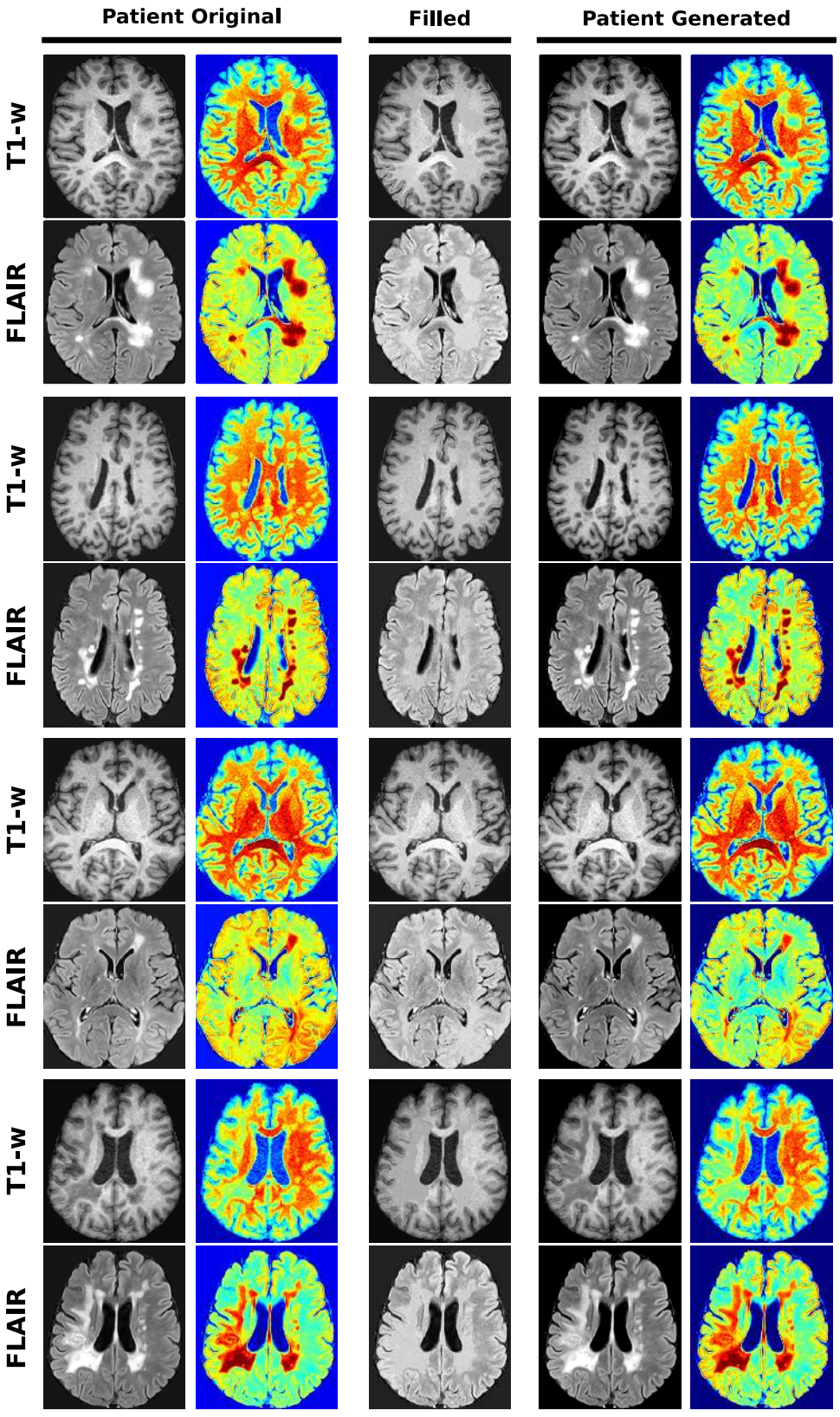


Figure 5.5: Qualitative assessment of the proposed MS lesions generator on cross-sectional clinical MS dataset. Slices are also displayed using jet color maps to visually enhance the intensities.

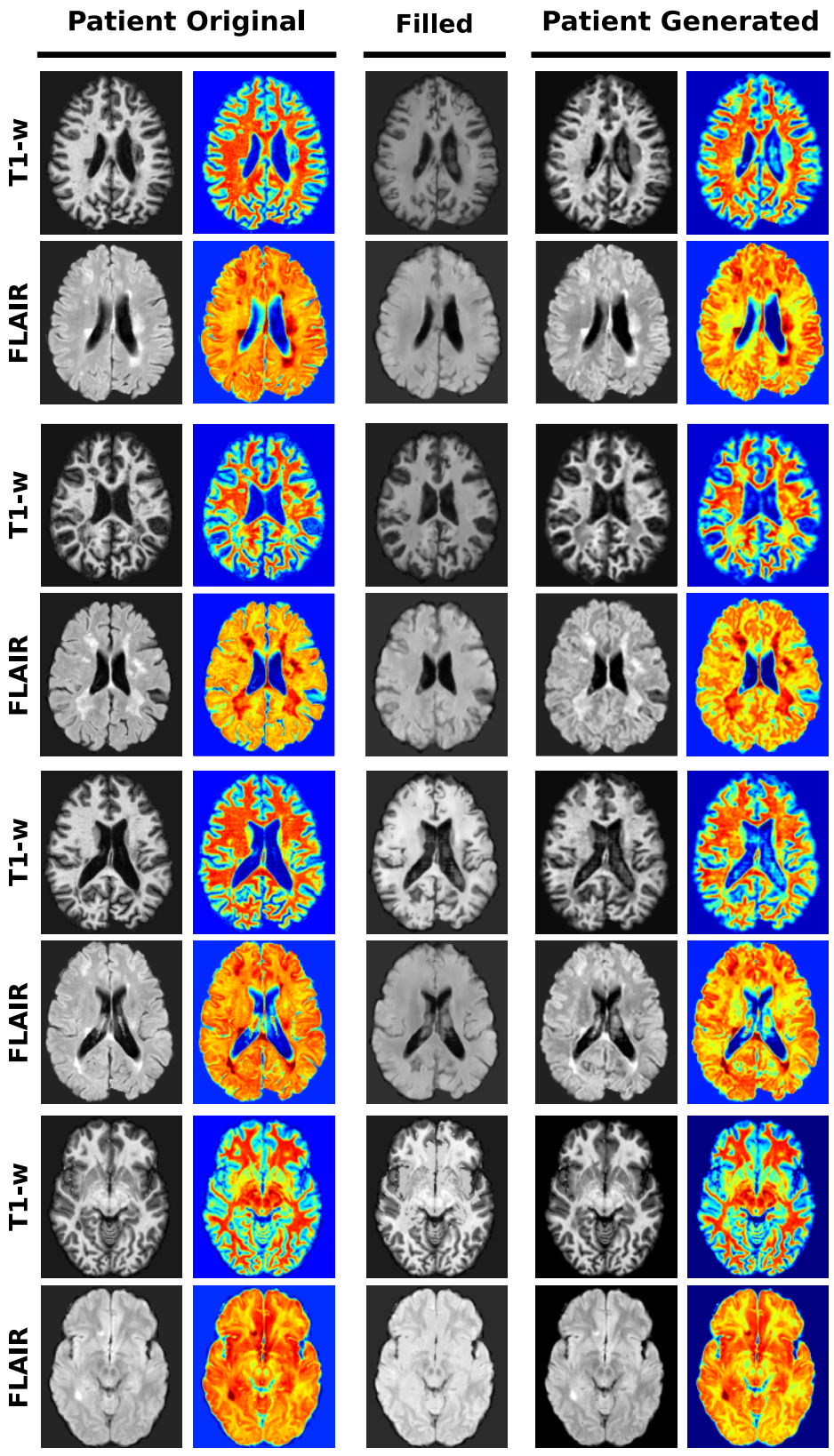


Figure 5.6: Qualitative assessment of the proposed MS lesions generator on ISBI2015 dataset. Slices are also displayed using jet color maps to visually enhance the intensities.

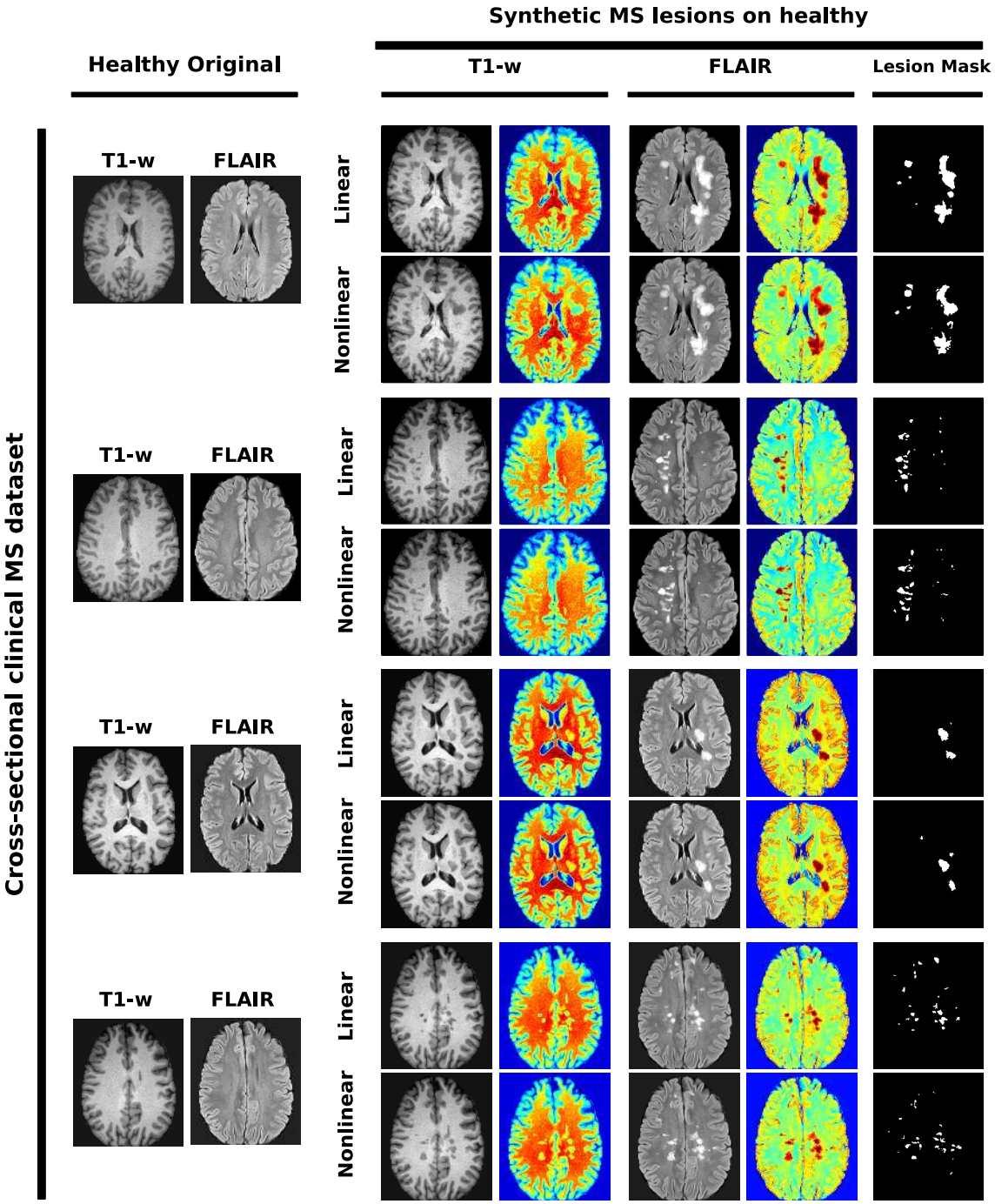


Figure 5.7: Synthetic MS lesions generated on a healthy subject using linear/nonlinear registration. Slices are also displayed using jet color maps to visually enhance the intensities.

