Universitat de Girona

AUTOMATED DETECTION OF NEW MULTIPLE SCLEROSIS LESIONS IN LONGITUDINAL BRAIN RESONANCE IMAGING

Onur Ganiler

Dipòsit legal: Gi. 1797-2014
http://hdl.handle.net/10803/283552

PhD Thesis

Automated detection of new multiple sclerosis lesions in longitudinal magnetic resonance imaging

Onur Ganiler

2014

PhD Thesis

Automated detection of new multiple sclerosis lesions in longitudinal magnetic resonance imaging

Onur Ganiler

2014

Doctoral Programme in Technology

Supervised by:

Xavier Lladó

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

Acknowledgments

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 colleagues, friends, family and various institutions. At the end of my thesis I would like to thank all those people who made this thesis possible. I would like to thank my Master thesis advisor Dr. Ayla Sayli to encourage me to apply for this PhD. program in Girona and give me this opportunity to live this unforgettable experience.

At this moment of accomplishment, first of all, I would like to thank my PhD. advisor Dr. Xavier Lladó for supporting me all the time during these past four years. When you stuck in details, his faith and advices will always rescue you and make the things easier and fun for you. He has always been beside me during the happy and hard moments to motivate me. Simply, without the support of Dr. Lladó, this thesis would not have been 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 them. I also gratefully acknowledge Dr. Jordi Freixenet who was the head of the VICOROB research group when I started doing my PhD. You cannot feel down when you are beside him. His understanding, encouragement and personal attention provided good and smooth basis for my PhD. I would also like to thank to them for their friendship during this journey. Not only were they advisers for me but also good friends, and they will always be.

Most of the results described in this thesis would not have been obtained without a close collaboration with few institutions. It's my pleasure to acknowledge the neurologist Dr. Lluís Ramió-Torrentà from the Hospital Dr. Josep Trueta. He deserves special mention here for his constant support, help and suggestions. My special thanks are also due to the radiologist Dr. Brigitte Beltrán from the Hospital Dr. Josep Trueta for her support and suggestions. I am also very much thankful to the radiologist Dr. Àlex Rovira from

the Hospital Vall d'Hebron and radiologist Dr. Kai Vilanova from the Clínica Girona. Without their support and data this thesis certainly would not have been done.

A journey is easier when you travel together. Words are short to express my deep sense of gratitude towards my following colleagues Mariano Cabezas, Yago Diez, Eloy Roura and Sergi Valverde from the SALEM project. Without their collaboration this thesis would have been taken much more time. I would also like to thanks for their friendships. Since the beginning, we had been always in a great collaboration with Mariano which certainly helped to make this thesis better in a shorter time and we had also great time together, especially in the international conferences that we have participated. I am also very much thankful to Yago for his collaboration and friendship. I was able to bring this thesis to a higher level with his support and suggestions. I should also mention that I had great fun with Yago when we were running together through Girona's pathways. Thanks to Yago for not allowing me to feel alone in Girona.

I am also indebted to my many student colleagues from VICOROB group for providing a stimulating and fun filled environment. I would like to thank Shivav, Konstantin, Tudor, Ricards (Campos and Prados), Alberts (Pla and Gubern), Josep, Gerard, Guillem, Mojdeh, Sonia, Habib and many others for their friendships and supports. I would also like to thank to Joseta, Mireia and Aina for their every support and friendships. Thanks also to Patricia who helped me to find the house that I had lived in Girona for 4 years. My thanks also to Dr. Robert Martí, Dr. Joan Martí and Dr. Rafael Garcia for their helpful suggestions and comments during my progress report presentations.

Last but not least, I would like to thank my friends from Turkey and my family for their every single support. Especially to my mother Cahide, my father Atilla and my beatiful wife Esra. They have been always with me. Without my wife Esra I would not have made a great living in Girona and without her support this thesis would have been always lacked something. By the way she is still working on my cover page of the book :).

Thank you all !!!

Publications

Journals

- **[NR 2014]** O. Ganiler, A. Oliver, Y. Díez, J. Freixenet, J.C. Vilanova, B.Beltrán, Ll. Ramió-Torrentà, A. Rovira, and X Lladó. "A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies". Neuroradiology, 56(5), pp. 363-374. 2014. [JCR RNMMI IF 2.700,Q2(34/120)].
- **[NR 2012]** X Lladó, O. Ganiler, A. Oliver, R. Martí, J. Freixenet, L. Valls, J.C. Vilanova, Ll. Ramió-Torrentà, A. Rovira. "Automated detection of multiple sclerosis lesions in serial brain MRI". Neuroradiology, 54(8), pp 787-807, 2012. [JCR RNMMI IF 2.700, Q2(34/120)].
- **[CIBM 2014]** Y. Díez, A. Oliver, E. Roura, O. Ganiler, J.C. Vilanova, Ll. Ramió-Torrentà, Alex Rovira, X Lladó. "Are skull stripping and bias filtering necessary for brain MRI longitudinal registration in MS patients?". Computers in Biology and Medicine, submitted

Conferences

- **[SEN 2012]** X. Lladó, M. Cabezas, O. Ganiler, A. Oliver, Y. Donoso, J. Freixenet, L. Valls, A. Quiles, G. Laguillo, D. Pareto, J.C. Vilanova, A. Rovira, Ll. Ramió-Torrentà. "SALEM: Herramientas informáticas para la detección de lesiones de esclerosis múltiple en estudios longitudinales mediante imágenes de resonancia magnética del cerebro". Sociedad Española de Neurología. Neurología, 27(Num. Esp.Congreso), pp 195-196. Barcelona, Spain, 2012. [JCR CN IF:1.322 Q3(142/191)]
- **[ECTRIMS 2012]** O. Ganiler, X. Lladó, A. Oliver, Y. Díez, J. Freixenet, J.C. Vilanova, A. Quiles, G. Laguillo, Ll. Ramió-Torrentà, D. Pareto, and A. Rovira.

"Detecting evolving white matter MS lesions in serial brain MRI studies: analysis of a subtraction approach". European Committee for Treatment and Research in Multiple Sclerosis conference. Multiple Sclerosis. 18(S4), pp 385. Lyon, France. October 2012. [JCR CN IF:4.255 Q1(27/191)]

- [**ECTRIMS 2012**] Y. Díez, X. Lladó, A. Oliver, R. Martí, E. Roura, M. Cabezas, O. Ganiler, J. Freixenet, J.C. Vilanova, L. Valls, Ll. Ramió-Torrentà, D. Pareto, and A. Rovira. "Registration of serial brain MRI scans from multiple sclerosis patients. Analysis of 3D intensity-based methods". European Committee for Treatment and Research in Multiple Sclerosis conference. Multiple Sclerosis. 18(S4), pp 384-385. Lyon, France. October 2012. [JCR CN IF:4.255 Q1(27/191)]

- [**ECTRIMS 2011**] X. Lladó, O. Ganiler, A. Oliver, M. Cabezas, J. Freixenet, J.C. Vilanova, L. Valls, Ll. Ramió-Torrentà, and A. Rovira. "Computer-assisted strategies to automated quantification of multiple sclerosis lesion evolution on brain magnetic resonance imaging". European Committee for Treatment and Research in Multiple Sclerosis conference. Multiple Sclerosis, 17(10S), pp 161-162. Amsterdam, Holland, 2011. [JCR CN IF:4.255 Q1(27/191)]

- [**SEN 2011**] X. Lladó, O. Ganiler, A. Oliver, M. Cabezas, J. Freixenet, J.C. Vilanova, L. Valls, Ll. Ramió-Torrentà, and A. Rovira. "Técnicas automáticas de segmentación de lesiones de EM y de cuantificación volumétrica en estudios temporales". Sociedad Española de Neurología. Neurología, 26(Num. Esp.Congreso), pp 181-182. Barcelona, Spain, 2011. [JCR CN IF:0.790 Q4(161/191)]

- [**ECTRIMS 2010**] X. Lladó, M. Cabezas, O. Ganiler, A. Oliver, J. Freixenet, J.C. Vilanova, A. Quiles, Ll. Ramió-Torrenta, A. Rovira. "Strategies for Automated Segmentation of Multiple Sclerosis Lesions on Brain Magnetic Resonance Imaging". European Committee for Treatment and Research in Multiple Sclerosis conference. Multiple Sclerosis, 16(10), pp S256. Gothenburg, Sweden, 2010. [JCR CN IF:4.230 Q1(29/185)]

Workshops

- [**JEMGI 2014**] A. Oliver, X. Lladó, J. Freixenet, M. Cabezas, O. Ganiler, Y. Diez, E. Roura, S. Valverde. "Neuroimatge en EM: de les noves tecnologies a la capçalera

del pacient". Ponència invitada a la IV Jornada d'Esclerosis Múltiple de Girona celebrada a Pals (Girona), el 4/07/2014.

- [**JEMGI 2012**] J. Freixenet, X. Lladó, A. Oliver, M. Cabezas, O. Ganiler, Y. Diez, E. Roura, S. Valverde. "Anàlisi d'imatge automatitzat en Esclerosi Múltiple: Identificació de lesions i atròfia". Ponència invitada a la II Jornada d'Esclerosis Múltiple de Girona celebrada a Pals (Girona), el 13/07/2012.

- [**TIC Salut 2012**] X. Lladó, A. Oliver, J. Freixenet, M. Cabezas, O. Ganiler. "Projecte SALEM: Segmentació de lesions d'esclerosi múltiple". In III Jornada R+D+i en TIC Salut realitzada en el Parc Científic i Tecnològic de la Universitat de Girona. 7 i 8 de juny 2012.

- [**MICCAT 2011**] O. Ganiler, A. Oliver, J. Freixenet, X. Lladó. "Determining MS Lesion Evolution on Brain MRI". In Medical Image Computing in Catalunya: Graduate Student Workshop (non-indexed). Girona, October 2011.

- [**JEMGI 2011**] X. Lladó, A. Oliver, J. Freixenet, M. Cabezas, O. Ganiler. "Técnicas de segmentación automática de imágenes de RM Neurológicas". Ponència invitada a la I Jornada d'Esclerosi Múltiple de Girona celebrada a Pals (Girona), el 15/07/2011.