Discussion

We demonstrated the similarity between the synthetic and real lesions qualitatively and quantitatively on patient and healthy subjects. Synthetic images are very similar to the real ones in terms of the two similarity metrics for nonbackground and $\gamma = 0.5$ WMH mask voxels for both datasets. Regarding the MS lesion segmentation results, the experiments show how similar the training is using real or synthetic images in terms of MS lesion detection. Regarding the MS clinical dataset, the performance is 2% less in terms of DSC and precision when training with the synthetic images than training with the real images. However, similar results were obtained when training with real images and testing on synthetic images. From the results obtained, synthetic images could be used as training or testing images. We only used synthetic images as testing images to evaluate how good they are when training with real images. Regarding the ISBI2015 datasets, the performance was very similar in terms of the three evaluation metrics. Regarding the training using synthetic MS lesions generated on healthy subjects, good segmentation and detection results were obtained when training with synthetic images generated on healthy subjects. The performance is also very similar when training with synthetic images generated using linear, nonlinear registration or both.

5.4.2 Data augmentation experiments

In these experiments, we evaluated the use of the proposed MS lesion generator as a data augmentation method by generating the lesion masks on healthy images from the same domain using registration as described in section 5.2.2. The two deformed generated lesion masks (from linear and nonlinear registration) and the correspondent two synthetic images were added to the original patient image during training as data augmentation.

Evaluation

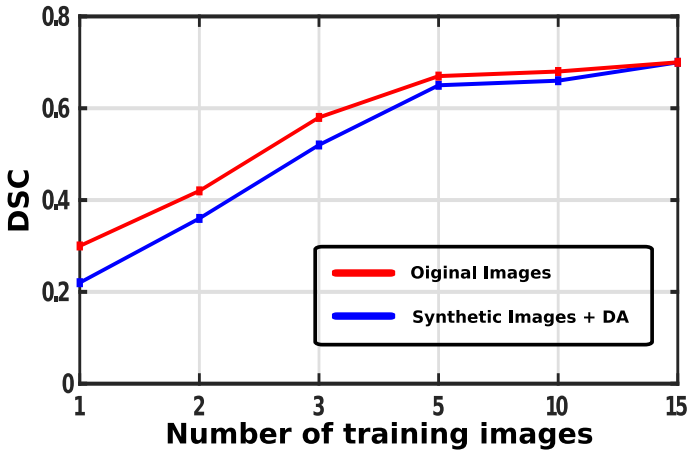
Cross-sectional clinical MS dataset: For each patient image from the VHtrain set, we created two synthetic images with lesions on a healthy image from the VHhealthy set (VHGenLinear and VHGenNonlinear) as described in section 5.2.2. Those two synthetic images were used together with the original image as data augmentation in the following experimental tests: 1) to analyze the effect of the synthetic data augmentation images on the segmentation performance while training with different number of training images, two models were trained using 1, 2, 3, 5, 10 or all of the available training images, with one model using the original images and the other using the same original images plus their synthetic data augmentation images; and 2) to simulate a situation with limited training data, we analyzed the effect of the synthetic data augmentation on the segmentation performance in the scenario of having only one-image for training. Using a single training image with a lesion volume in the range of $0.34 - 49.4$ ml, two models were trained. One

model used the original image (i.e., from VHtrain) and the other used the same original image plus the two synthetic images generated on the healthy image (i.e., from VHGenLinear and VHGenNonlinear).

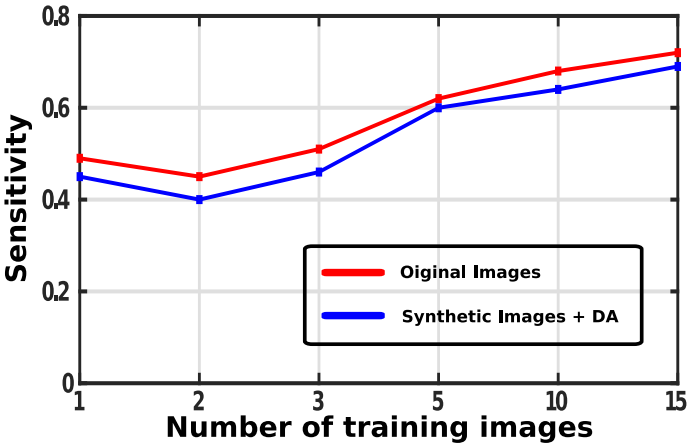
ISBI2015 dataset: To simulate a situation with limited training data, we analyzed the effect of the synthetic data augmentation images on the segmentation performance in the one-image training scenario on the overall performance of the testing set. To do so, we chose a single training image from each training subject (ISBItrain), which led to 5 different training sets with a varying number of lesions and a total lesion volume in the range $2.3 - 26.8$ ml. Since there were no healthy subjects available from this challenge, we chose the fourth training subject (this image has the smallest lesion load; ≈ 2.3 ml) and filled it as described in section 5.2.1 (but only MS lesions were filled instead of the WMH areas). We considered this image as a healthy subject and we refer to it as ISBI-H. The MS lesions of each of the four selected ISBI images were generated on the ISBI-H using linear and nonlinear registration, as described in section 5.2.2, yielding, for each patient image from the selected four, two generated images and their correspondent lesion masks that were used as data augmentation. Based on this, we undertook the following experiments. 1) To simulate a situation with limited training data, we analyzed the effect of the synthetic data augmentation images on the segmentation performance in the one-image training scenario. Using a single training image from the four images selected, two models were trained, one using the original image and the other using the original image plus its two synthetic images generated on ISBI-H using linear and nonlinear registration. 2) To determine the performance of all the models trained on the blind test set, all trained models from the previous experiment were sent to the challenge's evaluation platform, comparing its accuracy to those of the other submitted MS lesion segmentation pipelines fully trained using the entire available training set. Among the set of evaluated coefficients computed in the challenge, only the DSC, sensitivity and precision metrics are shown for comparison.

Results

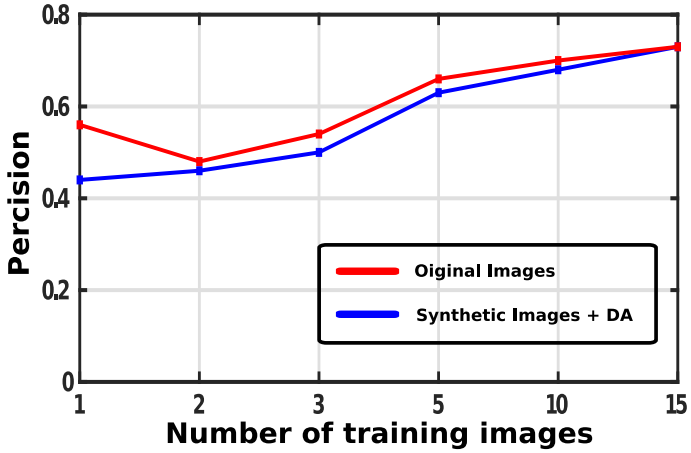
Regarding the cross-sectional clinical MS dataset, Figure 5.8 shows the DSC, sensitivity and precision coefficients of different models trained using different number of training images, which ranged from 1 to 15 images. Table 5.5 shows the DSC, sensitivity and precision coefficients of the models under the one-image training scenario. Regarding the ISBI2015 dataset, Table 5.6 shows the performance of each of the one-image scenario models when trained on different images with varying degrees of lesion size. Table 5.7 shows the performance of the models trained with ISBI02 plus DA against different top rank participant challenge strategies. From the list of compared methods, the best five strategies were based on CNN models (Andermatt et al. [264], Salehi et al. [265], Valverde et al. [28], Birenbaum and Greenspan [266]), while the others were based on either other supervised learning techniques (Valcarcel et al. [267], Deshpande et al. [268], Sudre et al. [269]) or unsupervised intensity models (Shiee et al. [270], Jain et al. [271]).



(a)



(b)



(c)

Figure 5.8: Effect of the number of training images and their DA images on the DSC, sensitivity and precision coefficients when evaluated on the cross-sectional clinical MS dataset. The represented value for each configuration is computed as the mean DSC, sensitivity and precision scores over the 14 VHtest images.

Table 5.5: One-image scenario for the cross-sectional clinical MS dataset: *DSC*, *sensitivity* and *precision* coefficients for two models, one model trained using a single original image (ORG) and the other one trained using same single image plus its synthetic data augmentation images (DA) with varying degrees of lesion load. For each coefficient, the reported values are the mean ± standard deviation when evaluated on the VHtest set. Significant results are shown in bold ($p < 0.05$).

lesion vol (num lesions)		*DSC*	*sensitivity*	*precision*
0.34 ml (18 lesions)	ORG	0.18 ± 0.11	0.43 ± 0.13	0.41 ± 0.11
	DA	$\mathbf{0.29 \pm 0.14}$	$\mathbf{0.48 \pm 0.15}$	$\mathbf{0.61 \pm 0.20}$
1.0 ml (6 lesions)	ORG	0.35 ± 0.25	0.23 ± 0.15	0.35 ± 0.29
	DA	$\mathbf{0.47 \pm 0.25}$	$\mathbf{0.27 \pm 0.14}$	0.37 ± 0.21
2.0 ml (25 lesions)	ORG	0.53 ± 0.19	0.43 ± 0.16	0.62 ± 0.26
	DA	$\mathbf{0.57 \pm 0.20}$	$\mathbf{0.54 \pm 0.14}$	0.64 ± 0.28
5.5 ml (15 lesions)	ORG	0.28 ± 0.15	0.28 ± 0.12	0.32 ± 0.14
	DA	$\mathbf{0.32 \pm 0.12}$	$\mathbf{0.35 \pm 0.13}$	$\mathbf{0.38 \pm 0.12}$
7.6 ml (42 lesions)	ORG	0.57 ± 0.25	0.41 ± 0.16	0.53 ± 0.23
	DA	$\mathbf{0.63 \pm 0.20}$	$\mathbf{0.50 \pm 0.16}$	$\mathbf{0.65 \pm 0.21}$
21.5 ml (181 lesions)	ORG	0.61 ± 0.21	0.61 ± 0.18	0.54 ± 0.14
	DA	0.63 ± 0.20	$\mathbf{0.66 \pm 0.14}$	$\mathbf{0.60 \pm 0.14}$
49.4 ml (53 lesions)	ORG	0.57 ± 0.25	0.58 ± 0.20	0.56 ± 0.22
	DA	0.58 ± 0.25	$\mathbf{0.67 \pm 0.18}$	$\mathbf{0.60 \pm 0.13}$

Table 5.6: One-image scenario for the ISBI2015 dataset: *DSC*, *sensitivity*, *precision*, and overall score coefficients for two models, one model trained using a single original image (ORG) and the other one trained using same single image plus its synthetic data augmentation images (DA). The reported values are extracted from the challenge results board. For each coefficient, the reported values are the mean ± standard deviation when evaluated on the ISBItest set. Significant results are shown in bold ($p < 0.05$).

lesion vol (num lesions)		*DSC*	*sensitivity*	*precision*	*score*
ISBI01	ORG	0.41 ± 0.13	0.30 ± 0.12	0.75 ± 0.19	87.60
	DA	$\mathbf{0.54 \pm 0.13}$	$\mathbf{0.45 \pm 0.15}$	0.75 ± 0.17	89.54
ISBI02	ORG	0.53 ± 0.18	0.44 ± 0.19	0.76 ± 0.21	88.60
	DA	$\mathbf{0.59 \pm 0.15}$	$\mathbf{0.51 \pm 0.19}$	$\mathbf{0.78 \pm 0.18}$	90.05
ISBI03	ORG	0.49 ± 0.13	0.39 ± 0.14	0.74 ± 0.18	88.67
	DA	0.49 ± 0.12	0.39 ± 0.14	$\mathbf{0.77 \pm 0.15}$	89.55
ISBI05	ORG	0.39 ± 0.13	0.29 ± 0.13	0.73 ± 0.17	88.02
	DA	$\mathbf{0.42 \pm 0.13}$	0.30 ± 0.12	$\mathbf{0.79 \pm 0.16}$	88.66

Table 5.7: ISBI2015 challenge: *DSC*, *sensitivity*, *precision* and overall score coefficients for the best one-image scenario with the data augmentation model (ISBI02 + DA). The obtained results are compared with different top rank participant strategies and also with the same model fully trained on all the available data. For each method, the reported values are extracted from the challenge results board. The reported values are the mean (standard deviation) when evaluated on the 61 testing images. The performance of the methods with an overall $score \geq 90$ is considered to be similar to human performance.