List of Acronyms

4D-CCA 4D Connected Component Analysis

ANN Artificial Neural Network

BL Black Hole Lesion

BNR Basal Neighborhood Ratio

BSE Brain Surface Extractor

BZS Basal Z-Score

CC Correlation Coefficient

CNR Contrast-To-Noise Ratio

COV Coefficient of Variation

CSE Conventional Spin Echo

CSF Cerebrospinal Fluid

DE Dual Echo

DFM Deformation Field Morphometry

DIS Dissemination in Space

DIT Dissemination in Time

DSC Dice Similarity Coefficient

DSCR Region-wise Dice Similarity Coefficient

DSCV Voxel-wise Dice Similarity Coefficient

EL Enhancing Lesion

EM Expectation Maximization

FCM Fuzzy C-Mean

FCS Fuzzy-Connectedness Segmentation

FDR False Discovery Rate

FFE Fast Field Echo

FLAIR Fluid-Attenuated Inversion-Recovery

FMRIB Functional MRI of the Brain

FN False Negative

FNR Follow-up Neighborhood Ratio

FOV Field of View

FP False Positive

FPR False Positive Rate

FSE Fast Spin Echo

Gd Gadolinium

GE Gradient Echo

GEL Gadolinium Enhancing Lesion

GLCM Gray Level Co-occurence Matrix

GLRT Generalized Likelihood Ratio Test

GM Gray Matter

GML Gray Matter Lesion

HL Hyperintense Lesion

HMRF Hidden Markov Random Fields

ICC Intracranial Cavity

IR Inversion Recovery

ITK Insight Segmentation and Registration Toolkit

KNN K-Nearest Neighbor

MI Mutual Information

MP-RAGE Magnetization Prepared - Rapid Acquisition with Gradient Echo

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

MS Multiple Sclerosis

NMI Normalized Mutual Information

PD-w Proton Density Weighted

PPMS Primary Progressive Multiple Sclerosis

PRMS Progressive Relapsing Multiple Sclerosis

PV Partial Volume

PVC Partial Volume Correction

PVE Partial Volume Effect

PVEC Partial Volume Effect Correction

RARE Rapid Acquisition with Refocused Echoes

RBF Radial Basis Function

RF Radio Frequency

RFC Random Forest Classification

RLM Run Length Matrix

ROI Region of Interest

RR Relapsing Remitting

RRMS Relapsing Remitting Multiple Sclerosis

SE Spin Echo

SNR Signal-To-Noise Ratio

SPMS Secondary Progressive Multiple Sclerosis

SSD Sum of Squared Differences

SVM Support Vector Machines

T1-w T1-weighted

T2-w T2-weighted

TDS Template Driven Segmentation

TE Echo Time

TN True Negative

TP True Positive

TPR True Positive Rate

TR Repetition Time

TSE Turbo Spin Echo

VDF Vector Displacement Field

VE Variational Echo

VI Visual Inspection

WM White Matter

WML White Matter Lesion

WMM White Matter Masking

WMSA White Matter Signal Abnormalities

List of Figures

1.1 MS prevalence by country 2

1.2 Progression types of MS 3

1.3 Intensity of various tissues at T1-w and T2-w imaging 5

1.4 Different MR images of the brain 8

1.5 Example of regions in which lesions are typically seen in MS 9

2.1 An example of an MS lesion serial analysis 20

2.2 Proposed classification of MS lesion serial analysis 24

2.3 Flowchart of the lesion detection approaches 28

2.4 Generated 3D volume with MS lesions segmented by two different experts . 33

2.5 Flowchart of the main change detection categories 35

2.6 An example of the pitfalls of an evaluation methodology 48

2.7 Inter-observer agreement of the subtraction-based approaches 50

3.1 Flowchart of the unsupervised pipeline 60

3.2 Illustration of the validation spaces 61

3.3 Skull stripping 63

3.4 Bias field correction 65

3.5 Histogram Matching 66

3.6 Subtracted image and WM masking 71

3.7 BNR, FNR and BZS features on a FP caused by CSF 76

3.8 BNR, FNR and BZS features on a FP caused by low signal on baseline image 76

3.9 BNR, FNR and BZS features on a FP caused by non-brain tissue in T2-w . 77
3.10 BNR, FNR and BZS features on a FP caused by non-brain tissue in PD-w . 77
3.11 Flowchart of the supervised pipeline 80
3.12 An example of MS lesion detection postprocessing 81

4.1 Comparison of the registration methods on different validation spaces . . . 90
4.2 Analysis of the preprocessing steps' performance in the pipeline 92
4.3 Lesion Sensitivity vs. False Discovery Rate for different operating points . . 93
4.4 Comparison of the automated thresholding methods 94
4.5 Analysis of the different registration methods 96
4.6 Comparison of the white matter masks 97
4.7 Comparison of the WM masking methods 98
4.8 Analysis of Gaussian filter performance in the pipeline 99
4.9 Visual examples of automated detections in unsupervised pipeline 102
4.10 Analysis of the postprocessing steps 105
4.11 Visual examples of automated detections in supervised pipeline 107

5.1 A prototype of the pipeline implemented by C++.NET. 116

A.1 Patient1 (12M): Visual examples of automated detections of new lesions . . 122
A.2 Patient2 (12M): Visual examples of automated detections of new lesions . . 123
A.3 Patient3 (12M): Visual examples of automated detections of new lesions . . 124
A.4 Patient4 (12M): Visual examples of automated detections of new lesions . . 125
A.5 Patient5 (12M): Visual examples of automated detections of new lesions . . 126
A.6 Patient6 (12M): Visual examples of automated detections of new lesions . . 127
A.7 Patient7 (12M): Visual examples of automated detections of new lesions . . 128
A.8 Patient8 (12M): Visual examples of automated detections of new lesions . . 129
A.9 Patient9 (12M): Visual examples of automated detections of new lesions . . 130
A.10 Patient10 (12M): Visual examples of automated detections of new lesions . 131
A.11 Patient11 (48M): Visual examples of automated detections of new lesions . 132

A.12 Patient12 (48M): Visual examples of automated detections of new lesions . 133
A.13 Patient13 (48M): Visual examples of automated detections of new lesions . 134
A.14 Patient14 (48M): Visual examples of automated detections of new lesions . 135
A.15 Patient15 (48M): Visual examples of automated detections of new lesions . 136
A.16 Patient16 (48M): Visual examples of automated detections of new lesions . 137
A.17 Patient17 (48M): Visual examples of automated detections of new lesions . 138
A.18 Patient18 (48M): Visual examples of automated detections of new lesions . 139
A.19 Patient19 (48M): Visual examples of automated detections of new lesions . 140
A.20 Patient20 (48M): Visual examples of automated detections of new lesions . 141

List of Tables

2.1 Classification of the lesion evolution methods 25

2.2 Summary of the results obtained by different lesion detection approaches . . 51

4.1 Study population for the 12M dataset 87

4.2 Study population for the 48M dataset 88

4.3 Performance of the unsupervised pipeline by lesion size 100

4.4 Performance of the unsupervised pipeline per patient 101

4.5 Impact of image combination, postprocessing and WM masking 104

4.6 Supervised pipeline using BNR, FNR and BZS features 108

4.7 Results for supervised pipeline using GLCM features 109

4.8 Comparison of the supervised and unsupervised pipelines 109

A.1 Patient1 (12M): Performance of the pipeline per lesion size 122

A.2 Patient2 (12M): Performance of the pipeline per lesion size 123

A.3 Patient3 (12M): Performance of the pipeline per lesion size 124

A.4 Patient4 (12M): Performance of the pipeline per lesion size 125

A.5 Patient5 (12M): Performance of the pipeline per lesion size 126

A.6 Patient6 (12M): Performance of the pipeline per lesion size 127

A.7 Patient7 (12M): Performance of the pipeline per lesion size 128

A.8 Patient8 (12M): Performance of the pipeline per lesion size 129

A.9 Patient9 (12M): Performance of the pipeline per lesion size 130

A.10 Patient10 (12M): Performance of the pipeline per lesion size 131

A.11 Patient11 (48M): Performance of the pipeline per lesion size 132
A.12 Patient12 (48M): Performance of the pipeline per lesion size 133
A.13 Patient13 (48M): Performance of the pipeline per lesion size 134
A.14 Patient14 (48M): Performance of the pipeline per lesion size 135
A.15 Patient15 (48M): Performance of the pipeline per lesion size 136
A.16 Patient16 (48M): Performance of the pipeline per lesion size 137
A.17 Patient17 (48M): Performance of the pipeline per lesion size 138
A.18 Patient18 (48M): Performance of the pipeline per lesion size 139
A.19 Patient19 (48M): Performance of the pipeline per lesion size 140
A.20 Patient20 (48M): Performance of the pipeline per lesion size 141

Contents

1 Introduction **1**

1.1 Multiple sclerosis 1

1.2 MRI, how it works? 2

1.3 MRI parameters and image contrast 4

1.4 MRI sequences 5

1.5 Why MRI for MS? 7

1.6 Conventional brain MR imaging in MS 7

1.6.1 T2-w lesions 8

1.6.2 T1-w lesions 10

1.6.3 Enhancing lesions 10

1.6.4 Gray matter involvement in MS lesions 11

1.7 Issues with MRI 11

1.8 The role of serial brain MRI in diagnosing MS 12

1.8.1 Dissemination in space of lesions (DIS) 12

1.8.2 Dissemination in time of lesions (DIT) 12

1.9 The role of computer-assisted technologies in serial brain MRI 13

1.10 Scope of the research 13

1.11 Objectives 15

1.12 Document structure 16

2 A review of automated detection of MS lesions in serial brain MRI 19

2.1 MS lesion detection 19

2.2 Classification of MS lesion detection and quantification in serial brain MRI 23

2.2.1 Proposed classification of lesion detection approaches 23

2.2.2 Proposed classification of lesion quantification 26

2.2.3 General problems in MRI 27

2.3 Lesion detection approaches 27

2.3.1 Supervised methods 30

2.3.2 Unsupervised methods 33

2.4 Change detection approaches 34

2.4.1 Intensity-based approaches 34

2.4.2 Deformation field-based approaches 38

2.5 Classification of MS lesion quantification in serial brain MRI 40

2.6 Experimental validation 41

2.6.1 Data preparation 42

2.6.2 Ground-truth preparation 43

2.6.3 Validation with the ground truth 44

2.6.4 Validation without a ground truth 45

2.7 Evaluation strategies for change detection 46

2.7.1 Visual inspection of outcome images 46

2.7.2 Validation by segmented lesions in baseline and follow-up images . . 47

2.7.3 Validation by segmented new lesions in follow-up images 47

2.8 Analysis of the reported results 49

2.8.1 Improvements and further trends 50

2.9 Conclusion 54

3 Temporal analysis proposal on MS lesion detection 57

3.1 Overview . . . 57

3.2 The proposed framework . . . 57

3.3 Validation spaces . . . 59

3.4 Preprocessing . . . 61

3.4.1 Skull stripping . . . 62

3.4.2 Bias field correction . . . 64

3.4.3 Histogram matching . . . 65

3.5 Registration . . . 66

3.5.1 Rigid registration . . . 66

3.5.2 Rigid halfway registration . . . 67

3.5.3 Non-rigid registration . . . 68

3.6 Tissue segmentation and WM masking . . . 70

3.6.1 Tissue segmentation using single modalities . . . 70

3.6.2 Atlas based multi-modal tissue classification . . . 71

3.7 3D subtraction . . . 72

3.8 Thresholding and locating of WM candidate lesions . . . 72

3.8.1 Gaussian filtering . . . 72

3.8.2 Unsupervised thresholding . . . 73

3.8.3 Supervised thresholding . . . 73

3.9 Postprocessing: refining candidate lesions . . . 74

3.9.1 Unsupervised pipeline . . . 74

3.9.2 Supervised pipeline . . . 78

3.10 Summary . . . 79

3.10.1 Preprocesing pipeline . . . 81

3.10.2 Registration of the images . . . 82

3.10.3 White matter segmentation . . . 82

3.10.4 Thresholding of the subtraction image . . . 82

3.10.5 Postprocessing 83

4 Experimental results **85**

4.1 Evaluation 85

4.2 Study population 85

4.3 Evaluation measures 87

4.4 Evaluating the validation space 89

4.5 Evaluating the preprocessing steps 91

4.6 Automated thresholding 93

4.6.1 Unsupervised thresholding 93

4.6.2 Supervised thresholding 94

4.7 Evaluating the registration methods 95

4.8 Evaluating the white matter masking methods 95

4.9 Evaluating the pipeline with different Gaussian filters 96

4.10 Detailed evaluation of the unsupervised pipeline for 12M and 48M datasets 98

4.10.1 Impact of image combination, postprocessing and WM masking . . . 101

4.11 Comparison with state of the art methods 105

4.12 Evaluating the pipeline when using supervised classifiers 106

4.13 Evaluating the supervised pipeline when using GLCM features 108

4.14 Comparing the unsupervised and supervised pipelines 108

4.15 Discussion 110

5 Conclusions **113**

5.1 Summary of the thesis 113

5.1.1 Contributions 115

5.2 Future work 117

5.2.1 Short term future improvements 117

5.2.2 Future research lines 118

A Detailed evaluation results **121**

A.1 Detailed performance per patient and lesion size 121

Bibliography **143**

Resum

Aquesta tesi es centra en la detecció automàtica de lesions noves d'esclerosi múltiple (EM) en estudis longitudinals del cervell mitjançant l'ús d'imatges de ressonància magnètica (RM). Aquesta malaltia es caracteritza per la presència de lesions al cervell, predominantment en el teixit de la matèria blanca, i la detecció i la quantificació de les noves lesions són elements crucials per al seguiment dels pacients. No obstant això, la detecció manual d'aquestes noves lesions no només requereix de molt temps, sinó que també és propensa a la variabilitat intra- i inter-observador. Per tant, el desenvolupament de tècniques automàtiques per a la detecció de lesions d'EM és un gran repte.

Després d'un anàlisi exhaustiu de l'estat de l'art de les diferents tècniques de detecció de lesions d'EM, en aquesta tesi es presenta una nova classificació assenyalant-ne les principals fortaleses i debilitats. També es proporciona una avaluació quantitativa complementària d'alguns dels mètodes més rellevants en la literatura. Posteriorment, es presenta una nova proposta, que combina diverses característiques de les diferents modalitats d'imatges de RM, basada en la subtracció d'imatges per tal de determinar els canvis entre una imatge basal i una de seguiment. En primer lloc, s'inclouen en la proposta mètodes de preprocessament, per tal de millorar la qualitat de les imatges de RM, així com mètodes de registre d'imatges rígids i no rígids. S'analitza en molt deteniment l'efecte d'aquests preprocessats en el resultat final de la proposta. També s'aplica un pas d'emmascarament de la matèria blanca amb la finalitat de reduir l'espai de cerca de les lesions només dins de la màscara. Posteriorment, s'aplica un llindar a les imatges de resta. Tot i que la determinació del llindar pot ser realitzada pels experts, en aquesta tesi es proposa un procés automatitzat de detecció del llindar òptim que proporciona un compromís satisfactori entre sensibilitat i especificitat. Finalment, es refinen les lesions candidates detectades utilitzant les característiques de la lesió, sobretot amb la finalitat de reduir la detecció de falsos positius. Per a aquest propòsit, s'inclouen les imatges basals i de seguiment, i es fusionen els resultats obtinguts a partir d'imatges PD-w i T2-w d'una manera supervisada i també no super-

visada. Els resultats experimentals s'avaluen en una base de dades de 20 pacients amb EM amb una càrrega variable de la lesió, on es disposa també de la segmentació manual proporcionada pels experts. L'avaluació s'ha realitzat de forma qualitativa i quantitativa, incloent una comparació dels diferents processos i usant diverses mètriques per a la detecció i segmentació.

Resumen

Esta tesis se centra en la detección de lesiones nuevas de esclerosis múltiple (EM) en estudios longitudinales del cerebro mediante el uso de imágenes de resonancia magnética (RM). Esta enfermedad se caracteriza por la presencia de lesiones en el cerebro, predominantemente en el tejido de la materia blanca. La detección y cuantificación de las lesiones nuevas son cruciales para el seguimiento de los pacientes con EM. Por consiguiente, la detección manual de estas lesiones nuevas no sólo requiere de mucho tiempo por parte del experto, sino que también es propensa a la variabilidad intra- e inter-observador. Por lo tanto, el desarrollo de técnicas automatizadas para la detección lesiones de la EM es un gran desafío.

Después de un análisis exhaustivo del estado del arte de las distintas técnicas de detección de lesiones de EM, en esta tesis se presenta una nueva clasificación de éstas señalando sus principales fortalezas y debilidades. También se proporciona una evaluación cuantitativa complementaria de algunos de los métodos más relevantes en la literatura. Posteriormente, se presenta una nueva propuesta basada en un enfoque de detección de cambios, que combina varias características de las diferentes modalidades de imágenes de RM. En primer lugar, varios métodos de preprocesamiento se incluyen en la propuesta para mejorar la calidad de las imágenes de RM. Analizamos estos procesos en detalle, así como diferentes métodos de registro rígidos y no rígidos de imágenes. La sustracción de las imágenes basal y de seguimiento se utiliza para determinar los cambios entre las dos imágenes. Por otra parte, se aplica un paso de enmascaramiento de la materia blanca con el fin de reducir el espacio de búsqueda de las lesiones sólo dentro de la máscara. Posteriormente, se aplica un umbral a las imágenes de sustracción. Aunque la determinación del umbral puede ser realizada por los expertos, en esta tesis se propone un proceso automatizado de detección del umbral óptimo que proporciona un compromiso satisfactorio entre sensibilidad y especificidad. Por último, las lesiones candidatas detectadas se refinan utilizando las características de la lesión, sobre todo con el fin de reducir la detección de

falsos positivos. Para este propósito, se incluyen las imágenes basales y de seguimiento, y se fusionan los resultados obtenidos a partir de imágenes PD-w y T2-w de una forma supervisada y también no supervisada. Los resultados experimentales se evalúan en una base de datos de 20 pacientes con EM con carga variable de lesión, donde se dispone también de la segmentación manual proporcionada por expertos. La evaluación, se ha realizado de forma cualitativa y cuantitativa, incluyendo una comparación de los distintos procesos y usando varias métricas para la detección y segmentación.

Abstract

This thesis deals with the detection of new multiple sclerosis (MS) lesions in longitudinal brain magnetic resonance (MR) imaging. This disease is characterized by the presence of lesions in the brain, predominantly in the white matter (WM) tissue of the brain. The detection and quantification of new lesions are crucial to follow-up MS patients. Moreover, the manual detection of these new lesions is not only time-consuming, but is also prone to intra- and inter-observer variability. Therefore, the development of automated techniques for the detection MS lesions is a major challenge. After a thorough analysis of the state-of-the art in MS lesion detection approaches, we present a new classification of techniques pointing out their main strengths and weaknesses. A complementary quantitative evaluation of some of the most remarkable methods in the literature is also provided. Subsequently, we present a new proposal based on a change detection approach, which combines various characteristics of different MR image modalities. Firstly, several preprocessing methods are included in the pipeline to improve the quality of MR images. We analyze these processes as well as several rigid and non-rigid image registration methods in detail. The subtraction of the baseline and follow-up images is used to determine changes between the images. Moreover, we apply a WM masking step in order to reduce the search space for lesions only within WM. Afterwards, we apply a threshold to the subtraction images. Although determining the threshold can be done by experts, we propose an automated thresholding process which provides a satisfactory trade-off between sensitivity and specificity. Finally, we refine the candidate lesions detected using lesion features, particularly in order to reduce false positive lesions. For this purpose, including the baseline and follow-up images, we join both results obtained from PD-w and T2-w images in a supervised and an unsupervised manner. Experimental results are evaluated on a database of 20 MS patients with a variable lesion load, where manual segmentation provided by experts was available. The evaluation, carried out in a quantitative and qualitative manner, includes a comparison and uses several metrics for detection and segmentation.

Chapter 1

Introduction

1.1 Multiple sclerosis

Multiple sclerosis (MS) is one of the world's most common neurological disorders affecting the central nervous system (CNS) and is generally considered to be autoimmune. Pathologically, MS is an inflammatory-demyelinating and neurodegenerative disease, clinically defined by demyelinating lesions and characterized by areas of inflammation, demyelination, axonal loss, and gliosis scattered throughout the CNS [27, 29]. 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 [28].

MS is considered to be caused by an interplay between genetic factors and the environment [28], however, the distribution of MS cannot be explained by environmental exposure and genetic susceptibility alone [27, 28]. As the estimated number of people with MS was 2.3 million in 2013 (33 per 100,000), it is more common in northern Europeans regions (140 per 100,000) (see Figure 1.1). Furthermore, the prevalence of MS also varies within regions. For instance, the highest prevalence in Europe is 189 per 100,000 in Sweden, whereas the lowest is 22 per 100,000 in Albania. On the other hand, age seems to be another important factor in MS since it is frequently seen in young adults. According to the Atlas of MS 2013, MS is usually diagnosed during early adulthood with an average age of MS onset of 30 years. Moreover, as with other autoimmune disorders, MS is approximately twice as common in women as in men, though, in some regions, the ratio of women to men is considerably higher, such as in East Asia where the female-to-male ratio is 3.0 [123].

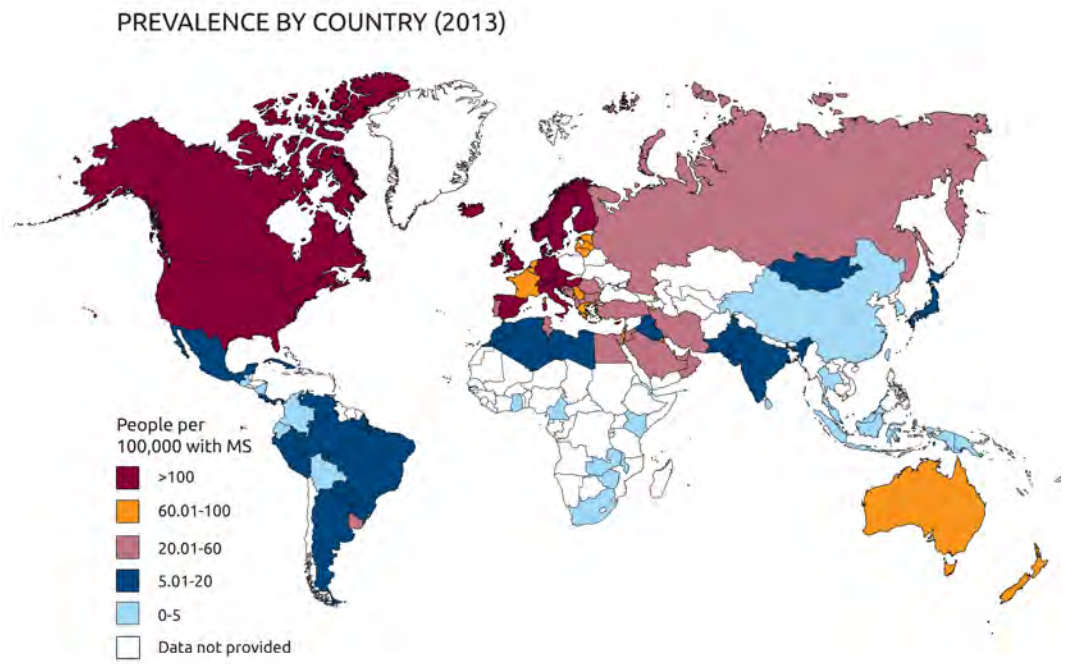


Figure 1.1: MS prevalance by country 2013 [123].

Several patterns of progression (subtypes) have been described in MS. In 1996, the United States National Multiple Sclerosis Society described four clinical courses: relapsing remitting MS (RRMS), secondary progressive MS (SPMS), primary progressive MS (PPMS) and progressive relapsing MS (PRMS) [72] (see Figure 1.2). At the time of diagnosis, around 85% of patients are diagnosed with a relapsing-remitting form of MS, while a small subset of patients (10%) are diagnosed with PPMS and 5% with PRMS. The majority of people (80%) diagnosed with RRMS will eventually go on to develop a more progressive form, SPMS [123].

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 for clinical purposes.