Method	*DSC*	*sensitivity*	*precision*	*score*
Andermatt et al. [264]	0.63 ± 0.14	0.54 ± 0.19	0.84 ± 0.10	92.07
Salehi et al. [265]	0.66 ± 0.11	0.67 ± 0.20	0.71 ± 0.16	91.52
Valverde et al. [28]	0.64 ± 0.12	0.57 ± 0.17	0.79 ± 0.15	91.44
Birenbaum and Greenspan [266]	0.63 ± 0.14	0.55 ± 0.18	0.80 ± 0.15	91.26
Deshpande et al. [268]	0.60 ± 0.13	0.55 ± 0.17	0.73 ± 0.18	89.81
Jain et al. [271]	0.55 ± 0.14	0.47 ± 0.15	0.73 ± 0.20	88.74
Shiee et al. [270]	0.55 ± 0.19	0.54 ± 0.15	0.70 ± 0.29	88.46
Valcarcel et al. [267]	0.57 ± 0.13	0.57 ± 0.18	0.61 ± 0.16	87.71
Sudre et al. [269]	0.52 ± 0.14	0.46 ± 0.15	0.66 ± 0.18	86.44
Full train	0.63 ± 0.13	0.55 ± 0.16	0.79 ± 0.14	91.33
ISBI02 + DA	0.59 ± 0.15	0.51 ± 0.19	0.78 ± 0.18	90.05

Discussion

We demonstrated the effect of data augmentation on the MS lesion segmentation performance when increasing the number of the training images. The difference in performance between training with original images and original images plus DA decreases in terms of the three metric coefficients as the number of the training images increases. The DA images generated from linear and nonlinear registration do not give more variability to the training data when increasing the number of training images. Furthermore, to simulate a situation with limited training data, we analyzed the effect of one-image training scenario. Regarding the MS clinical dataset, significant improvement was obtained in terms of the three metric coefficients with a lesion volume in the range of $0.34 - 49.4$ ml. Regarding the ISBI2015 dataset, a significant improvement was obtained in terms of the three metric coefficients, except for ISBI03, where only a significant improvement in precision was obtained. Comparing the accuracy of the best performing model (ISBI02+DA) to those of the other submitted MS lesion segmentation pipelines fully trained using the entire available training set, the proposed one image plus its data augmentation images reported a performance similar to that of the same fully trained cascaded CNN architecture (score 91.44) [28], which shows the improvement of the proposed data augmentation strategy to the training used with limited training data.

5.5 Longitudinal: Experiments and results

5.5.1 Longitudinal synthetic lesions

In these experiments, qualitative and quantitative evaluations were undertaken in terms of longitudinal MS lesion detection using our approach (SimLearnedDFs) proposed in chapter 4 and the evaluation metrics described in section 5.3.3.

Evaluation

First, the lesion masks of the three cross-sectional datasets (cross-sectional clinical MS (Group B), MICCAI2016, and ISBI2015) were used to generate synthetic MS lesions on the follow-up images using the LongNoNewLesions images (the 24 longitudinal images with no new MS lesions). The baseline images of LongNoNewLesions were also generated but without adding any synthetic lesions. For ISBI2015, we chose a single training image from each training subject (ISBItrain). Some lesion masks from ISBItrain and MICCAI2016 datasets were duplicated to have 24 lesion masks. The generator was trained using the 36 cross-sectional images (Group A) described in table 5.1. So, we had three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI, and Long-ISBI) with 24 images each yielding from cross-sectional clinical MS (Group B), MICCAI2016, and ISBI2015 datasets, respectively. The evaluation of the proposed MS generator was performed by training our FCNN-based model (SimLearnedDFs) proposed in chapter 4 with each of the three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI, and Long-ISBI) and testing on the LongNewLesions images (the 36 longitudinal images with new MS lesions). We also included two unsupervised state-of-the-art approaches for comparison [33, 149]. A paired t-test at the 5% level was used to evaluate the significance of the results of using the synthetic longitudinal datasets and two unsupervised state-of-the-art approaches [33, 149].

Results

Figure 5.9 shows the qualitative assessment of the proposed MS lesion generator of the synthetic MS lesions generated on the follow-up images with no new MS lesions using linear/nonlinear registration. The slices are also displayed using jet color maps to show the similarity of intensities inside the original and the synthetic lesions. Table 5.8 summarizes the new MS lesion detection results, showing the obtained mean values when training with the three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI2016, and Long-ISBI2015) and testing on the LongNewLesions set. Moreover, Figure 5.10 shows a visual example of new MS lesion detection from the obtained results.

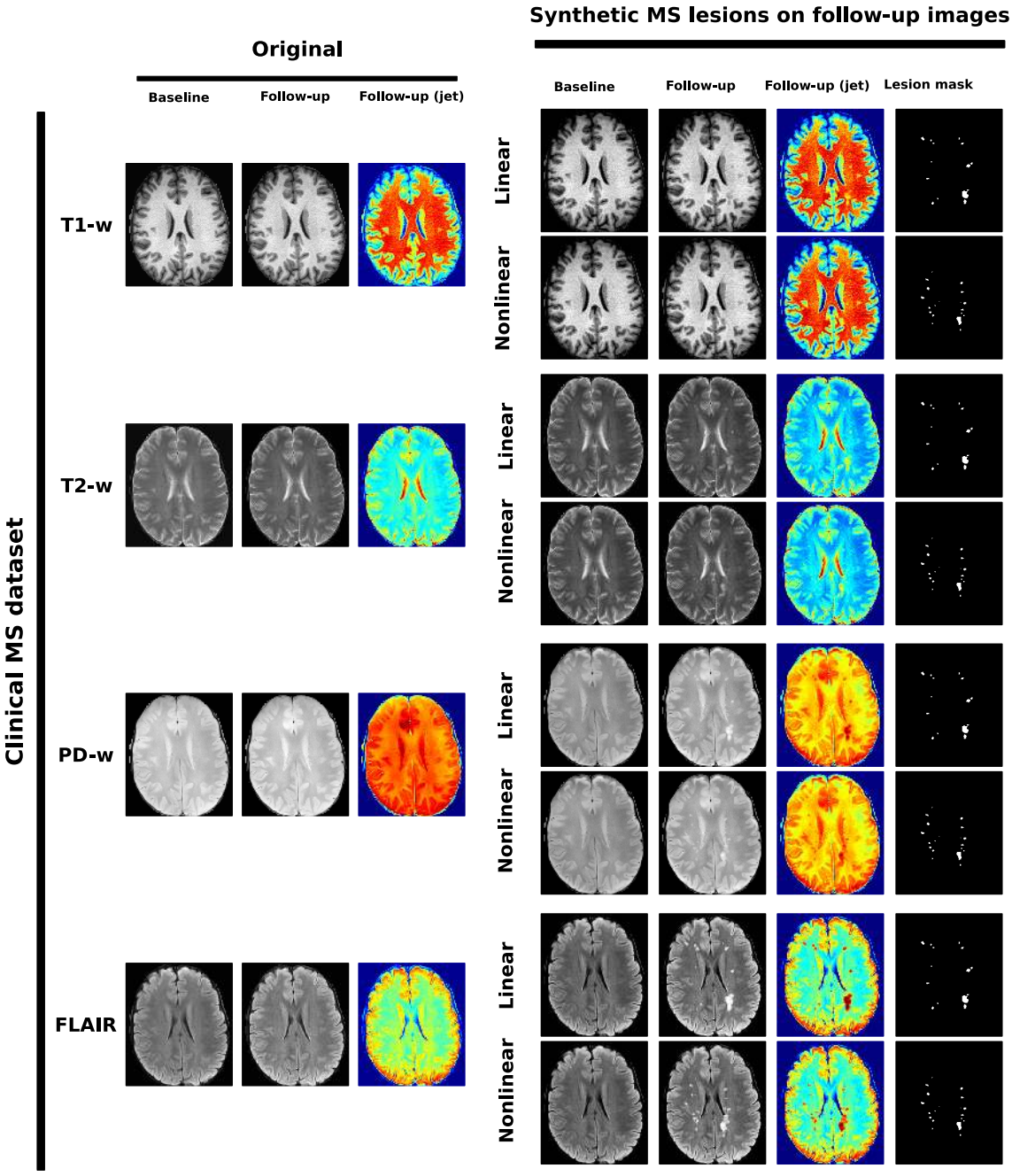


Figure 5.9: Synthetic MS lesions generated on the follow-up images with no new MS lesions using linear/nonlinear registration. Slices are also displayed using jet color maps to visually enhance the intensities.

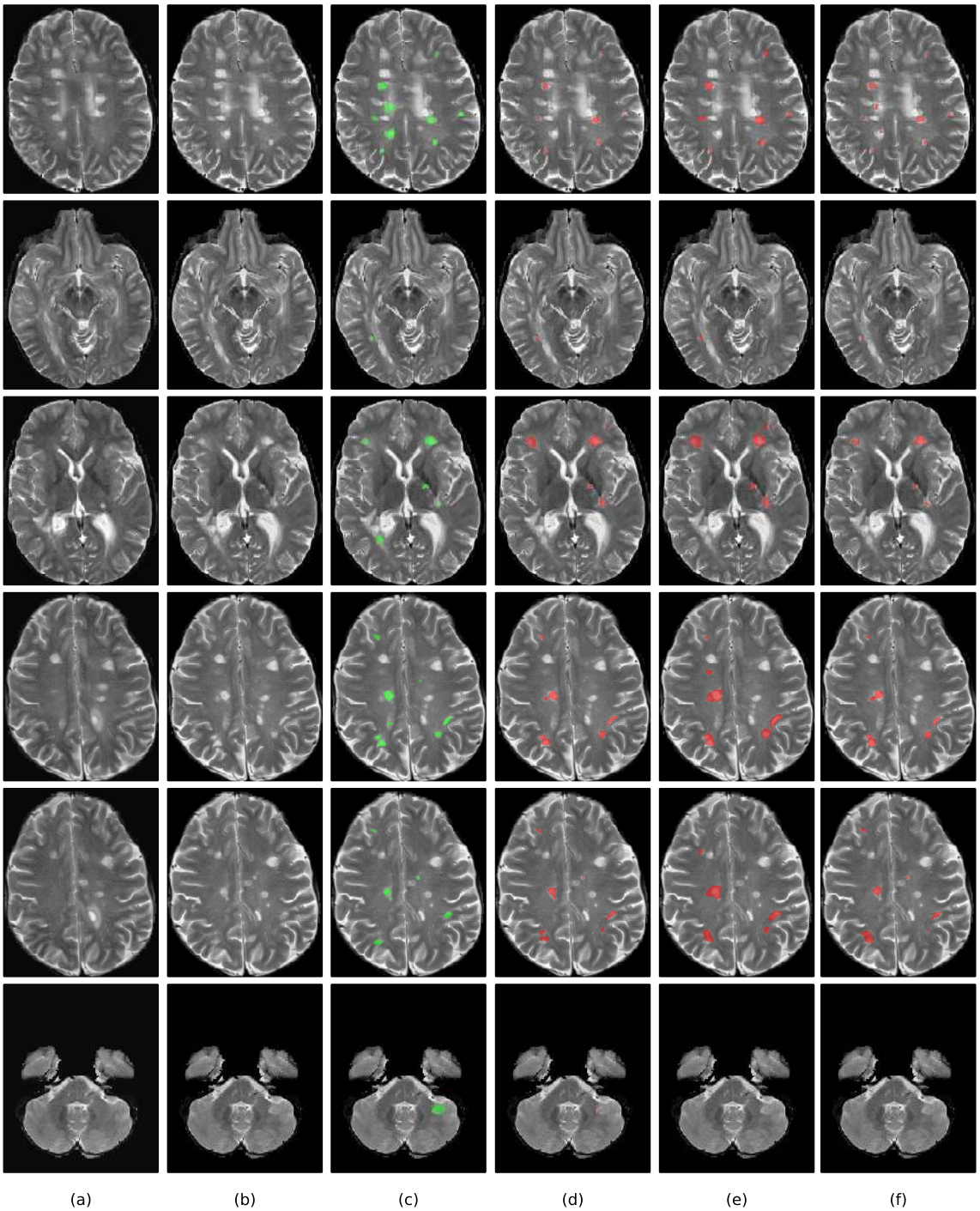


Figure 5.10: An example of new MS lesion detection when training with synthetic longitudinal datasets. (a) and (b) show one axial slice of the T2-w image at baseline and follow-up, respectively. (c) shows the new MS lesions annotations performed by an expert (GT). (d), (e), and (f) show the segmentation when training the SimLearnedDFs model with each of Long-Clinical, Long-MICCAI, and Long-ISBI datasets, respectively. The GT and the segmentations are overlaid in green and red, respectively, on the follow-up T2-w image.