1.2 MRI, how it works?

As 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. Due to fact that a large pro-

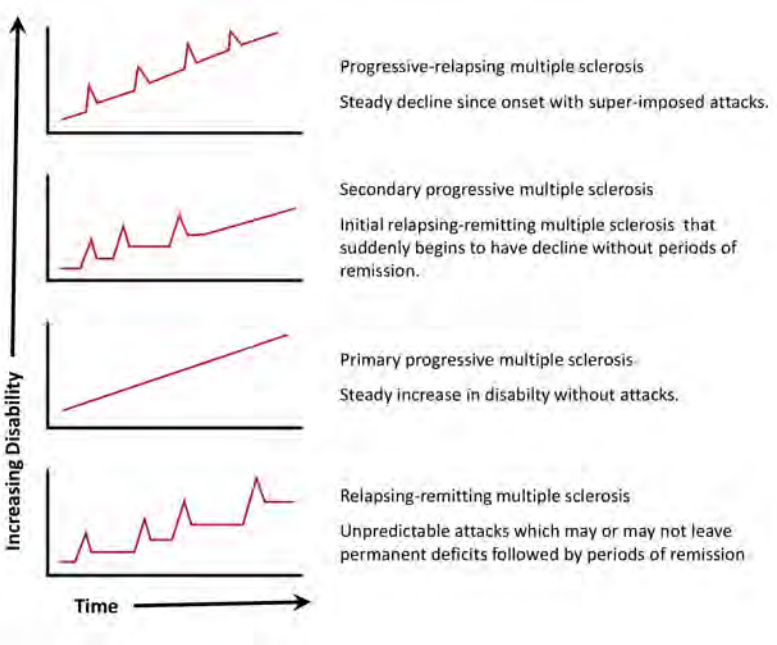


Figure 1.2: Progression types of MS.

portion of the human body is made up of fat and water, both of which contain lots of hydrogen atoms, the hydrogen atom is commonly used in MRI studies.

As each nuclei spins around its own axis, this motion induces a magnetic field and when the nuclei are exposed to an external magnetic field, the interaction between the two magnetic fields urges the nuclei to align with the magnetic field. When the nuclei, initially precessing with a wobble at various angles, aligns with the external magnetic field (9 atoms per million in a 1.5T system), this event creates magnetic moments. Different 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].

1.3 MRI parameters and image contrast

Two key parameters, repetition time (TR) and echo time (TE), are key to the creation of image contrast [14].

- *TR:* is the time between the application of an RF excitation pulse and the start of the next RF pulse.
- *TE:* refers to the time between the application of the RF pulse and the peak of the echo detected.

For instance, the difference in relaxation time 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. In this sense, TR relates to T1 and affects contrast in T1-weighted images and TE relates to T2 and affects contrast in T2-weighted images. Hence, both parameters affect contrast in MR images because they provide varying levels of sensitivity to differences in relaxation time between various tissues [14]. Consequently, for instance, a tissue with a long T1 and T2 (like water) is dark in the T1-weighted (*T1-w*) image and brighter in the T2-weighted (*T2-w*) image, whereas a tissue with a short T1 and a long T2 (like fat) is bright in the T1-weighted image and gray in the T2-weighted image (see also Figure 1.3). On the other hand, when the TR is long and the TE is short, the differences in magnetization recovery and in signal decay between fat and water are not distinguishable.

- *T1-w sequences:* short TR, short TE (TR $<$ 1000ms, TE$<$ 30 ms).
- *T2-w sequences:* long TR, long TE (TR $>$ 2000ms, TE $>$ 80 ms).
- *PD-w sequences:* long TR, short TE (TR $>$ 2000ms, TE $<$ 30 ms).

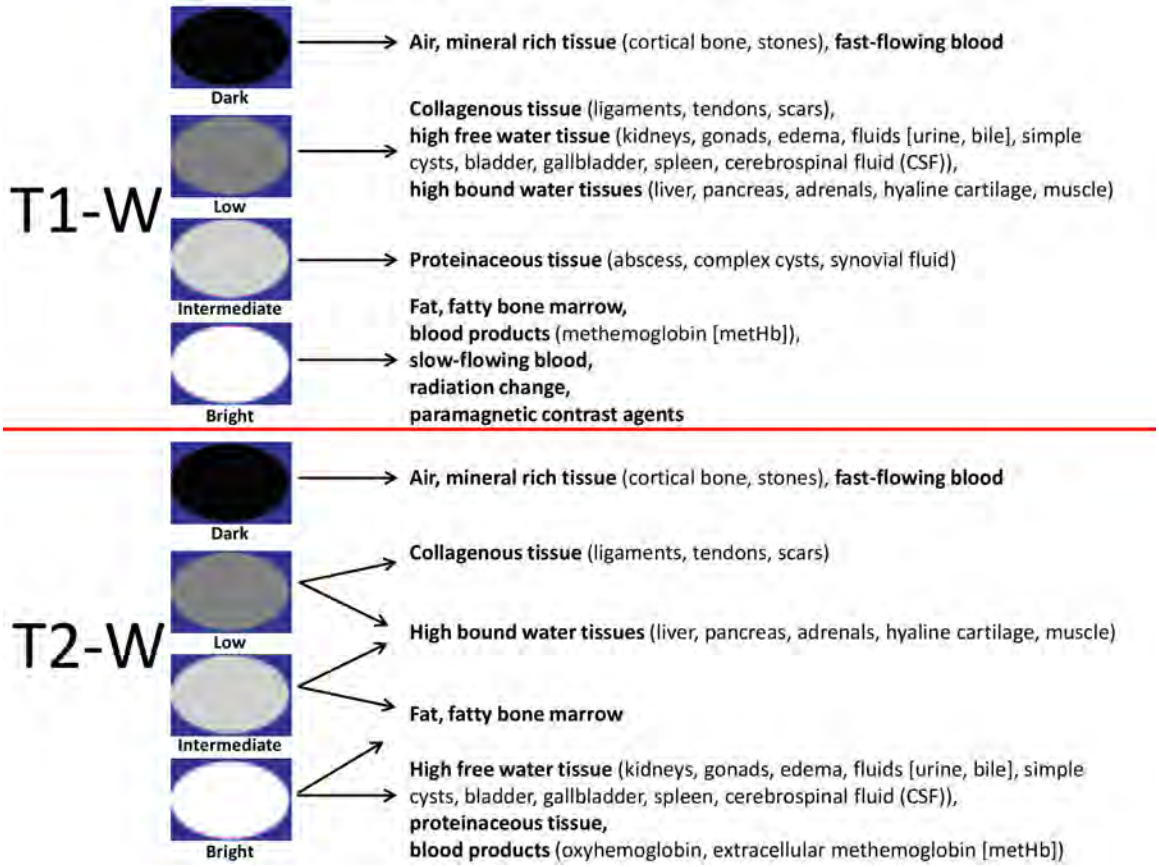


Figure 1.3: Diagram shows the signal intensity of various tissues at T1- and T2-weighted imaging [14].

Hence, the contrast observed in MR images depends on the difference in proton density so that tissues with more protons have a higher signal intensity and vice-versa [14]. This is called proton density weighted (*PD-w*) imaging. Note that, in all weighted MR imaging, the images show all types of contrast, however, T1 contrast is accentuated in T1-w and T2 contrast is accentuated in T2-w, while proton density is accentuated in PD-w imaging. Note that the images formed by MR pulse sequences can be 2D and 3D.

1.4 MRI sequences

A pulse sequence describes a series of RF pulses applied to a sample. A spin echo is the refocusing of spin magnetization by a pulse of resonant electromagnetic radiation. In MR image acquisition, additional gradient pulses are being applied by switching magnetic fields that exhibit a space-dependent gradient that can be used to reconstruct, after Fourier

Transform, spatially resolved images [13]. Therefore, in MRI studies, there are only two fundamental types of MR pulse sequences: Spin Echo (SE) and Gradient Echo (GE) sequences. All other MR sequences are variations of these two sequences obtained by using different parameters [14]. SE based sequences cover conventional SE (CSE), fast (turbo) sequences (FSE or TSE) and additionaly inversion recovery (IR) methods.

The hydrogen atoms are first hit with an excitation pulse that tips the atoms 90° from their original orientation into the transverse plane so that they spin and can be detected by catching the signals from their spin. This is followed by another excitation pulse that flips the hydrogen atoms' position 180° so that they are synchronized and we can collect the maximum signal. However, the additional 180° refocusing pulse has to be repeated several times in order to keep them synchronized, therefore, there is only one 90°, but many 180° excitations per TR. As a result, particularly in CSE, more time is needed due to having to collect more data per TR. In this sense, FSE is a way of manipulating the CSE technique to save time. FSE allows high resolution imaging in a reasonable amount of time with less severe motion artifacts, better signal-to-noise ratio (SNR), but with a decreased number of slices [48]. In FSE imaging, cerebrospinal fluid (CSF) is brighter in PD-w images and fat is brighter in T2-w images. Moreover, some MS plaques and other lesions at the brain/CSF interface might be missed in FSE due to the fact that distinguishing between CSF and periventricular high-intensity plaques is more difficult [48].

The Inversion Recovery technique is an SE sequence using a 180° flip of the atoms in order to null the signal from a specific tissue or a particular entity (like water) so that no signal is generated for that particular tissue. This is done by applying an inversion pulse before the normal pulse sequence (SE, FSE, etc) [14]. Particulary for MS patients, the Fluid-Attenuated Inversion-Recovery (*FLAIR*) is used. In FLAIR sequences, the signal from CSF is nulled and appears dark, which can be useful for some lesions that are not easily distinguishable due to CSF.

On the other hand, Gradient Echo sequences (*GE*) are an alternative technique to spin echo sequences. They use gradient fields to generate transverse magnetization and flip angles of less than 90°. The Gradient Echo sequences show a wide range of variations compared to the Spin Echo and Inversion Recovery sequences. Gradients are used to dephase and rephase transverse magnetization. GE is useful when fast scans are needed but does not correct for local magnetic field inhomogeneities, which translates into the presence of artifacts in the image [15]. This technique is particularly helpful in diagnosing hemorrhagic contusions such as cerebral hemorrhagic contusions [14]. Note also that T1-w

rapid gradient-echo (MP-RAGE) is considered to have better image quality and contrast between gray and white matter than the T1-weighted spin echo sequence [18].

1.5 Why MRI for MS?

With MRI it is possible to detect contrast differences in soft tissues. Furthermore, by manipulating the MR parameters, one can optimize the pulse sequence for certain pathologies such as adjusting the TR and TE to emphasize a particular type of contrast [14]. Additionally, it has been demonstrated that MRI is highly sensitive for detecting MS plaques. Hence, 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 [96]. For instance, as a clinically isolated syndrome (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 [96, 100]. Moreover, a number of trials have reported that MRI is useful in monitoring early treatment of MS and offers a opportunity to reduce the disease's activity and may slow disability progression [103]. 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 [103, 96, 24].

1.6 Conventional brain MR imaging in MS

Conventional MR (*cMRI*) sequences used in MS, covering T1-w, gadolinium-enhanced T1-w, PD/T2-w [103] and FLAIR [105] spin-echo sequences, are accepted in standard protocols for diagnosis and treatment outcome measures in clinical trials [96]. Particularly in RRMS and SPMS patients, disease activity is detected more frequently with cMRI than with clinical assessment of relapses [105]. Therefore, cMRI-derived metrics have become established as the most important paraclinical tool for MS patients [103, 96, 24].