Table 5.8: Comparison between training the SimLearnedDFs model with each of the three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI, and Long-ISBI). The results represent the mean detection TPF, FPF, DSC_d and mean segmentation DSC_s when testing the 36 MS patients (LongNewLesions images). The automatic segmentation masks were obtained by thresholding the probability maps with 0.5 (using argmax), and all automatic lesions with a size lower than three voxels were removed.

Training dataset	TPF	FPF	DSC_d	DSC_s
Long-Clinical	78.16 ± 24.90	25.27 ± 27.62	0.71 ± 0.25	0.50 ± 0.23
Long-MICCAI	69.13 ± 30.64	17.06 ± 26.38	0.68 ± 00.29	0.46 ± 00.24
Long-ISBI	64.44 ± 35.23	25.31 ± 27.96	0.60 ± 0.32	0.38 ± 0.25
Cabezas et al. [33]	70.93 ± 34.48	17.80 ± 27.96	0.68 ± 0.33	0.53 ± 0.24
Schmidt et al. [149]	68.66 ± 35.26	31.89 ± 36.10	0.62 ± 0.34	0.40 ± 0.25

Discussion

The experiments show that the longitudinal synthetic datasets generated using cross-sectional MS lesions could be used alone for training a new lesion detection model. From the obtained results, training with Long-Clinical was significantly better than the two unsupervised approaches [33, 149] in terms of TPF while training with the other two datasets (Long-MICCAI, and Long-ISBI) were similar to the unsupervised approaches ($p < 0.05$).

5.5.2 Data augmentation experiments

In these experiments, we evaluated the use of the proposed MS lesion generator as a longitudinal data augmentation method. The three synthetic longitudinal datasets (Long-Clinical, Long-ISBI, and Long-MICCAI) were used as data augmentation.

Evaluation

The LongNewLesions images (the 36 longitudinal images with new MS lesions) were randomly split into training set (LongTrain) and testing set (LongTest). Each set had 18 images. The evaluation of the proposed MS generator was performed by training and testing the SimLearnedDFs model as follows: 1) training with the LongTrain set and testing on the LongTest set; 2) training with the LongTrain set together with each of the three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI, and Long-ISBI) as data augmentation and testing on the LongTest set. A paired t-test at the 5% level was used to evaluate the significance of the results of using the synthetic longitudinal datasets as data augmentation.

Table 5.9: Longitudinal synthetic datasets as data augmentation. The results represent the mean detection TPF, FPF, DSC_d and mean segmentation DSC_s for training the SimLearnedDFs model with LongTrain set and with LongTrain plus each of the three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI, and Long-ISBI) as data augmentation. For each coefficient, the reported values are the mean ± standard deviation when evaluated on the LongTest set. The automatic segmentation masks were obtained by thresholding the probability maps with 0.5 (using argmax), and all automatic lesions with a size lower than three voxels were removed. Significant results are shown in bold ($p < 0.05$).

Training dataset	TPF	FPF	DSC_d	DSC_s
LongTrain	52.09 ± 34.72	12.62 ± 22.13	0.57 ± 0.35	0.39 ± 0.28
LongTrain +				
(Long-Clinical)	**73.32 ± 30.56**	10.72 ± 18.09	**0.75 ± 0.27**	**0.50 ± 0.24**
(Long-MICCAI)	**65.60 ± 34.64**	10.10 ± 15.64	**0.68 ± 0.31**	**0.46 ± 0.28**
(Long-ISBI)	**60.51 ± 34.12**	10.83 ± 20.43	**0.65 ± 0.32**	0.43 ± 0.27

Results

Table 5.9 summarizes the new T2-w lesion detection and segmentation mean results for training the SimLearnedDFs with the three synthetic longitudinal datasets (Long-Clinical, Long-MICCAI, and Long-ISBI) as data augmentation.

Discussion

Regarding the data augmentation experiments, we showed the effect of data augmentation on the longitudinal MS lesion detection performance when adding the longitudinal synthetic datasets to the training images. In terms of TPF, adding the longitudinal synthetic datasets was significantly better ($p < 0.05$). In terms of FPF, the results were not significantly better (about 2% improvement).

5.6 Discussion

We proposed a synthetic MS lesion generator pipeline that generates synthetic images with MS lesions. The use of the intensity level masks enabled us to train the model without the need of ground truth. Furthermore, the intensity level masks help the MS lesion generator to preserve the intensity gradients inside the synthetic MS lesion. The proposed pipeline was used to improve the cross-sectional and longitudinal MS lesion approaches. In the cross-sectional analysis, the pipeline was used to generate MS lesions on T1-w and FLAIR images using only two encoders and two decoders, while in the longitudinal analysis, it was extended to generate MS lesions on T1-w, T2-w, PD-w, and FLAIR images through the addition of two other encoder/decoder.

In conclusion, the obtained results indicate that the proposed pipeline is able to generate useful synthetic images with MS lesions that do not differ from real images. Also, the combination of the synthetic MS lesions generated on healthy images and original patient images from the same domain increases the segmentation and detection accuracy of MS lesions. Moreover, the proposed pipeline is also able to generate synthetic MS lesions on the follow-up images of longitudinal images with no new MS lesions. These images could be used for training or for increasing the performance of new MS lesion detection approaches.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Summary and contributions of the thesis

The aim of this PhD thesis has been the proposal of novel and fully automated methods for the detection of new T2-w lesions in MR images of MS patients. After reviewing the state-of-the-art of the new T2-w MS lesions detection, we observed the importance of using prior knowledge. Prior knowledge is important to guide the lesion detection and segmentation. Supervised approaches that rely on similar segmented cases usually outperform unsupervised strategies. Moreover, we noticed the effect of the tissue transformation and the mass effect for new T2-w lesion detection. We realized that deformation field-based algorithms have allowed the mass effect of the lesions to be considered. Therefore, we analyzed a basic supervised machine learning approach based on DF as a starting point for new MS lesions detection in longitudinal analysis. Following the same objectives defined in the Introduction, in what follows we summarize the main conclusions and contributions of this PhD thesis:

- We proposed and evaluated a novel fully automated supervised framework with intensity subtraction and deformation field for the detection of new T2-w lesions. For each modality (T1-w, T2-w, PD-w, and FLAIR), an affine transformation from baseline to follow-up was computed and the images were subtracted. The DF were obtained using the multi-resolution Demons registration approach from ITK v.4. The DF information was incorporated as features, in particular, we computed three DF operators (Jacobian, Divergence, and NormDiv) at each voxel. Then, a voxel-level logistic regression classifier was trained to predict the lesion probability of each voxel using the baseline and follow-up intensities, subtraction values, and the DF operators for T1-w, T2-w, PD-w, and FLAIR images. As a post-processing, the probab-

ilistic maps were thresholded to obtain a binary segmentation where all lesions smaller than three voxels were removed. We evaluated the performance of the model following a leave-one-out cross-validation scheme using an in-house clinical dataset (60 different patients: 36 of the patients confirmed MS with new T2-w lesions, while 24 patients did not present new T2-w lesions) from our collaborators. The obtained results were compared with those of recent state-of-the-art approaches. The performance of our model was significantly higher than state-of-the-art methods. Moreover, we studied the impact of both the deformation field operators and the baseline intensities features in the detection and segmentation of new T2-w lesions by analyzing the different models (LR-NDFNB, LR-NDF, LR-DFNB, LR-DF). We also studied the performance of these models and the state-of-the-art methods according to the different lesion sizes. In conclusion, the obtained results indicate that the combination of DFs and supervised classification increases the accuracy when detecting new T2-w lesions. We released a public version of the proposed method that can be downloaded for free from our research team web page [1]. This software is already being used in the collaborating hospitals. Furthermore, this entire analysis has been published in the following paper:

Paper published in the **NeuroImage: Clinical**
Title: A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis.
Volume: 17, Pages: 607-615, Published: November, 2017.
DOI: 10.1016/j.nicl.2017.11.015
[JCR N IF 3.943, Q1(3/14)]

- We proposed and evaluated a new deep learning based model to detect new T2-w lesions in longitudinal brain MR images. The aim of this approach was to eliminate the feature extraction step that was needed to extract important features from input images before the training process. The method was also based on DF features. The DFs were not computed using classic registration approaches that establish a dense nonlinear correspondence between a pair of 3D brain scans but using a learning-based method that learns a parametrized registration function from a collection of images during training. The proposed network was an FCNN that takes four image modalities (T1-w, T2-w, PD-w, and FLAIR) in both baseline and follow-up as inputs and outputs the new T2-w lesion segmentation. The network consisted of two parts. The first part of the network consisted of U-Net blocks that learned the deformation fields (DFs) and nonlinearly registered the baseline image to the follow-up image for each input modality. The second part of the network was another U-Net that performed the final detection and segments the new T2-w lesions. The DFs and the new T2-w lesions were learned simultaneously using a combined loss

[1] https://github.com/NIC-VICOROB/LR-T2-w-Lesions

function. The loss function used in this work was the summation of two loss functions. The first function was an unsupervised loss function that controls the registration part of the network and the second function was a supervised loss function CrossEntropy that controls the segmentation part of the network and penalizes differences between the segmentation and ground truth. We evaluated the performance of the model following a leave-one-out cross-validation scheme using the in-house clinical dataset (60 different patients: 36 of the patients confirmed MS with new T2-w lesions, while 24 patients did not present new T2-w lesions) from our collaborators. The performance of our model was significantly better compared to the state-of-the-art methods. Similarly to the evaluation of the LR-based model proposed, we also studied the performance of the model according to the different lesion sizes. Moreover, we demonstrated the contribution of simultaneously learning both the DF and the segmentation of new T2-w lesions by analyzing three other models (SepLearnedDFs, DemonsDFs, and NDFs). As the MRI criteria for dissemination in space consider the lesion type and location, we also studied the performance of the proposed model, the three variants (SepLearnedDFs, DemonsDFs, NDFs), and the state-of-the-art approaches on different brain regions (periventricular, juxtacortical, infratentorial, and deep white matter). We also analyzed the generalization and the performance of the proposed approach when tested in images from a different scanner and image acquisition protocol, we performed a new experiment with data from another collaborating Hospital. In conclusion, the obtained results indicate that the proposed end-to-end training model increases the accuracy of the new T2-w lesion detection. The results also indicate that the DL based model is better that the proposed LR-based model. Given the sensitivity and limited number of false positives, we strongly believe that the proposed method may be used in clinical studies in order to monitor the progression of the disease. This software is going to be used in the collaborating hospitals. Furthermore, this entire analysis has been submitted in the following manuscript:

Paper published in the **NeuroImage: Clinical**
Title: A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis.
Volume: 25, 102149, Published: December, 2019.
DOI: 10.1016/j.nicl.2019.102149
[JCR N IF 3.943, Q1(3/14)]