Furthermore, cMRI modalities offer a high contrast between the main brain tissues, gray matter (GM), formed by neuron nuclei, white matter (WM), formed by neuronal axons, and CSF which is the colorless bodily fluid that provides protection and cerebral autoregulation of cerebral blood flow. The CSF appears dark in both T1-w and FLAIR images while it is the brightest tissue in T2-w and has relatively similar intensities to

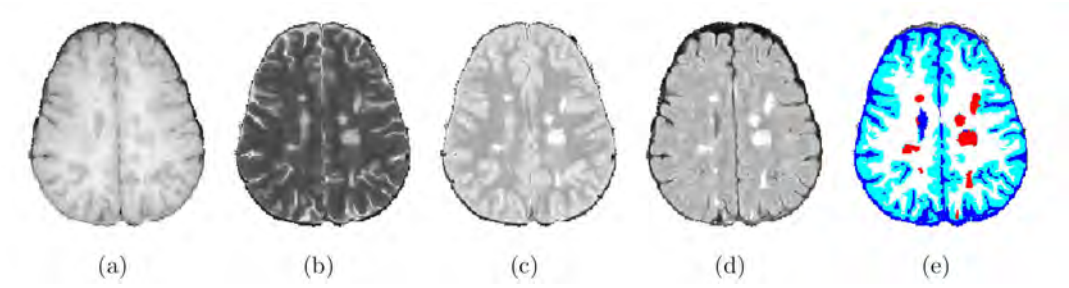


Figure 1.4: Different MR images of the brain: a) T1-w image, b) T2-w image, c) PD-w image and d) FLAIR image and their e) tissue segmentation: CSF appears dark blue, GM appears blue, WM appears white and lesions appear red.

GM tissue in PD-w images. On the other hand, WM is the brightest tissue in T1-w and has the lowest signal in both PD-w and T2-w images while showing an intermediate gray level in FLAIR images similar to GM. Lastly, GM tissue has an intermediate gray level in T1-w and T2-w images in comparison with other tissues. All sequences come with their 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 and fluid areas appear brighter in T2-w images [14]. 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 (see Figure 1.4)

1.6.1 T2-w lesions

T2-w SE sequences are created by a long TR and consist 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 [80]. In T2-w sequences, the characteristic appearance of MS is bright hyperintense lesions (HL), reflecting their increased water content. T2-weighted lesions do not have a pathological specificity and can be caused by inflammation, demyelination, gliosis, edema or axonal loss. Both acute and chronic lesions appear in T2-w images. They are typically discrete and focal in the early stages of the disease, however, more subtle as the disease progresses.

These lesions are more frequent in periventricular areas and also typically seen in juxtacortical, infratentorial and temporal regions (see Figure 1.5), as well as in the corpus callosum. Note that periventricular lesions are more easily identified in PD-w images [83, 105] since they give better contrast between periventricular MS lesions and CSF when com-

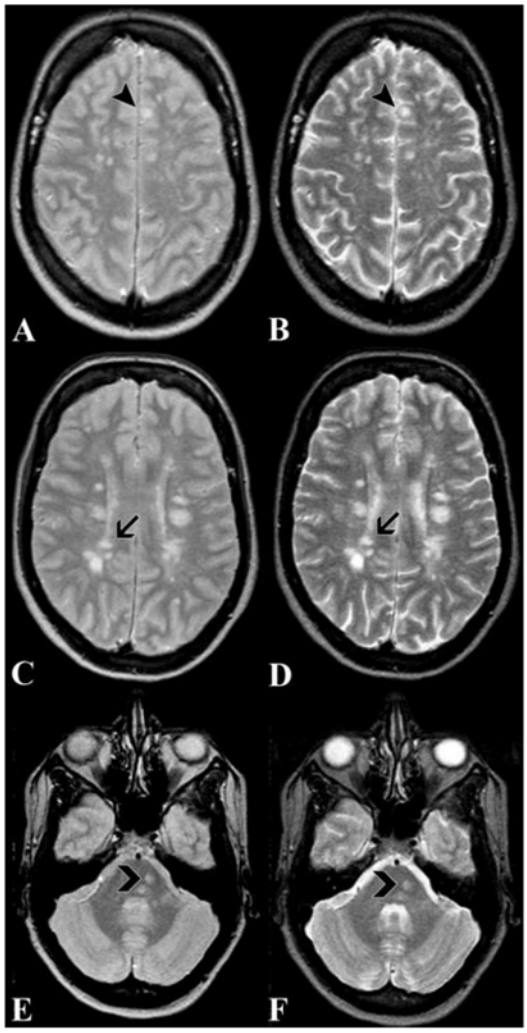


Figure 1.5: Example of regions in which lesions are typically seen in MS. ACE: Proton-density (short-echo) images; BDF: T2 (long-echo) images. Arrowheads: Juxtacortical lesion; Arrows: Periventricular lesion; Delta arrows: Infratentorial lesion. (Moraal et al. 2010)

pared to long TE T2-w images, but suffer more from flow artifacts, particularly in the posterior cranial fossa, which makes it difficult to identify infratentorial lesions. Also note that, as FLAIR images produce heavily T2-w images by nulling 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 [105].

1.6.2 T1-w lesions

Unlike T2-w lesions, MS lesions in T1-w sequences can be both hyperintense and hypointense. Though the majority of MS lesions seen in T2-w images are isointense or only slightly hypointense lesions in T1-w images (HL lesions) [19], around 10% to 20% of T2-w hyperintensities are also seen in T1-w images as hypointense areas compared with normal-appearing white matter (NAWM), so called black holes (BL) [103]. The study by Vavasour et al. [134] suggests that myelin loss occurs equally in both chronic isointense and hypointense lesions but hypointense lesions are distinguished by increased extracellular water.

Black holes are considered to be acute when they coincide with contrast-enhancing lesions, otherwise they are considered as chronic or persistent. True chronic black holes are usually defined as T1-hypointense lesions and provide a more accurate correlation with disability compared to T2-w lesions [80, 105]. Moreover, chronic black holes are more frequent in the progressive stage than in RRMS and more frequently seen in supratentorial WM as compared to the infratentorial WM area [80].

1.6.3 Enhancing lesions

Enhancing lesions (EL) reflects the blood-brain barrier (BBB) disruption and is considered to be the inflammatory phase of lesion development [19]. Note that the BBB breakdown is a consistent early feature of new lesion development in RRMS and SPMS patients. In this sense, after the injection of gadolinium (Gd) in combination with T1-w images, which suppresses a normal brain but not enhancement, a subset of T2-w lesions shows this enhancement, so called Gd-enhancing lesions [80]. Approximately 65-80% of contrast enhancing lesions have a corresponding hypointensity in native T1-w images [105] and these acute hypointense lesions may become isointense or develop into BL lesions.

Gadolinium-enhanced T1-w imaging detects disease activity more frequently than clinical evaluation of relapses, suggesting that most enhancing lesions are silent. Although contrast enhancement is more sensitive than T2-weighted images in detecting disease activity, it is expensive and there are more false positive lesions (i.e. small vessels) and/or flow artifacts (i.e. around the brainstem and posterior fossa), while delaying scanning interferes with the patient's throughput [80]. Note that, while all new lesions seen in T2-w images are initially observed as areas of the BBB disruption, a few enhancing lesions appear without an accompanying T2-w lesion.

1.6.4 Gray matter involvement in MS lesions

Although MS is predominantly a disease of the white matter, it is also characterized by lesions in the gray matter [23]. Between 5% and 10% of the lesions may involve gray matter, including the cerebral cortex and basal ganglia [105]. For instance, cortical gray matter lesions (*GML*) and deep gray matter lesions can comprise approximately 5% of the total lesion volume [23]. Having a less severe degree of inflammation and an intermediate high signal intensity, GM lesions are more obscure and thus more difficult to detect in MR imaging when compared to white matter lesions (*WML*) [105].

1.7 Issues with MRI

The detection of an MS lesion is highly related to the contrast-to-noise ratio (CNR) and signal-to-noise ratio (SNR) for a region of interest (ROI). In this sense, voxel (3D) or pixel (2D) size, slice thickness and many other parameters play an important role. For example, smaller lesions require a higher CNR and thinner slices, however, a reduction of slice thickness decreases the SNR. Likewise, when the image voxel size is larger, a greater SNR is obtained, however, for small lesions, a smaller voxel size is needed [80]. Also note that 3D sequences have intrinsically higher SNRs when compared to 2D sequences, and therefore, may reveal more subtle changes in lesions over time compared to 2D sequences.

On the other hand, noise and image artifacts due to the scanner's performance [68], such as radio-frequency (RF) artifacts [17], are other issues in MRI to deal with. Repositioning errors, motion artifacts due to inadvertent head movement, inconsistent objects over time such as blood and cerebrospinal fluid flow artifacts [85] can also affect the detection and quantification accuracy of MS lesions [68]. Additionally, inhomogeneities in the magnetic field due to the imperfections in the image acquisition process, as well as patient properties, can cause a smooth inhomogeneity field across the image, known as the bias field error. Furthermore, partial volume effects (PVE), where a single voxel contains a mixture of multiple tissue values, may also introduce errors into the MS quantification and tissue segmentation processes [61].

Apart from the imaging problems, MRI measures are limited in their sensitivity and specificity, failing to provide a comprehensive assessment of the underlying pathology [9]. The correlations between disability and conventional MRI parameters are relatively modest [80] mostly due to the fact that MRI lesions are often clinically silent and MRI changes do not necessarily correlate well with clinical disability [105]. Moreover, a moderate cor-

relation has been demonstrated between the degree of clinical disability and the mean of enhancing lesions in RRMS and SPMS patients [105] and there is much less MRI activity in the primary progressive group [80]. Therefore, more studies are still needed for MR imaging.

1.8 The role of serial brain MRI in diagnosing MS

The criteria for establishing the diagnosis of MS relies on the principle of demonstration of demyelinating lesions disseminated in space (DIS) and time (DIT) [96]. These principals were first codified in 1983 by the Poser committee [98], then in 1997 Barkhof et al. [10] proposed a four-parameter MRI model to predict the development of CDMS in patients presenting a CIS, which was later modified by Tintore et al. [125]. After an international panel on the diagnosis of MS [77], accepted the Barkhof/Tintore criteria [10, 39, 125] into their scheme for demonstrating DIS, they made use of advances in magnetic resonance imaging (MRI) techniques for diagnostic criteria for multiple sclerosis, known as the McDonald Criteria. Afterwards, the McDonald criteria underwent revisions in 2005 [97], which simplified the MRI evidence required for DIT and in 2010 [96], taking into consideration the MAGNIMS research group's studies [117, 104, 84].

As a result, a diagnosis of multiple sclerosis in patients who present a CIS for the first time can be established with MRI, if the MRI demonstrates demyelinating lesions with dissemination in space (DIS) and dissemination in time (DIT) according to the principles defined in Sections 1.8.1 and 1.8.2 [96].

1.8.1 Dissemination in space of lesions (DIS)

According to the last revised McDonald Criteria [96], DIS can be demonstrated with at least 1 T2 lesion in at least 2 out of 4 typical locations for MS as specified in the original McDonald criteria; periventricular, subcortical, infratentorial and spinal cord areas. Note that in the event of brainstem spinal cord syndromes, lesions in these regions do not contribute to the demonstration of DIS.

1.8.2 Dissemination in time of lesions (DIT)

The panel [96], abandoning the requirement for an extra reference MRI after 30 days, allows a new T2 lesion to establish DIT 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 non-enhancing lesions 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, irrespective of the timing of the baseline MRI (serial MRI imaging).

1.9 The role of computer-assisted technologies in serial brain MRI

The most common reason for falsely attributing a patient's symptoms to MS is faulty interpretation of the MRI [105]. The manual detection of change is not only time-consuming, but is also prone to intra- and inter-observer variability [111]. Therefore, using conventional MRI modalities, an automated lesion detection and quantification method, without doubt, will help neuroradiologists to improve the diagnosis and follow-up of MS patients. In fact, over the last few years, there have been numerous studies dealing with these issues. Furthermore, using advance MRI techniques can improve the understanding of the natural history of the disease and monitoring the efficacy of experimental treatments as well as enhance our understanding of tissue damage in MS.

1.10 Scope of the research

The Computer Vision and Robotics group (VICOROB) of the University of Girona has been working on medical image analysis since 1996, mainly in segmentation and registration of mammographic images. Thanks to their previous knowledge acquired through several medical projects, the group started to focus their research on brain MRI analysis. This new line of research started with the segmentation of MS lesions and has expanded to other fields such as temporal analysis, registration (temporal and intersubject) or atrophy analysis.

All these studies have been carried out within the funded research projects CEMCAT2011 "AVALEM: Avaluació de l'atròfia en pacients amb lesions d'esclerosi múltiple", PI09/91918 "SALEM: Segmentación Automática de Lesiones de Esclerosis Múltiple en imágenes de resonancia magnética" awarded by the Instituto Carlos III, and the VALTEC09-1-0025 "Salem: toolkit para la segmentación automática de lesiones de esclerosis múltiple en resonancia magnética" awarded in 2009 by the Generalitat de Catalunya

within the "Projectes de valorització VALTEC".

In the SALEM project, we proposed to develop and validate an automatic system for the detection, segmentation and description of MS lesions based on computer vision techniques. In particular, we aim to develop a computer aided tool in order to automatically detect and segment the MS lesions in MRI images, and to provide quantitative and qualitative descriptions for each patient. On the other hand, the AVALEM project aims to automatically evaluate and quantify the atrophy in the patients and their evolution in time.

This research has been carried out in close collaboration with the Dr. Josep Trueta and Vall d'Hebron Hospitals and Clínica Girona. The tools developed have been exhaustively tested and evaluated in the hospital centers involved in the project, which are reference centers in Catalonia within the multiple Sclerosis research field.

The goal of both projects is twofold: to create a novel dataset with imaging data from hospitals and to study and develop techniques to detect and segment new MS lesions that can be passed to experts for clinical use in evaluating the evolution and quantification of MS lesions. Within these projects, for which this PhD was the starting point of research, there has been a strong relationship with medical expert teams in the field of multiple sclerosis. Specifically:

- 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 several research projects funded by public and private institutions in the last few years, Dr. Pareto and technicians Huerga and Corral. This group is part of the MAGNIMS network, a European network of centers that share an interest in the MS study through MRI.

- From the Clínica Girona: 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.

- From the Hospital Dr. Josep Trueta: Dr. Ramió-Torrentà, who is the current coordinator of the "Unitat de Neuroimmunologia i Esclerosi Múltiple", as well as radiologists. Quiles, Valls and Beltrán, who work in the radiology unit. The relationship with this hospital arose from a previous collaboration with Dr. Gich within the EM-Line project that studied MS rehabilitation through interactive activities and games.

1.11 Objectives

As part of the SALEM framework, this PhD thesis' main goal is

the proposal of a new pipeline capable of detecting new MS lesions in serial brain magnetic resonance imaging.

This objective refers to the the detection and quantification of new MS lesions as well as lesion load change using time-series sequences of brain MRI.

This general goal can actually be divided into several sub-goals focused on the different stages of this thesis. The first goal is to provide a comprehensive state of the art of MS lesion detection and quantification methods in serial brain MRI. This objective aims to review the current MS lesion detection and change detection of these strategies with an eye detecting new MS lesions in order to better understand the advantages and drawbacks. With this analysis, we have seen that change detection techniques suffer from several artifacts as well as registration errors, therefore, tissue classification, using multi modal information and some post-processing steps are particularly necessary in order to reduce false positive lesions. These post-processing steps can be either supervised or unsupervised.

Following this idea, our second goal is to establish an automated general framework for automatic detection of new MS lesions. As we chose our change detection approach based on a subtraction pipeline, this framework should also cover a reliable validation approach, pre-processing steps to reduce image artifacts, noise and bias error caused by the scanner, a registration method to bring consecutive MR images into the same space, tissue segmentation for focusing on WM lesions, and finally, post processing steps to refine detected lesions.

The preprocessing steps can be divided into five main groups: noise reduction due to the capturing process, the correction of the bias field inherent to this image modality, intra-subject intensity normalization by means of histogram matching, skull stripping to remove non-brain tissue that can bias segmentation results and WM tissue segmentation. After the registration process, subtraction between consecutive images is used to find the areas of change. The subtraction image obtained undergoes a thresholding step in order to find any candidate lesions. Afterwards, multi-modal information can be used to refine the MS lesion candidates. Finally, either using prior knowledge or statistical methods, some

post-processing steps are applied to the pipeline in order to reduce more false positives and refine the lesion detection.

We validate our pipeline with real data, obtained from the hospitals, which have different time intervals (12 and 48 months) and lesion loads. We propose and prepare a reliable ground truth, which focuses only on the presence of new MS lesions prepared by using the manual annotations of the experts on the follow-up image. We also evaluate the validation spaces to determine whether they affect the validation accuracy.

1.12 Document structure

This thesis is structured as follows:

- **Chapter 1. Introduction.** This chapter presents the background, objectives and planning of this thesis project.

- **Chapter 2. A review of automated detection of MS lesions in serial brain MRI.** After stating the problem in chapter 1, we will review the most recent techniques dealing with this problem, focusing on advantages and drawbacks. A classification of the approaches for automatic monitoring of MS lesion evolution and quantification will also be introduced, emphasizing the approaches for detecting new MS lesions. Finally, the results will be gathered, along with the most common evaluation measures followed by our conclusions.

- **Chapter 3. Temporal analysis proposal on MS lesion detection** After the review in chapter 2, a multi-modal change detection strategy based on a subtraction pipeline for automatic detection of new multiple sclerosis lesions in longitudinal studies is proposed and analysed in detail. We provide an overview of the most important steps: preprocessing steps such as skull stripping and bias field correction; registration of the images, white matter segmentation, automated thresholding of the subtraction images, refining detected lesions via supervised or unsupervised methods using lesion features obtained from different sequences and determining a reliable validation method with the real MR images of MS patients.

- **Chapter 4. Experimental results.** The methods implemented will be tested and evaluated with real data using common similarity measures. In this chapter, we present our results, pointing out strenghts and weaknesses. We will also present a

comparison of the state of the art methods and a discussion of the results obtained, pointing out the important aspects of the proposed contributions.

- **Chapter 5. Conclusions.** In this final chapter, conclusions summarizing the work developed are presented. Based on these conclusions, possible improvements are also introduced as future work.

Chapter 2

A review of automated detection of MS lesions in serial brain MRI

2.1 MS lesion detection

This chapter presents a review of approaches that deal with time-series analysis of brain MRI to detect active MS lesions and quantify the lesion load change. We provide a comprehensive reference source for researchers in which several approaches to change detection and quantification of MS lesions are investigated and classified. We also analyze the results provided by the approaches, discuss open problems and point out possible future trends.

Conventional magnetic resonance imaging (MRI) techniques, such as T2-weighted (T2-w) and gadolinium-enhanced T1-weighted (T1-w) sequences, are highly sensitive in detecting MS plaques and can provide a quantitative assessment of inflammatory activity and lesion load. MRI-derived metrics have become the most important paraclinical tool for diagnosing MS, understanding the natural history of the disease and monitoring the efficacy of experimental treatments [133]. Quantitative analysis have become invaluable in the assessment of the disease's progression [103, 104, 75] and activity [124] and the evaluation of therapies over the last 25 years [42, 22]. Figure 2.1 shows two scans (T1-w, T2-w, and FLAIR images) of a damaged brain taken with a year's difference, together with the manual annotations made by an expert. The last column in the figure illustrates the total 3D lesion load in the baseline exploration and the new lesions appearing in the follow-up scan.

While there are many articles focusing on the lesion detection problem, most do not incorporate an automated method to interpret the lesion's evolution. The most common approach to the detect changes in serial imaging is visual inspection, which is typically

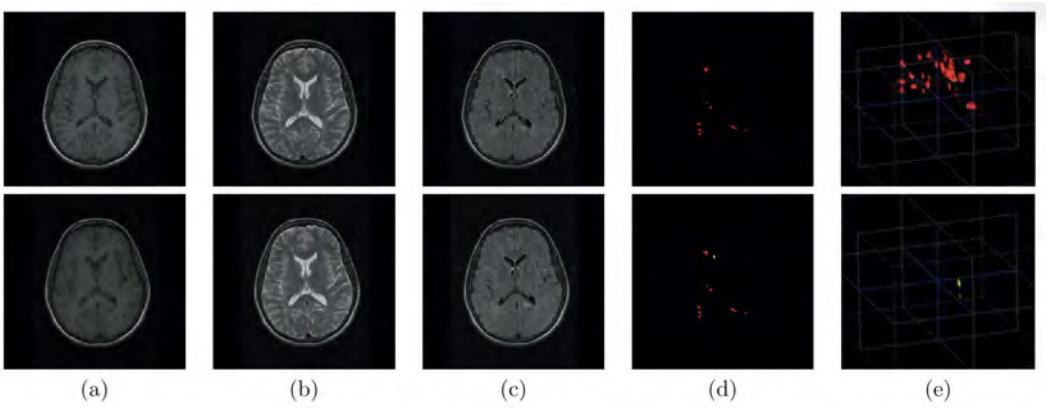


Figure 2.1: An example of MS lesion serial analysis. The upper row shows a slice of the baseline control, while the lower row shows the corresponding slice from the following exploration, made 12 months later. (a), (b), and (c) show, respectively, T1-w, T2-w, and FLAIR images. (d) shows the manual lesion annotations of the slices performed by an expert radiologist. In the baseline exploration all the lesions are annotated (in red) while in the following one only new lesions are marked (in green). Finally, the upper image of (e) represents the 3D lesion load in the baseline exploration, while the lower image shows the 3D representation of the new lesions in the follow-up exploration.

performed manually by experts [93]. The processed data, i.e already detected lesions, are presented to radiologists in order to obtain a decision with respect to the lesion load change [104, 96]. Experts use their anatomical and prior knowledge to identify lesion and its evolution. The manual detections of lesion and any change however, are not only time-consuming, but are also prone to intra-observer and inter-observer variability [111]. Although automated lesion detection techniques reduce this disagreement, an automated change detection method is still necessary to increase diagnostic precision [34, 35]. Moreover, it has been established that automated systems may outperform any human expert. For instance, as reported by Bosc et al. [17], that while many small and subtle changes in lesion evolution were missed by the expert, the automated change detection algorithm did not. Therefore, we believe that a comprehensive summary of the literature on automated lesion detection and quantification is important for researchers who want to improve upon previous work or develop new automated methods for progressive neurological disease analysis.

Change detection techniques can be divided into two categories: methods considering large structural changes and methods for smaller, more localized changes [17]. In accordance with this classification, lesion detection and quantification methods involve algorithms that must consider both small and large localized structural changes (i.e. tu-

mors). General problems associated with these techniques are the lesion's shape, which is usually ambiguous and has ill-defined boundaries, and the lesion's position, since the lesion can appear or disappear arbitrarily and may shrink or enlarge over time. In addition, their growth rates are not well characterized, and there can be great similarity between lesions and normal tissues, so they may not always be easily distinguishable. Moreover, 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*) [122]. Thus, observing the lesion's evolution without change in intensity but with displacement in the surrounding tissues (deformation) is more difficult. 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. Furthermore, detecting real image changes is hard work due to noise and residual artifacts in MR images, and also because the images of a patient at different times are not always directly comparable due to patient movement. In many cases, a robust image registration algorithm must be used [49, 74, 153]. Notice that in this case the quantification accuracy will depend on the alignment's accuracy [99]. Therefore, change detection techniques should be tuned to these facts accordingly.