- Finally, we proposed and evaluated a novel deep learning based approach model for MS lesion synthesis on MR images with the final aim to improve the performance of supervised machine learning algorithms, therefore avoiding the problem of the lack of available ground truth. We proposed a FCNN model for MS lesion synthesis in MR images. The lesion information was encoded as discrete binary intensity level masks passed to the model and stacked with

the input images. The model was trained end-to-end without the need for manually annotating the lesions in the training set. We then performed the generation of synthetic lesions on healthy images via registration of patient images, which were subsequently used for data augmentation to increase the performance for supervised MS lesion detection algorithms. The proposed model was evaluated on improving the cross-sectional and longitudinal MS lesion detection and segmentation approaches. Regarding the cross-sectional evaluation, our pipeline was evaluated on MS patient data from an in-house clinical dataset and the public ISBI2015 challenge dataset. The evaluation was based on measuring the similarities between the real and the synthetic images as well as in terms of lesion detection performance by segmenting both the original and synthetic images individually using a state-of-the-art segmentation framework. We also demonstrated the usage of synthetic MS lesions generated on healthy images as data augmentation. We analyzed a scenario of limited training data (one-image training) to demonstrate the effect of the data augmentation on both datasets. Our results significantly showed the effectiveness of the usage of synthetic MS lesion images. For the ISBI2015 challenge, our one-image model trained using only a single image plus the synthetic data augmentation strategy showed a performance similar to that of other CNN methods that were fully trained using the entire training set, yielding a comparable human expert rater performance. Regarding the longitudinal evaluation, MS lesions were generated on only the follow-up scans. The follow-up image with the synthetic lesions with the untouched baseline were used as data augmentation to increase the longitudinal MS lesion detection performance. We released a public version of the proposed method that can be downloaded for free from our research team web page [2]. Furthermore, this entire analysis has been published in the following paper:

Paper published in the **IEEE Access**
Title: Multiple Sclerosis Lesion Synthesis in MRI Using an Encoder-Decoder U-NET.
Volume: 7, Pages: 25171-25184, Published: February, 2019.
DOI: 10.1109/ACCESS.2019.2900198
[JCR CSIS IF 4.098, Q1(23/155)]

6.2 Future work

The analysis of brain MR images for MS patients is a complex topic involving several aspects and multiple research lines. This notion is exemplified in this PhD thesis by the several steps that are involved in the new T2-w MS lesion detection process. Furthermore, some of the concepts applied to MS patients can be applied to other

[2]https://github.com/NIC-VICOROB/MS_Lesions_Generator

MRI fields or can be studied further. Besides this, other interesting topics arise from the needs of current clinical practice for MS patients.

Hence, future directions are presented in two categories: those related to improve our proposal, and long term future research lines departing from this thesis.

6.2.1 Short-term proposal improvements

In this PhD thesis, we presented two supervised methods for the detection of the new T2-w MS lesions. One was based on the conventional machine learning techniques and the other was based on deep learning. The two methods were validated on in-house clinical dataset from our collaborating hospitals. However, in chapter 4 we analyzed the generalization and the performance of the proposed approach when tested in images from a different scanner and image acquisition protocol. Our future work is to increase the longitudinal database to validate the two proposals again for various scenarios like training and testing on data from other scanners.

As seen in chapter 4, the sensitivity of CNN methods was remarkably lower in the infratentorial region due to a lack of training data. A short-term improvement could be increasing the sensitivity of these methods in the infratentorial region by obtaining more samples containing infratentorial lesions or by generating infratentorial synthetic lesions using our MS generator with the aim to improve the performance of the proposed methods at that location. We believe that more training data or the use of synthetic MS data may increase the sensitivity of the methods while reducing FP lesions. As Generative adversarial networks (GANs) have been used widely to provide anatomically-plausible and diverse samples for augmentation and other applications [272, 273], another short-term work is to use GANs to generate MR images with synthetic MS lesions and to compare this GAN-based model with our Encoder-Decoder-based MS lesion generator model proposed in chapter 5.

Another short-term improvement could be to build an automated pipeline that allows complete control on lesions quantities, volumes, and location. One way might be to build a lesion database which contains for each lesion the information of GT, 8 intensity level mask, size, and location. The automated pipeline could generate random lesions at specified regions with specified lesion load. Furthermore, an active shape model could be used to learn the shape of the lesions at each region. The model could be used to generate variability in the lesions. This model could be integrated to our MS lesion generator to be able to automatically generate synthetic lesions at specific regions.

6.2.2 Future research lines

In the long term, there are several new research lines departing from this PhD thesis that could be studied in our research group. As already seen during the development of this thesis, there is a lack of public longitudinal databases of MS patients with both manual lesion annotations in baseline and the different time points on

which automatic segmentation algorithms can be trained and tested. Therefore, the construction of a standardized and publicly available dataset of patients with different diseases, including MS, with reliable annotations of lesions, would not only be of special interest for training and testing all the methods presented in this thesis, but also very useful to the scientific community. Furthermore, with the increasing success of the deep learning approaches for medical image analysis, which need large amounts of data for their training and testing, the creation of the mentioned database would be an incentive for new method proposals able to achieve more accurate results.

Moreover, the methods and concepts presented here could also be applied to the study of other diseases with similar properties. A deeper study involving different diseases and lesion properties would be interesting. Patients with lupus, with stroke, or with WM hyperintensities may be interesting to study.

Finally, the ultimate future goal should be to provide state-of-the-art tools for the collaborating hospitals involved in these research projects that may be useful not only to diagnose and monitor the progression of this disease, but also to evaluate new treatments for MS patients. Related to that, the tools developed in this thesis should be integrated with other tools developed in our group in order to implement this complete system capable of providing robust and useful biomarkers in MS such as the number of lesions, lesion volume, brain tissue volume or brain atrophy.

BIBLIOGRAPHY

[1] P. Brodal. *The Central Nervous System*. Oxford University Press, 2010.

[2] G. Sperber. Clinically oriented anatomy. *Journal of anatomy*, 208(3):393, 2006.

[3] C. Confavreux, G. Aimard, and M. Devic. Course and prognosis of multiple sclerosis assessed by the computerized data processing of 349 patients. *Brain: A journal of neurology*, 103(2):281–300, 1980.

[4] MSIF. Atlas of MS 2013: Mapping Multiple Sclerosis Around the World. *Multiple Sclerosis International Federation*, page 1–28, 2013.

[5] S. Love. Demyelinating diseases. *Journal of clinical pathology*, 59(11):1151–1159, 2006.

[6] A. Compston and A. Coles. Multiple sclerosis. *The Lancet*, 372(9648):1502 – 1517, 2008.

[7] C. Confavreux, S. Vukusic, T. Moreau, and P. Adeleine. Relapses and progression of disability in multiple sclerosis. *New England Journal of Medicine*, 343(20):1430–1438, 2000.

[8] W. Brownlee, T. Hardy, F. Fazekas, and D. Miller. Diagnosis of multiple sclerosis: Progress and challenges. *The Lancet*, 389(10076):1336–1346, 2017.

[9] T. Geva. Magnetic resonance imaging: Historical perspective. *Journal of Cardiovascular Magnetic Resonance*, 8(4):573–580, 2006.

[10] R. Edelman and S. Warach. Magnetic Resonance Imaging. *New England Journal of Medicine*, 328(10):708–716, 1993.

[11] À. Rovira, M. Wattjes, M. Tintoré, C. Tur, T. Yousry, M. Sormani, N. de Stefano, M. Filippi, C. Auger, M. Rocca, F. Barkhof, F. Fazekas, L. Kappos,

C. Polman, D. Miller, and X. Montalban. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. *Nature Reviews Neurology*, 11(August): 1–12, 2015.

[12] M. Filippi, M. Rocca, O. Ciccarelli, N. de Stefano, N. Evangelou, L. Kappos, À. Rovira, J. Sastre-Garriga, M. Tintoré, J. Frederiksen, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology*, 15(3):292 – 303, 2016.

[13] A. van der Kolk, J. Hendrikse, J. Zwanenburg, F. Visser, and P. Luijten. Clinical applications of 7T MRI in the brain. *European Journal of Radiology*, 82(5):708 – 718, 2013.

[14] R. Bitar, G. Leung, R. Perng, S. Tadros, A. Moody, J. Sarrazin, C. McGregor, M. Christakis, S. Symons, A. Nelson, et al. MR pulse sequences: What every radiologist wants to know but is afraid to ask. *Radiographics*, 26(2):513–537, 2006.

[15] C. Polman, S. Reingold, B. Banwell, M. Clanet, J. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Annals of Neurology*, 69(2):292–302, 2011.

[16] M. Rocca, N. Anzalone, A. Falini, and M. Filippi. Contribution of magnetic resonance imaging to the diagnosis and monitoring of multiple sclerosis. *La radiologia medica*, 118(2):251–264, Mar 2013.

[17] À. Rovira and A. León. MR in the diagnosis and monitoring of multiple sclerosis: An overview. *European Journal of Radiology*, 67(3):409 – 414, 2008.

[18] A. Ceccarelli, R. Bakshi, and M. Neema. MRI in multiple sclerosis: A review of the current literature. *Current opinion in neurology*, 25(4):402–409, 2012.

[19] W. McDonald, A. Compston, G. Edan, D. Goodkin, H. Hartung, F. Lublin, H. McFarland, D. Paty, C. Polman, S. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. van Den Noort, B. Weinshenker, and J. Wolinsky. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, 50(1):121–127, 2001.

[20] C. Polman, S. Reingold, G. Edan, M. Filippi, H. Hartung, L. Kappos, F. Lublin, L. Metz, H. McFarland, P. O'Connor, M. Sandberg-Wollheim, A. Thompson, B. Weinshenker, and J. Wolinsky. Diagnostic criteria for multiple sclerosis: 2005 revisions to the McDonald criteria. *Annals of Neurology*, 58(6):840–846, 2005.

[21] A. Thompson, B. Banwell, F. Barkhof, W. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. Freedman, K. Fujihara, S. Galetta, H. Hartung, L. Kappos, F. Lublin, R. Marrie, A. Miller, D. Miller, X. Montalban, E. Mowry, P. Sorensen, M. Tintoré, A. Traboulsee, M. Trojano, B. Uitdehaag, S. Vukusic, E. Waubant, B. Weinshenker, S. Reingold, and J. Cohen. Diagnosis of multiple sclerosis: 2017 revisions of the Mcdonald criteria. *The Lancet Neurology*, 17(2):162 – 173, 2018.

[22] E. Roura, A. Oliver, M. Cabezas, J. Vilanova, À. Rovira, L. Ramió-Torrentà, and X. Lladó. MARGA: Multispectral adaptive region growing algorithm for brain extraction on axial MRI. *Computer Methods and Programs in Biomedicine*, 113(2):655 – 673, 2014.

[23] S. Valverde, A. Oliver, and X. Lladó. A white matter lesion-filling approach to improve brain tissue volume measurements. *NeuroImage: Clinical*, 6:86–92, 2014.

[24] M. Cabezas, A. Oliver, E. Roura, J. Freixenet, J. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding. *Computer Methods and Programs in Biomedicine*, 115(3):147 – 161, 2014.

[25] M. Cabezas, A. Oliver, S. Valverde, B. Beltran, J. Freixenet, J. Vilanova, L. Ramió-Torrentà, À Rovira, and X Lladó. BOOST: A supervised approach for multiple sclerosis lesion segmentation. *Journal of Neuroscience Methods*, 237:108–117, 2014.

[26] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira. Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. *Information Sciences*, 186 (1):164–185, March 2012.

[27] E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. A toolbox for multiple sclerosis lesion segmentation. *Neuroradiology*, 57(10):1031–1043, 2015.

[28] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159 – 168, 2017.

[29] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21:101638, 2019.

[30] M. Salem, S. Valverde, M. Cabezas, D. Pareto, A. Oliver, J. Salvi, À. Rovira, and X. Lladó. Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET. *IEEE Access*, 7:25171–25184, 2019.

[31] X. Lladó, O. Ganiler, A. Oliver, R. Martí, J. Freixenet, L. Valls, J. Vilanova, L. Ramió-Torrentà, and À. Rovira. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*, 54(8):787–807, 2012.

[32] O. Ganiler, A. Oliver, Y. Díez, J. Freixenet, J. Vilanova, B. Beltran, L. Ramió-Torrentà, À. Rovira, and X. Lladó. A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology*, 56(5):363–374, 2014.

[33] M. Cabezas, J. Corral, A. Oliver, Y. Díez, M. Tintoré, C. Auger, X. Montalban, X. Lladó, D. Pareto, and À. Rovira. Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. *American Journal of Neuroradiology*, 37(10):1816–1823, 2016.