Numerous approaches to lesion detection and quantification have been proposed in the literature [45, 99, 17, 94, 146, 86]. Despite the variety of approaches, none provide a fully automatic procedure that includes all the required steps for the diagnosis and treatment follow-up. For instance, some of the studies that introduce automated methods for lesion detection, typically based on segmentations, do not always provide an automated method for quantifying the lesion's evolution [137, 5, 4]. On the other hand, some of the studies that focus on change detection do not always provide an automatic lesion detection method and need user interaction to locate lesions [122] because they are not good enough to segment lesions after the change detection [99]. Furthermore, some of the change detection algorithms provide only a resulting image which then has to be interpreted visually by experts [119, 85], and a final expert decision is required to assess the lesion's evolution [17]. Note also that the change detection algorithms do not cover the detection of static lesions. Combining the advantages of different techniques may compensate some of the missing elements in some strategies and may enable the development of less subjective and more automated approaches.

The aim of this chapter is to point out the capabilities of the approaches developed and provide an up-to-date state-of-the-art review of automated MS lesion detection and

quantification methods in serial MRI. Furthermore, we classify the different techniques according to the strategy used, as well as describing the most representative studies in this field. We analyze numerous articles and provide a detailed classification of lesion detection and change detection techniques based on the main characteristics of each strategy, pointing out the challenging parts of each method. In addition to introducing and classifying these approaches, we also describe the algorithms used to detect and quantify the lesions as well as the features and type of MR images used. Furthermore, we compare the results of the studies analyzed in terms of accuracy and robustness.

Few articles have reviewed MS lesion detection and quantification methods in brain MRI serial analysis. For instance, Patriarche and Erickson [93] provided a review of the change detection techniques in time-series analysis. However, this review did not particularly focus on the purpose of MS lesion detection. Bosc et al. [17] also provided a simple classification of inter-image comparisons considering lesion evolution. Nevertheless, this study was not a complete review. Recently, Mortazavi et al. [87], Lladó et al. [71], and Garcia et al. [41] have presented a review of MS lesion detection at a single time point, without taking into account change detection, lesion evolution, or quantification. Even though some articles have given information about either MS lesion detection or lesion evolution quantification methods [112, 94, 64, 58], none have proposed a comprehensive review. Furthermore, none of them tried to quantitatively compare the results, as it would be difficult to guess the performance of all these detection and quantification approaches. Ideally, methods should be applied to a common database and compared to a ground truth. This, however, is very difficult due to the lack of common public databases of real image scans at different time-points along with their ground truth and the fact that only a few methods are publicly available. Here we will quantitatively compare the detection approaches accordingly to their reported results in the literature. We will describe the most typical measures used for evaluating MS lesion detection and quantification in time-series MRI, comparing in a qualitative and quantitative way the results of the studies analyzed. As a consequence, we review the most relevant studies in the time-series analysis from both an MS detection and quantification point of view and which also provides an evaluation of the experimental results.

2.2 Classification of MS lesion detection and quantification in serial brain MRI

In this section we propose a classification to categorize the state-of-the-art automated serial MS detection and quantification methods in time-series analysis. Afterwards, we also analyze the general problems encountered in the segmentation and quantification processes.

2.2.1 Proposed classification of lesion detection approaches

In order to classify the MS lesion detection approaches, we considered the different classifications proposed by Bosc et al. in 2003 [17] along with the one proposed by Patriarche and Erickson in 2004 [93]. From this starting point, and also from the information collected from the newest studies [112, 64, 94, 58], we propose a new classification of the categories and subcategories shown in Figure 2.2. In particular, we classify the detection approaches in two primary categories, according to their main principle and characteristics:

- *Lesion detection methods.* We consider lesion detection methods to be those that aim to detect both static and dynamic MS lesions in a single time MR volume of a patient. These segmentation-based methods, which can be supervised or unsupervised algorithms, rely on the intensity homogeneities of the tissues and typically apply data mining techniques (clustering, classification) to distinguish lesions from normal tissues. In time-series analysis, the use of segmentation-based methods mostly involves a subsequent lesion quantification approach that computes the volumetric changes of each segmented lesion between two time points in order to determine the MS lesion's evolution.

- *Change detection methods.* These approaches are not based on the analysis of a single time point (one control of a patient) but rely on analyzing the differences between successive MRI controls at both a 2D and 3D image levels. From this classification, we further subclassify the main strategies. The *intensity based methods* consist of analyzing two successive scans by means of subtraction techniques. Among these methods, we further distinguish between *deterministic* approaches, which typically cover the subtraction methods using direct intensity differences between the scans, and *statistical* approaches, which are used for compensating the interpretation problems of point-to-point comparison. The *temporal analysis* approaches are based on

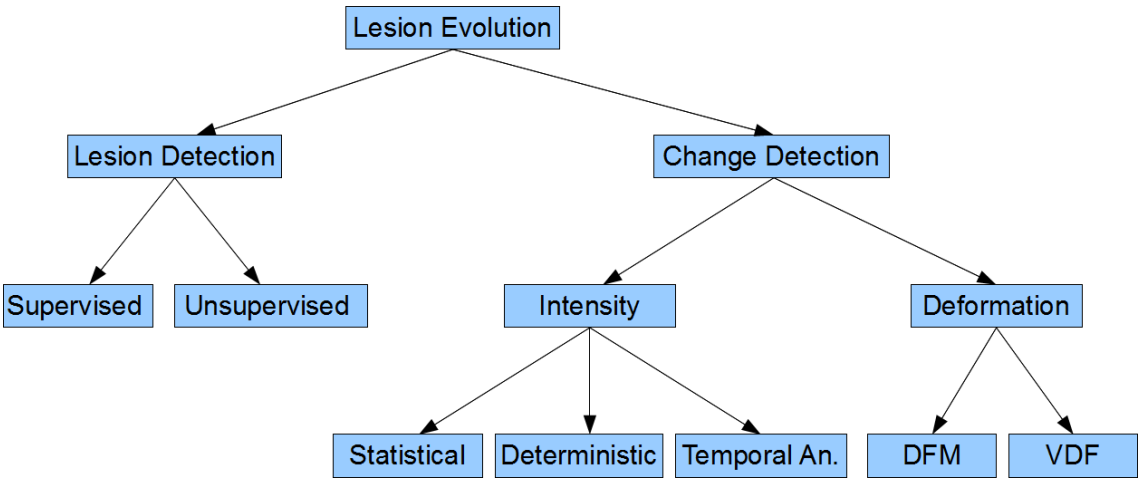


Figure 2.2: Proposed classification of MS lesion serial analysis. We clearly distinguish between lesion detection and lesion change detection techniques. The acronyms VDF and DFM stand for vector displacement field and deformation field morphometry, respectively.

detecting active voxels through a time-series analysis of more than two successive scans. Finally, the *deformation-based* approaches aim to obtain a deformation field from a non-rigid registration process between successive controls that can be used directly to perform the lesion detection and evolution. We have subclassified these approaches according to the way the deformation field is used: *vector displacement field* and *deformation field morphometry*. Note that, depending on the technique used, these approaches may or may not require a subsequent analysis of the quantification.

The approaches reviewed in this study are summarized in Table 2.1, which offers a compact, at-a-glance overview of these studies. Moreover, the most important features and properties of all the approaches have also been taken into account. Namely, the main characteristics of each approach analyzed: The detection strategy and quantification algorithm used, the type of automation (semi automated or fully automated), whether the method uses a template (an atlas) to improve the accuracy, such as a template driven segmentation (TDS), or methods that use healthy control images to compare and correct their results. Finally, we have also included the image types used (T1-w, T2-w, PD-w, FLAIR, etc) and the lesion types the method can deal with. It should be noted that not all the studies analyzed always specify the particular type of lesion.

Table 2.1: Classification of the lesion evolution methods. The different acronyms refer to: **DETECTION METHODS:** **FCS:** **F**uzzy-**C**onnectedness **S**egmentation, **KNN:** **K**-**N**earest **N**eighbor, **ANN:** **A**rtificial **N**eural **N**etwork, **EM:** **E**xpectation **M**aximization, **PVEC:** **P**artial **V**olume **E**ffect **C**orrection, **SVM:** **S**uport **V**ector **M**achines, **FCM:** **F**uzzy **C**-**M**ean, **SDF:** **S**tructure **D**ifference **F**iltering, **SNM:** **S**tructure **N**oise **M**ap, **GLRT:** **G**eneralized **L**ikelihood **R**atio **T**est, **LVR:** **L**ocal **V**olume **R**atio, **RFC:** **R**andom **F**orest **C**lassification, **RLF:** **R**egion **L**evel **F**iltering, **STPC:** **S**ingle **T**ime **P**oint **C**lassification, **GLCM:** **G**ray **L**evel **C**o-occurence **M**atrix, **RL:** **R**un **L**ength Matrix, **AR:** **A**uto **R**egressive Model, **WA:** **W**avelet **A**nalysis Model, **QUANTIFICATION METHODS:** **VI:** **V**isual Inspection, **SCD:** **S**tatistical **C**hange **D**etection, **4DCCA** **4D** **C**onnected **C**omponent **A**nalysis. **LESION TYPES:** **WML:** **W**hite **M**atter **L**esion, **WMSA:** **W**hite **M**atter **S**ignal **A**bnormalities, **GML:** **G**ray **M**atter **L**esion, **GEL:** **G**adolinium **E**nhancing **L**esion, **FCDL:** **F**ocal **C**ortical **D**ysplasia **L**esion.