[34] M. Salem, M. Cabezas, S. Valverde, D. Pareto, A. Oliver, J. Salvi, À. Rovira, and X. Lladó. A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage: Clinical*, 17:607 – 615, 2018.

[35] Y. Díez, A. Oliver, M. Cabezas, S. Valverde, R. Martí, J. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. Intensity based methods for brain MRI longitudinal registration. A study on multiple sclerosis patients. *Neuroinformatics*, 12(3):365–379, 2014.

[36] E. Roura, T. Schneider, M. Modat, P. Daga, N. Muhlert, D. Chard, S. Ourselin, X. Lladó, and C. Wheeler-Kingshott. Multi-channel registration of FA and T1-w images in the presence of atrophy: Application to multiple sclerosis. *Functional Neurology*, 30(4), 2015.

[37] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. Bach-Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, 104(3):e158–e177, 2011.

[38] S. Valverde, A. Oliver, E. Roura, D. Pareto, J. Vilanova, L. Ramió-Torrentà, J.Sastre-Garriga, X. Montalban, À. Rovira, and X. Lladó. Quantifying brain tissue volume in multiple sclerosis with automated lesion segmentation and filling. *NeuroImage: Clinical*, 9:640 – 647, 2015.

[39] S. Valverde, A. Oliver, E. Roura, S. González-Villà, D. Pareto, J. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. Automated tissue segmentation of MR brain images in the presence of white matter lesions. *Medical Image Analysis*, 35:446 – 457, 2017.

[40] J. Bernal, K. Kushibar, M. Cabezas, S. Valverde, A. Oliver, and X. Lladó. Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access*, 7: 89986–90002, 2019.

[41] A. Elster. Gradient-echo MR imaging: Techniques and acronyms. *Radiology*, 186(1):1–8, 1993.

[42] D. Miller, R. Grossman, S. Reingold, and H. McFarland. The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain*, 121(1):3–24, 01 1998.

[43] M. Sahraian and A. Eshaghi. Role of MRI in diagnosis and treatment of multiple sclerosis. *Clinical Neurology and Neurosurgery*, 112(7):609 – 615, 2010.

[44] I. Kilsdonk, L. Jonkman, R. Klaver, S. van Veluw, J. Zwanenburg, J. Kuijer, P. Pouwels, J. Twisk, M. Wattjes, P. Luijten, F. Barkhof, and J. Geurts. Increased cortical grey matter lesion detection in multiple sclerosis with 7T MRI: A post-mortem verification study. *Brain*, 139(5):1472–1481, 03 2016.

[45] P. Molyneux, D. Miller, M. Filippi, T. Yousry, E. Radü, H. Ader, and F. Barkhof. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. *Neuroradiology*, 41(12):882–888, 1999.

[46] H. Cline, W. Lorensen, R. Kikinis, and F. Jolesz. Three-dimensional segmentation of MR images of the head using probability and connectivity. *Journal of computer assisted tomography*, 14(6):1037—1045, 1990.

[47] G. Gerig, J. Martin, R. Kikinis, O. Kubler, M. Shenton, and F. Jolesz. Unsupervised tissue type segmentation of 3D dual-echo MR head data. *Image and Vision Computing*, 10(6):349 – 360, 1992. Information Processing in Medical Imaging.

[48] T. Kapur, W. Grimson, W. Wells, and R. Kikinis. Segmentation of brain tissue from magnetic resonance images. *Medical Image Analysis*, 1(2):109 – 127, 1996.

[49] D. Sha and J. Sutton. Towards automated enhancement, segmentation and classification of digital brain images using networks of networks. *Information Sciences*, 138(1):45 – 77, 2001.

[50] J. Iglesias, C. Liu, P. Thompson, and Z. Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, 2011.

[51] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, and J. Gee. N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010.

[52] J. Acosta-Cabronero, G. Williams, J. Pereira, G. Pengas, and P. Nestor. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. *NeuroImage*, 39(4):1654 – 1665, 2008.

[53] J. Lee, U. Yoon, S. Nam, J. Kim, I. Kim, and S. Kim. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. *Computers in Biology and Medicine*, 33(6):495 – 507, 2003.

[54] S. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17 (3):143–155, 2002.

[55] D. Shattuck, S. Sandor-Leahy, K. Schaper, D. Rottenberg, and R. Leahy. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856 – 876, 2001.

[56] S. Eskildsen, P. Coupé, V. Fonov, J. Manjón, K. Leung, N. Guizard, S. Wassef, L. Østergaard, and D. Collins. BEaST: Brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362 – 2373, 2012.

[57] U. Vovk, F. Pernus, and B. Likar. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26 (3):405–421, March 2007.

[58] J. Arnold, J. Liow, K. Schaper, J. Stern, J. Sled, D. Shattuck, A. Worth, M. Cohen, R. Leahy, J. Mazziotta, and D. Rottenberg. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *NeuroImage*, 13(5):931 – 943, 2001.

[59] Z. Hou. A review on MR image intensity inhomogeneity correction. *International journal of biomedical imaging*, 2006, 2006.

[60] J. Sled, A. Zijdenbos, and A. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, Feb 1998.

[61] L. Nyúl, J. Udupa, and X. Zhang. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000.

[62] E. Roura, T. Schneider, M. Modat, P. Daga, N. Muhlert, D. Chard, S. Ourselin, X. Lladó, and C. Wheeler-Kingshott. Multi-channel registration of fractional anisotropy and T1-weighted images in the presence of atrophy: Application to multiple sclerosis. *Functional neurology*, 30(4):245, 2015.

[63] J. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: A survey. *IEEE Transactions on Medical Imaging*, 22(8): 986–1004, Aug 2003.

[64] L. Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376, December 1992.

[65] M. Holden. A review of geometric transformations for nonrigid body registration. *IEEE Transactions on Medical Imaging*, 27(1):111–128, Jan 2008.

[66] R. Rabbitt, J. Weiss, G. Christensen, and M. Miller. Mapping of hyperelastic deformable templates using the finite element method. In *Vision Geometry IV*, volume 2573, pages 252–265. International Society for Optics and Photonics, 1995.

[67] G. Christensen. Deformable shape models for anatomy. *Ph.D. dissertation, The University of Iowa*, 1994.

[68] G. Christensen, R. Rabbitt, M. Miller, et al. Deformable templates using large deformation kinematics. *IEEE transactions on image processing*, 5(10): 1435–1447, 1996.

[69] J.-P. Thirion. Non-rigid matching using demons. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 245–251, June 1996.

[70] J.-P. Thirion. Image matching as a diffusion process: An analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243 – 260, 1998.

[71] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[72] D. Chard, G. Parker, R. Kapoor, A. Thompson, and D. Miller. Brain atrophy in clinically early relapsing-remitting multiple sclerosis. *Brain*, 125:327–337, 2002.

[73] M. Battaglini, M. Jenkinson, and N. De Stefano. Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Human Brain Mapping*, 33(9):2062–2071, 2012.

[74] D. Chard, J. Jackson, D. Miller, and C. Wheeler-Kingshott. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. *Journal of Magnetic Resonance Imaging*, 32(1):223–228, 2010.

[75] S. Magon, L. Gaetano, M. Chakravarty, J. Lerch, Y. Naegelin, C. Stippich, L. Kappos, E. Radue, and T. Sprenger. White matter lesion filling improves the accuracy of cortical thickness measurements in multiple sclerosis patients: A longitudinal study. *BMC Neuroscience.*, 15(1):106, January 2014.

[76] M. Sdika and D. Pelletier. Nonrigid registration of multiple sclerosis brain images using lesion inpainting for morphometry or lesion mapping. *Human Brain Mapping*, 30(4):1060–1067, 2009.

[77] V. Popescu, N. Ran, F. Barkhof, D. Chard, C. Wheeler-Kingshott, and H. Vrenken. Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks. *NeuroImage: Clinical*, 4:366–373, 2014.

[78] A. Ceccarelli, J. Jackson, S. Tauhid, A. Arora, J. Gorky, E. Dell'Oglio, A. Bakshi, T. Chitnis, S. Khoury, H. Weiner, C. Guttmann, R. Bakshi, and M. Neema. The impact of lesion in-painting and registration methods on voxel-based morphometry in detecting regional cerebral gray matter atrophy in multiple sclerosis. *American Journal of Neuroradiology*, 33(8):1579–85, September 2012.

[79] M. Steenwijk, J. Geurts, M. Daams, B. Tijms, A. Wink, L. Balk, P. Tewarie, B. Uitdehaag, F. Barkhof, H. Vrenken, and P. Pouwels. Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant. *Brain*, 139(1):115–126, 12 2015.

[80] C. Xu, D. Pham, M. Rettmann, D. Yu, and J. Prince. Reconstruction of the human cerebral cortex from magnetic resonance images. *IEEE Transactions on Medical Imaging*, 18(6):467–480, 1999.

[81] D. MacDonald, N. Kabani, D. Avis, and A. Evans. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. *NeuroImage*, 12(3): 340–356, 2000.

[82] J. de Bresser, M. Portegies, A. Leemans, G. Biessels, L. Kappelle, and M. Viergever. A comparison of MR based segmentation methods for measuring brain atrophy progression. *Neuroimage*, 54(2):760–768, 2011.

[83] N. Kovacevic, N. Lobaugh, M. Bronskill, B. Levine, A. Feinstein, and S. Black. A robust method for extraction and automatic segmentation of brain images. *Neuroimage*, 17(3):1087–1100, 2002.

[84] B. Cherradi, O. Bouattane, M. Youssfi, and A. Raihani. Brain extraction and fuzzy tissue segmentation in cerebral 2D T1-weigthed magnetic resonance images. *International Journal of Computer Science Issues*, 8(3):215–223, 2011.

[85] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. Schmid, C. Zimmer, B. Hemmer, and M. Mühlau. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, 59(4):3774 – 3783, 2012.

[86] M. Filippi, P. Preziosa, M. Copetti, G. Riccitelli, M. Horsfield, V. Martinelli, G. Comi, and M. Rocca. Gray matter damage predicts the accumulation of disability 13 years later in MS. *Neurology*, 81(20):1759–1767, 2013.

[87] E. Fisher, J. Lee, K. Nakamura, and R. Rudick. Gray matter atrophy in multiple sclerosis: A longitudinal study. *Annals of Neurology*, 64(3):255–265, 2008.

[88] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, Jan 2001.

[89] J. Ashburner and K. Friston. Unified segmentation. *NeuroImage*, 26(3):839 – 851, 2005.

[90] K. Pohl, J. Fisher, W. Grimson, R. Kikinis, and W. Wells. A bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228 – 239, 2006.

[91] S. Roy, A. Carass, P. Bazin, S. Resnick, and J. Prince. Consistent segmentation using a rician classifier. *Medical Image Analysis*, 16(2):524 – 535, 2012.

[92] B. Caldairou, N. Passat, P. Habas, C. Studholme, and F. Rousseau. A non-local fuzzy segmentation method: Application to brain MRI. *Pattern Recognition*, 44(9):1916 – 1927, 2011. Computer Analysis of Images and Patterns.

[93] H. Vrooman, F. van der Lijn, and W. Niessen. Auto-kNN: brain tissue segmentation using automatically trained knearest-neighbor classification. In *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS'13)*, 2013.

[94] A. van Opbroek, F. van der Lijn, and M. de Bruijne. Automated brain-tissue segmentation by multi-feature SVM classification. In *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS'13)*, 2013.

[95] M. Rajchl, J. Baxter, J. McLeod, J. Yuan, W. Qiu, T. Peters, and A. Khan. Hierarchical max-flow segmentation framework for multi-atlas segmentation with kohonen self-organizing map based gaussian mixture modeling. *Medical image analysis*, 27:45–56, 2016.

[96] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015.

[97] P. Moeskops, M. Viergever, A. Mendrik, L. de Vries, M. Benders, and I. Išgum. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261, May 2016.

[98] H. Chen, Q. Dou, L. Yu, and P. Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, 2016.

[99] S. Valverde, A. Oliver, Y. Díez, M. Cabezas, J. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. Evaluating the effects of white matter multiple sclerosis lesions on the volume estimation of 6 brain tissue segmentation methods. *American Journal of Neuroradiology*, 36(6):1109–1115, 2015.

[100] V. Popescu, N. Ran, F. Barkhof, D. Chard, C. Wheeler-Kingshott, and H. Vrenken. Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks. *NeuroImage: Clinical*, 4:366–373, 2014.

[101] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira. Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. *Information Sciences*, 186 (1):164 – 185, 2012.

[102] M. Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J-P Thiran. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *IEEE transactions on medical imaging*, 24 (12):1548–1565, 2005.

[103] R. Khayati, M. Vafadust, F. Towhidkhah, and M. Nabavi. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. *Computers in biology and medicine*, 38(3):379–390, 2008.

[104] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE transactions on medical imaging*, 18(10):897–908, 1999.

[105] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[106] J. Udupa, L. Nyul, Y. Ge, and R. Grossman. Multiprotocol MR image segmentation in multiple sclerosis: Experience with over 1,000 studies. *Academic Radiology*, 8(11):1116–1126, 2001.

[107] D. Kroon, E. van Oort, and C. Slump. Multiple sclerosis detection in multispectral magnetic resonance images with principal components analysis. In *Proceedings of the MICCAI 2008 Workshop on MS Lesion Segmentation*, pages 1–14, 2008.

[108] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, P. Suetens, et al. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE transactions on medical imaging*, 20(8):677–688, 2001.

[109] Y. Wu, S. Warfield, I. Tan, W. Wells III, D. Meier, R. van Schijndel, F. Barkhof, and C. Guttmann. Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. *NeuroImage*, 32(3):1205–1215, 2006.

[110] A. Zijdenbos, R. Forghani, A. Evans, et al. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. *IEEE Trans. Med. Imaging*, 21(10):1280–1291, 2002.

[111] M. Prastawa and G. Gerig. Automatic MS lesion segmentation by outlier detection and information theoretic region partitioning. *Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge*, pages 1–8, 2008.

[112] A. Akselrod-Ballin, M. Galun, J. Gomori, M. Filippi, P. Valsasina, R. Basri, and A. Brandt. Automatic segmentation and classification of multiple sclerosis in multichannel MRI. *IEEE Transactions on biomedical engineering*, 56(10): 2461–2469, 2009.

[113] N. Shiee, P. Bazin, A. Ozturk, D. Reich, P. Calabresi, and D. Pham. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535, 2010.

[114] M. Kamber, R. Shinghal, D. Collins, G. Francis, and A. Evans. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. *IEEE Transactions on Medical Imaging*, 14(3):442–453, 1995.

[115] D. Goldberg-Zimring, A. Achiron, S. Miron, M. Faibel, and H. Azhari. Automated detection and characterization of multiple sclerosis lesions in brain MR images. *Magnetic resonance imaging*, 16(3):311–318, 1998.

[116] P. Anbeek, K. Vincken, G. Van Bochove, M. Van Osch, and J. van der Grond. Probabilistic segmentation of brain tissue in MR imaging. *Neuroimage*, 27(4): 795–804, 2005.

[117] S. Datta, B. Sajja, R. He, J. Wolinsky, R. Gupta, and P. Narayana. Segmentation and quantification of black holes in multiple sclerosis. *Neuroimage*, 29 (2):467–474, 2006.

[118] J. Lecoeur, J. Ferr, and C. Barillot. Optimized supervised segmentation of MS lesions from multispectral MRIs. *Image Anal. on Multiple Sclerosis*, pages 5–14, 2009.

[119] N. Subbanna, M. Shah, S. Francis, S. Narayanan, D. Collins, D. Arnold, and T. Arbel. MS lesion segmentation using Markov random fields. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, London, UK*, pages 15–26, 2009.

[120] N. Guizard, P. Coupé, V. Fonov, J. Manjón, D. Arnold, and D. Collins. Rotation-invariant multi-contrast non-local means for MS lesion segmentation. *NeuroImage: Clinical*, 8:376–389, 2015.

[121] H. Deshpande, P. Maurel, and C. Barillot. Classification of multiple sclerosis lesions using adaptive dictionary learning. *Computerized Medical Imaging and Graphics*, 46:2–10, 2015.

[122] O. Freifeld, H. Greenspan, and J. Goldberger. Lesion detection in noisy MR brain images using constrained GMM and active contours. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 596–599. IEEE, 2007.

[123] D. García-Lorenzo, S. Prima, D. Collins, D. Arnold, S. Morrissey, and C. Barillot. Combining robust expectation maximization and mean shift algorithms for multiple sclerosis brain segmentation. In *MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)(MIAMS'2008)*, pages 82–91, 2008.

[124] D. García-Lorenzo, S. Prima, L. Parkes, J. Ferré, S. Morrissey, and C. Barillot. The impact of processing workflow in performance of automatic white matter lesion segmentation in multiple sclerosis. In *MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)(MIAMS'2008)*, pages 104–112, 2008.

[125] D. Garcia-Lorenzo, S. Prima, S. Morrissey, and C. Barillot. A robust expectation-maximization algorithm for multiple sclerosis lesion segmentation. In *MICCAI Workshop: 3D Segmentation in the Clinic: A Grand Challenge II, MS Lesion Segmentation*, page 277, 2008.

[126] B. Bedell and P. Narayana. Automatic segmentation of gadolinium-enhanced multiple sclerosis lesions. *Magnetic resonance in medicine*, 39(6):935–940, 1998.

[127] A. Boudraa, S. Dehak, Y. Zhu, C. Pachai, Y. Bao, and J. Grimaud. Automated segmentation of multiple sclerosis lesions in multispectral MR imaging using fuzzy clustering. *Computers in biology and medicine*, 30(1):23–40, 2000.

[128] S. Datta, B. Sajja, R. He, R. Gupta, J. Wolinsky, and P. Narayana. Segmentation of gadolinium-enhanced lesions on MRI in multiple sclerosis. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 25(5):932–937, 2007.

[129] S. Saha and S. Bandyopadhyay. A new point symmetry based fuzzy genetic clustering technique for automatic evolution of clusters. *Information Sciences*, 179(19):3230–3246, 2009.

[130] X. Tomas-Fernandez and S. Warfield. A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 34(6):1349–1361, 2015.

[131] R. Harmouche, N. Subbanna, D. Collins, D. Arnold, and T. Arbel. Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood information. *IEEE transactions on Biomedical Engineering*, 62(5):1281–1292, 2014.

[132] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi. Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[133] M. Ghafoorian and B. Platel. Convolutional neural networks for MS lesion segmentation, method description of diag team. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[134] A. Birenbaum and H. Greenspan. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. *2nd International Workshop on Deep Learning in Medical Image Analysis, DLMIA 2016*, pages 58–67, 2016.

[135] S. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017.

[136] S. Roy, J. Butman, D. Reich, P. Calabresi, and D. Pham. Multiple sclerosis lesion segmentation from brain MRI via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172*, 2018.

[137] S. Aslani, M. Dayan, L. Storelli, M. Filippi, V. Murino, M. Rocca, and D. Sona. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196:1 – 15, 2019.

[138] À. Rovira, M. Wattjes, M. Tintore, C. Tur, T. Yousry, M. Sormani, N. De Stefano, M. Filippi, C. Auger, M. Rocca, et al. MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process (vol 11, pg 471, 2015). *NATURE REVIEWS NEUROLOGY*, 11(8), 2015.

[139] I. Pestalozza, C. Pozzilli, S. Di Legge, M. Piattella, P. Pantano, F. Caramia, P. Pasqualetti, and G. Lenzi. Monthly brain magnetic resonance imaging scans in patients with clinically isolated syndrome. *Multiple Sclerosis Journal*, 11 (4):390–394, 2005.

[140] J. Rio, J. Castillo, À. Rovira, M. Tintoré, J. Sastre-Garriga, A. Horga, C. Nos, M. Comabella, X. Aymerich, and X. Montalbán. Measures in the first year of therapy predict the response to interferon β in MS. *Multiple Sclerosis Journal*, 15(7):848–853, 2009.

[141] M. Sormani and N. de Stefano. Defining and scoring response to IFN-β in multiple sclerosis. *Nat Rev Neurol*, 9(9):504–512, 2013.

[142] M. Sormani, J. Rio, M. Tintoré, A. Signori, D. Li, P. Cornelisse, B. Stubinski, M. Stromillo, X. Montalban, and N. de Stefano. Scoring treatment response in patients with relapsing multiple sclerosis. *Multiple Sclerosis Journal*, 19(5): 605–612, 2013.

[143] L. Prosperini, C. Mancinelli, L. de Giglio, F. de Angelis, V. Barletta, and C. Pozzilli. Interferon beta failure predicted by EMA criteria or isolated MRI activity in multiple sclerosis. *Multiple Sclerosis Journal*, 20(5):566–576, 2014.

[144] M. Freedman, D. Selchen, D. Arnold, A. Prat, B. Banwell, M. Yeung, D. Morgenthau, Y. Lapierre, Canadian Multiple Sclerosis Working Group, et al. Treatment optimization in MS: Canadian MS working group updated recommendations. *Canadian Journal of Neurological Sciences*, 40(3):307–323, 2013.

[145] M. Stangel, I. Penner, B. Kallmann, C. Lukas, and B. Kieseier. Towards the implementation of 'no evidence of disease activity' in multiple sclerosis treatment: The multiple sclerosis decision model. *Ther Adv Neurol Disord*, 8 (1):3–13, 2015.

[146] E. Altay, E. Fisher, S. Jones, C. Hara-Cleaver, J. Lee, and R. Rudick. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurology*, 70(3):338–344, 2013.

[147] B. Moraal, M. Wattjes, J. Geurts, D. Knol, R. van Schijndel, P. Pouwels, H. Vrenken, and F. Barkhof. Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. *Radiology*, 255(1):154–163, 2010.

[148] B. Moraal, M. Wattjes, J. Geurts, D. Knol, R. van Schijndel, P. Pouwels, H. Vrenken, and F. Barkhof. Long-interval T2-weighted subtraction magnetic resonance imaging: A powerful new outcome measure in multiple sclerosis trials. *Annals of Neurology*, 67(5):667–675, 2010.

[149] P. Schmidt, V. Pongratz, P. Küster, D. Meier, J. Wuerfel, C. Lukas, B. Bellenberg, F. Zipp, S. Groppa, P. Sämann, F. Weber, C. Gaser, T. Franke, M. Bussas, J. Kirschke, C. Zimmer, B. Hemmer, and M. Mühlau. Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage: Clinical*, 23:101849, 2019.

[150] J-P Thirion and G. Calmon. Deformation analysis to detect and quantify active lesions in three-dimensional medical image sequences. *IEEE Transactions on Medical Imaging*, 18(5):429–441, 1999.

[151] B. Moraal, D. Meier, P. Poppe, J. Geurts, H. Vrenken, W. Jonker, D. Knol, R. van Schijndel, P. Pouwels, C. Pohl, et al. Subtraction MR images in a multiple sclerosis multicenter clinical trial setting. *Radiology*, 250(2):506–514, 2009.

[152] C. Elliott, D. Arnold, D. Collins, and T. Arbel. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Transactions on Medical Imaging*, 32(8):1490–1503, 2013.

[153] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[154] R. Gentleman, W. Huber, and V. Carey. Supervised machine learning. In *Bioconductor Case Studies*, pages 121–136. Springer, 2008.

[155] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct): 2825–2830, 2011.

[156] E. Sweeney, R. Shinohara, C. Shea, D. Reich, and C. Crainiceanu. Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *American Journal of Neuroradiology*, 34(1): 68–73, 2013.

[157] D. Rey, G. Subsol, H. Delingette, and N. Ayache. Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis. *Medical Image Analysis*, 6(2):163 – 179, 2002.

[158] M. Fartaria, T. Kober, C. Granziera, and M. Cuadra. Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. *NeuroImage: Clinical*, 23:101938, 2019.

[159] B. Erickson, P. Korfiatis, Z. Akkus, and T. Kline. Machine learning for medical imaging. *RadioGraphics*, 37(2):505–515, 2017.

[160] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-78019-5.

[161] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986.

[162] G. Seber and A. Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[163] D. Hosmer Jr, S. Lemeshow, and R. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[164] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[165] D. Lowd and P. Domingos. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 529–536. ACM, 2005.

[166] C. Zhou and Y. Chen. Improving nearest neighbor classification with cam weighted distance. *Pattern Recognition*, 39(4):635–645, 2006.

[167] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[168] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[169] K. Krishna and N. Murty. Genetic K-means algorithm. *IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics*, 29(3):433–439, 1999.

[170] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):603–619, 2002.

[171] D. Dueck and B. Frey. Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[172] S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[173] D. Birant and A. Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data and Knowledge Engineering*, 60(1):208–221, 2007.

[174] S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[175] R. Chellappa and A. Jain. Markov random fields: Theory and application. *Boston: Academic Press, 1993, edited by Chellappa, Rama; Jain, Anil*, 1993.

[176] J. Bezdek, R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10(2-3):191–203, 1984.

[177] V. Sharma, S. Rai, and A. Dev. A comprehensive study of artificial neural networks. *International Journal of Advanced research in computer science and software engineering*, 2(10), 2012.

[178] E. Li. Artificial neural networks and their business applications. *Information and Management*, 27(5):303–313, 1994.

[179] D. Siganos and C. Stergiou. Neural networks, the human brain, and learning. *Imperial College London: Surveys and Presentations in Information Systems Engineering*, 1996.

[180] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[181] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. Van Der Laak, B. Van Ginneken, and C. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[182] G. Forslid, H. Wieslander, E. Bengtsson, C. Wählby, J. Hirsch, C. Stark, and S. Sadanandan. Deep convolutional neural networks for detecting cellular changes due to malignancy. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 82–89, Oct 2017.

[183] K. Suzuki. Overview of deep learning in medical imaging. *Radiological Physics and Technology*, 10(3):257–273, Sep 2017.

[184] V. Romanuke. Appropriate number and allocation of ReLUs in convolutional neural networks. *Naukovi Visti NTUU KPI*, (1):69–78, 2017.

[185] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[186] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[187] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[188] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[189] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[190] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[191] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015.

[192] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[193] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

[194] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[195] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.

[196] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688.

[197] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[198] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303 – 312, 2017.

[199] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uder, F. de Leeuw, E. Marchiori, B. van Ginneken, and B. Platel. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1414–1417, April 2016.

[200] J. Charbonnier, E. van Rikxoort, A. Setio, C. Schaefer-Prokop, B. van Ginneken, and F. Ciompi. Improving airway segmentation in computed tomography using leak detection with convolutional networks. *Medical Image Analysis*, 36:52 – 60, 2017.

[201] M. van Grinsven, B. van Ginneken, C. Hoyng, T. Theelen, and C. Sánchez. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, May 2016.

[202] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[203] W. Yang, Y. Chen, Y. Liu, L. Zhong, G. Qin, Z. Lu, Q. Feng, and W. Chen. Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. *Medical Image Analysis*, 35:421 – 433, 2017.

[204] J. Bernal, K. Kushibar, D. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review. *Artificial Intelligence in Medicine*, 95: 64 – 81, 2019.

[205] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, 129:460 – 469, 2016.

[206] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen. Estimating CT image from MRI data using 3D fully convolutional networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 170–178, Cham, 2016. Springer International Publishing.

[207] X. Yang, R. Kwitt, and M. Niethammer. Fast predictive image registration. In *Deep Learning and Data Labeling for Medical Applications*, pages 48–57, Cham, 2016. Springer International Publishing.

[208] Y. Guo, G. Wu, L. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen. Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 308–315. Springer, 2014.

[209] T. Brosch, L. Tang, Y. Yoo, D. Li, A. Traboulsee, and R. Tam. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

[210] K. Kamnitsas, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36: 61–78, 2017.

[211] H. Suk and D. Shen. Deep ensemble sparse regression network for alzheimer's disease diagnosis. In *International Workshop on Machine Learning in Medical Imaging*, pages 113–121. Springer, 2016.

[212] S. Sarraf and G. Tofighi. Classification of alzheimer's disease using fMRI data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*, 2016.

[213] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, 22(1):173–183, Jan 2018.

[214] M. Tintoré, À. Rovira, J. Río, S. Otero-Romero, G. Arrambide, C. Tur, M. Comabella, C. Nos, M. Arévalo, L. Negrotto, et al. Defining high, medium and low impact prognostic factors for developing multiple sclerosis. *Brain*, 138 (7):1863, 2015.

[215] H. Johnson, M. McCormick, and L. Ibanez. The ITK software guide book 2: Design and functionality fourth edition updated for ITK version 4.7. *Kitware, Inc.(January 2015)*, 2015.

[216] M. Bosc, F. Heitz, J. Armspach, I. Namer, D. Gounot, and L. Rumbach. Automatic change detection in multimodal serial MRI: Application to multiple sclerosis lesion evolution. *NeuroImage*, 20(2):643–656, 2003.

[217] J-P Thirion. Image matching as a diffusion process: An analogy with Maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[218] M. Bro-Nielsen. Medical image registration and surgery simulation. *PhD thesis*, 1996.

[219] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, and J. Gee. N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

[220] M. Modat, D. Cash, P. Daga, G. Winston, J. Duncan, and S. Ourselin. Global image registration using a symmetric block-matching approach. *Journal of Medical Imaging*, 1(2):024003, 2014.

[221] M. Modat, G. Ridgway, Z. Taylor, M. Lehmann, J. Barnes, D. Hawkes, N. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, 98(3):278 – 284, 2010.

[222] J. Menke and T. Martinez. Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *2004 IEEE International Joint Conference on Neural Networks*, volume 2, pages 1331–1335, 2004.

[223] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214 – 224, 2015.

[224] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18 – 31, 2017.

[225] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. *Midas*, pages 1 – 6, 11 2008.

[226] *MSSEG challenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure*, Athènes, Greece, 2016.

[227] A. Carass, S. Roy, A. Jog, J. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. Sudre, M. Cardoso, N. Cawley, O. Ciccarelli, C. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. Iheme, D. Unay, S. Jain, D. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P. Bazin, P. Calabresi, C. Crainiceanu, L. Ellingsen, D. Reich, J. Prince, and D. Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148:77 – 102, 2017.

[228] S. Hashemi, S. Salehi, D. Erdogmus, S. Prabhu, S. Warfield, and A. Gholipour. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2019.

[229] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95 – 113, 2007.

[230] B. Avants, C. Epstein, M. Grossman, and J. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26 – 41, 2008.

[231] R. Bajcsy and S. Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1 – 21, 1989.

[232] M. Beg, M. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, Feb 2005.

[233] A. Dalca, A. Bobu, N. Rost, and P. Golland. Patch-based discrete registration of clinical brain images. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 60–67. Springer, 2016.

[234] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through MRFs and efficient linear programming. *Medical Image Analysis*, 12(6):731 – 741, 2008.

[235] X. Han, L. Hibbard, and V. Willcut. GPU-accelerated, gradient-free MI deformable registration for atlas-based MR brain image segmentation. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 141–148. IEEE, 2009.

[236] K. Punithakumar, P. Boulanger, and M. Noga. A GPU-accelerated deformable image registration algorithm with applications to right ventricular segmentation. *IEEE Access*, 5:20374–20382, 2017.

[237] J. Wu, X. Yang, Z. Zhang, G. Chen, and R. Mao. A performance model for GPU architectures that considers on-chip resources: Application to medical image registration. *IEEE Transactions on Parallel and Distributed Systems*, pages 1–1, 2019.

[238] E. Geremia, O. Clatz, B. Menze, E. Konukoglu, A. Criminisi, and N. Ayache. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378 – 390, 2011.

[239] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214 – 224, 2015.

[240] S. Pereira, A. Pinto, V. Alves, and C. Silva. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

[241] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[242] T. Brosch, L. Tang, Y. Yoo, D. Li, A. Traboulsee, and R. Tam. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

[243] H. Sokooti, B. de Vos, F. Berendsen, B. Lelieveldt, I. Išgum, and M. Staring. Nonrigid image registration using multi-scale 3D convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–239. Springer, 2017.

[244] X. Yang, R. Kwitt, M. Styner, and M. Niethammer. Quicksilver: Fast predictive image registration – A deep learning approach. *NeuroImage*, 158:378 – 396, 2017.

[245] B. de Vos, F. Berendsen, M. Viergever, M. Staring, and I. Išgum. End-to-End unsupervised deformable image registration with a convolutional neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 204–212, Cham, 2017. Springer International Publishing.

[246] H. Li and Y. Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1075–1078, April 2018.

[247] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, and A. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, pages 1–1, 2019.

[248] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. 2016.

[249] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[250] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[251] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané,

R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[252] M. Battaglini, F. Rossi, R. Grove, M. Stromillo, B. Whitcher, P. Matthews, and N. De Stefano. Automated identification of brain new lesions in multiple sclerosis using subtraction images. *Journal of Magnetic Resonance Imaging*, 39(6):1543–1549, 2014.

[253] C. Zhang, W. Tavanapong, J. Wong, P. de Groen, and J. Oh. Real data augmentation for medical image classification. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 67–76, 2017.

[254] G. van Tulder and M. de Bruijne. Why does synthesized data improve multi-sequence classification? In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 531–538, 2015.

[255] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R.and Gunn, A. Hammers, E. Sakka, D. Dickie, M. Herná dez, N. Royle, J. Wardlaw, H. Rhodius-Meester, B. Tijms, A. Lemstra, W. van der Flier, F. Barkhof, P. Scheltens, and D. Rueckert. Pseudo-healthy image synthesis for white matter lesion segmentation. In *Simulation and synthesis in medical imaging*, pages 87–96, 2016.

[256] J. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. van Leemput, and B. Fischl. Is synthesizing MRI contrast useful for inter-modality analysis? In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 631–638, 2013.

[257] A. Chartsias, T. Joyce, M. Giuffrida, and S. Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814, March 2018.

[258] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. Valdés-Hernández, D. Dickie, J. Wardlaw, and D. Rueckert. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, 17:918 – 934, 2018.

[259] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. Pop, P. Girard, R. Ameli, J. Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1):13650, 2018.

[260] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3D magnetic resonance images. *IEEE Transactions on Medical Imaging*, 27(4):425–441, April 2008.

[261] O. Commowick, N. Wiest-Daesslé, and S. Prima. Block-matching strategies for rigid registration of multimodal medical images. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 700–703, May 2012.

[262] J. Manjón and P. Coupé. volbrain: An online MRI brain volumetry system. *Frontiers in Neuroinformatics*, 10:30, 2016.

[263] D. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63 – 72, 2009.

[264] S. Andermatt, S. Pezold, and P. Cattin. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 31–42, 2018.

[265] S. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Machine Learning in Medical Imaging*, pages 379–387, 2017.

[266] A. Birenbaum and H. Greenspan. Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. *Engineering Applications of Artificial Intelligence*, 65:111 – 118, 2017.

[267] A. Valcarcel, K. Linn, S. Vandekar, T. Satterthwaite, J. Muschelli, P. Calabresi, D. Pham, Melissa L. Martin, and R. Shinohara. MIMoSA: An automated method for intermodal segmentation analysis of multiple sclerosis brain lesions. *Journal of Neuroimaging*, 28(4):389–398, 2018.

[268] H. Deshpande, P. Maurel, and C. Barillot. Classification of multiple sclerosis lesions using adaptive dictionary learning. *Computerized Medical Imaging and Graphics*, 46:2 – 10, 2015.

[269] C. Sudre, M. Cardoso, W. Bouvy, G. Biessels, J. Barnes, and S. Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10): 2079–2102, Oct 2015.

[270] N. Shiee, P. Bazin, A. Ozturk, D. Reich, P. Calabresi, and D. Pham. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524 – 1535, 2010.

[271] S. Jain, D. Sima, A. Ribbens, M. Cambron, A. Maertens, W. van Hecke, J. de Mey, F. Barkhof, M. Steenwijk, M. Daams, F. Maes, S. van Huffel, H. Vrenken, and D. Smeets. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage: Clinical*, 8:367 – 375, 2015.

[272] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama. GAN-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738. IEEE, 2018.

[273] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, page 101552, 2019.