			References	Detection	Quantification	A	Atlas	Sequences	Lesions
Lesion Detection	Supervised		[Udupa, 1997]	FCS	Volumetric	SA	-	T2;PD	WML
			[Warfield, 2000]	KNN	-	A	✓	T2	WML
			[Zijdenbos, 2002]	ANN	Volumetric	A	-	T1;T2;PD	WML
			[Wei, 2002]	Self Adaptive EM & PVEC	Volumetric	A	✓	T2;PD	WMSA
			[Ashton, 2003]	Bayesian	Volumetric	SA	-	T1;T2;PD	WML
			[Meier, 2003]	Self Adaptive EM & PVEC	Temporal Analysis	A	✓	T2	WMSA
			[Antel, 2003]	GLCM Features & Bayesian	-	A	-	T1	FCDL
			[Anbeek, 2004]	KNN	-	A	-	T1;T2;PD;FLAIR;IR	WML
			[Wu, 2006]	KNN	-	A	✓	T1;T2;PD;	WML & GEL
			[Duan, 2008]	PVEC & Thresholding & Manual	Volumetric	SA	✓	T2;PD	-
			[Zacharaki, 2008]	SVM	Volumetric	A	-	T1;T2;PD;FLAIR	WML
			[Shen, 2008]	FCM	Volumetric	A	✓	T1	IL
			[Zhang, 2008]	GLCM & RL &AR & WA Features & KNN & ANN	-	SA	-	T2	WML
			[Shiee, 2010]	FCM	-	A	✓	T1;T2;FLAIR	WML
			[Yamamoto, 2010]	Level Sets & SVM	-	A	-	T1;T2;FLAIR	WML
			[Cerasa, 2011]	ANN	-	A	-	FLAIR	WML
			[Geremia, 2011]	Random Decision Forest	-	A	-	T1;T2;FLAIR	WML
			[Rode, 2012]	GLCM Features & SVM & ANN	-	SA	-	-	WML
	UnSupervised		[Ettinger, 1994]	EM & Subtraction	Volumetric (4D-CCA)	A	-	T2;PD	WML
			[Lee, 1998]	Thresholding & Subtraction	Volumetric	SA	-	T2;T1	GEL
			[Guttmann, 1999]	EM & PVEC	Volumetric (4D-CCA)	A	-	T1;PD	WML
			[Kikinis, 1999]	EM & PVEC	Volumetric (4D-CCA)	A	-	T1;PD	WML
			[Weiner, 2000]	EM & PVEC	Volumetric (4D-CCA)	A	-	T1;PD	GEL
			[Hillary, 2009]	ISODATA	Volumetric	A	-	T1;FLAIR	-
			[Duan, 2008]	Thresholding & Manual	Volumetric	SA	-	T2;PD	-
			[Juang, 2010]	Histogram-Based Classification	VI	A	-	T2;T1	Tumor
Change Detection	Intensity	Det.	[Curati, 1996]	2D Subtraction	VI (Manual)	SA	-	T1	-
			[Tan, 2002]	2D Subtraction	VI (Manual)	A	-	T2	-
			[Moraal, 2009]	2D Subtraction	VI (Manual)	A	-	T1;T2;PD	WML
			[Moraal, 2010]	3D Subtraction	VI (Manual)	A	-	FLAIR;DIR;MP-RAGE	WML
			[Battaglini, 2013]	3D Subtraction & RLF	Volumetric(region-wise)	A	-	T1;T2;Pd;FLAIR	WML
		Stat.	[Lemieux, 1998]	2D Subtraction & SDF & SNM	VI (SCD)	A	-	T1	-
			[Bosc, 2003]	2D Subtraction & GLRT	VI (SCD)	A	-	T1;RARE;FLAIR	-
			[Elliot, 2013]	3D Subtraction & Bayesian (STPC) & lesion-level RFC	Volumetric(region-wise)	A	-	T1;T2;PD;FLAIR;T1c	-
			[Sweeney, 2013]	3D Subtraction & logistic regression	Volumetric(voxel-wise)	A	-	T1;T2;PD;FLAIR	-
		Tmp	[Gerig, 2000]	Temporal Analysis	Volumetric	A	-	FMRI	WML & GML
			[Welti, 2000]	Spatio Temporal Analysis	-	A	-	T1;T2;PD;FLAIR	WML & GML
	Deformation	VDF	[Thirion, 1999]	Norm and Divergence of Vector Fields	Warping	A	-	T2	-
			[Rey, 2002]	Flow Field & Jacobian Operator	Warping	A	-	T2;PD	-
		DFM	[Pieperhoff, 2008]	Flow Field & LVR	Warping	A	-	PD	-

2.2.2 Proposed classification of lesion quantification

As well as classifying the MS lesion detection approaches, we have categorized the methods according to the quantification of the lesion's evolution. Note that this quantification process is essential for radiologists and neurologists to analyze the patients' follow-ups [96]. We classify the quantification approaches into three main categories: *visual inspection*, *statistical change detection*, and *volumetric approaches*.

The visual inspection is a manual method to determine a lesion's evolution. The processed images, such as registered images or subtracted images, are analyzed and visually interpreted by a user or expert in order to arrive at a decision. Although this is a very subjective method, some improvements can be made to reduce the number of misinterpretations made by an expert. For instance, statistical change detection techniques using statistical correction [17] or structured noise maps [68] in order to reduce false positives in the subtracted images may be applied. In a different way, the volumetric approaches typically use already segmented lesions in order to quantify the lesion' evolution by means of its volume changes. These volumetric quantification approaches have proven to be useful in detecting positive and negative disease activity [140]. Notice that this quantification process can be done by either subtracting single lesion volumes or total lesion volumes between the time-series images. However, notice that when computing the total MS lesion volume of a patient, it is possible that some lesions will enlarge while others shrink. Therefore, this quantification process may not detect a change in lesion volume even if there are growing and shrinking lesions. As a result, comparing lesion volumes individually seems a more precise way of doing the quantification. Furthermore, when using volumetric measures one should note that the process relies on the results of a previous segmentation method that might not provide the desired result and introduce errors in the quantification. Note that we could also add the temporal analysis and the warping methods, which were also included as detection approaches, to this quantification classification. In fact, these methods produce the detection and quantification of the MS active lesions in a single step. For instance, the main property of the warping algorithms (also known as deformation field based approaches [17]) is that they are based on a one-to-one tissue correspondence and, as well as providing lesion detection, they allow the lesion mass effect to be quantified from the registration process between temporal studies of a patient.

2.2.3 General problems in MRI

Image intensities between corresponding tissues or structures in successive scans may differ. Thus, normalization algorithms are used to compensate global intensity changes between successive images before and/or after the registration processes [17]. This normalization process improves the alignment between images if used before the registration step and also allows a better comparison between the tissues analyzed, structures and lesions if used after registration. For instance, Bosc et al. [17] used a linear intensity normalization algorithm before each registration step and a non-linear joint normalization algorithm after registration.

Another well-known issue when processing MRI images is that noise and artifacts may be present due to the scanner's performance and may affect the detection and quantification accuracy [68]. For instance, Guttman et al. [46] and Kikinis et al. [61] used a non-linear anisotropic diffusion filtering, an edge-preserving noise reduction method, to overcome this problem. With a similar strategy, Bosc et al. [17] applied a low-pass Gaussian filter to the images obtained by subtracting successive registered and normalized images to eliminate residual artifacts such as radio-frequency (RF) artifacts [17].

Besides these difficulties, partial volume effects, where a single voxel contains a mixture of multiple tissue values, generally occur in medical imaging. This situation is particularly true for voxels on the boundaries [30] or brain surfaces that contain both brain tissue (skull bone) and cerebrospinal fluid [61] due to the particular intensity characteristics of PD-w and T2-w images. Thus, regions with similar intensity values to the lesions may introduce errors into the quantification process. Several approaches have been proposed to deal with this issue. For instance, some methods use a priori anatomical knowledge [46] to eliminate spurious lesions selectively [61].

2.3 Lesion detection approaches

Image segmentation is the process of assigning a label to every voxel in a single image so that voxels with the same label share certain visual characteristics typically indicating a particular object, namely tissue or lesion. To our knowledge, segmentation based approaches cover the largest area of methods for MS lesion detection and are still the largest active area of research. Various methods have been proposed for this purpose [46, 152, 4, 36, 107, 58], and some attempts to classify these automated MS lesion segmentation approaches have been made. Mortazavi et al. [87] have recently presented a

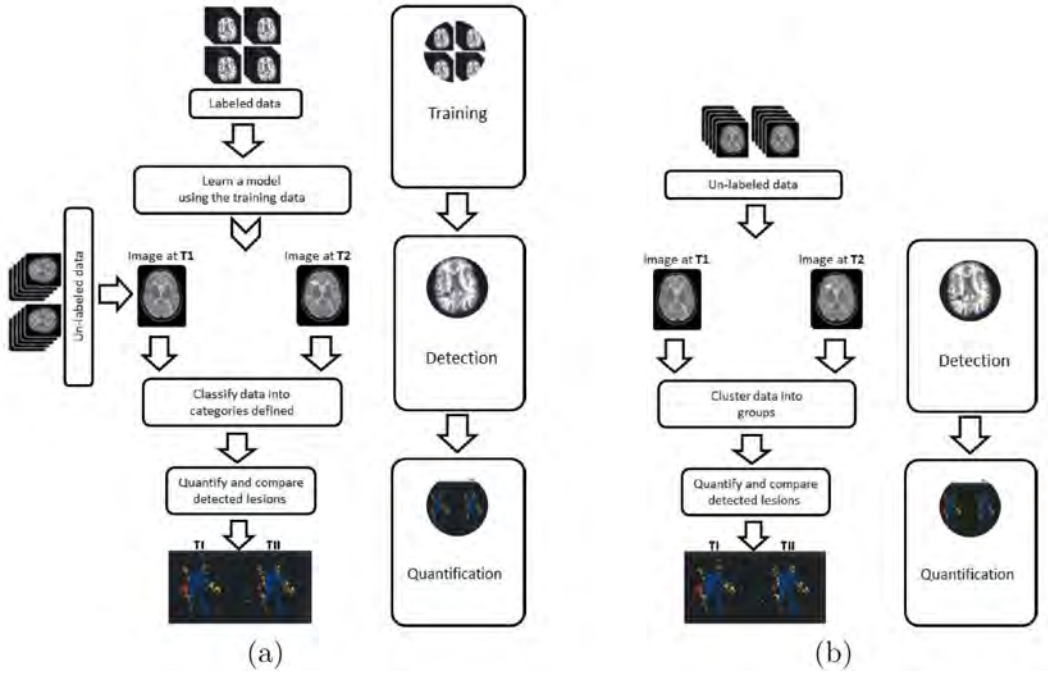


Figure 2.3: Flowchart of the lesion detection approaches: (a) supervised, (b) unsupervised. As clearly shown, the main difference between both strategies is the use or not of an initial training step.

review of the segmentation of multiple sclerosis lesions in MR images, providing a classification of the approaches reviewed into four different categories: data-driven, statistical, intelligent, and deformable methods. Even though numerous attempts have been made to solve this segmentation problem, due to the arbitrary shapes and locations of the lesions, automated segmentation is still an open issue and a challenging task [107].

Segmentation based approaches can be classified as manual outlining methods, semi-automated methods, and fully automated methods. In this study, we explore the fully automated methods, which do not require user interaction and reduce the intra/inter operator variability [112]. However, to provide a wider analysis, we also include some of the most relevant semi-automated methods [128, 65, 7, 36]. Figure 2.3 shows a flowchart of the general idea of the segmentation based approaches for brain MRI time-series analysis.

Automated MS lesion segmentation is a difficult task due to the similarity of intensity between lesions and normal tissues. For instance, gray matter lesions (GML) may share intensities with gray matter (GM) or cerebrospinal fluid (CSF) [107]. Thus, traditional segmentation methods, like region-based methods where the voxels are directly analyzed by means of a region growing strategy, or methods using thresholding techniques, may

not provide the desired results. Noise and residual artifacts also make lesion segmentation difficult, even for white matter lesions (WML).

Analyzing the literature, we have seen that tissue class segmentation based approaches (clustering methods) are commonly used for automated MS lesion segmentation. Tissue class segmentation, which uses tissue-class weights to consider the presence of lesions, can be considered as an estimation problem to determine intensity inhomogeneities [141]. These techniques use spatial information (position of the tissues and lesions) and inconsistency of lesions (intensity differences between lesions and normal tissue distributions) to detect and then quantify the lesions. Note that these are statistical segmentation approaches that use knowledge about tissue properties and, therefore, rely on the fact that the same tissues have the same intensity values.

Several techniques have been used to improve segmentation accuracy. It is well known that the use of prior knowledge of normal tissue distribution improves the capability of segmentation methods [139]. The main strategy is to use an anatomical template (atlas) to introduce spatial information into the statistical segmentation. Although this information can be introduced in different ways [21], the most common approach is based on template-driven segmentation (TDS), which mainly consists of a non-linear registration step [137, 114] to match MR images to the atlas. As reported by Warfield et al. [137], statistical classification and non-linear registration are often complementary since pathologic structures such as lesions are not modeled in an anatomical template. Lesions cannot be segmented directly with an anatomical template. Therefore, statistical methods are performed to compensate for this problem.

In addition to TDS, multi-spectral approaches are used to improve the segmentation's accuracy since different modalities of MR images (T1-w, T2-w, PD-w, FLAIR, etc.) have different signal characteristics that provide different information. However, multi-spectral anatomical images are not always available in clinical practice since the acquisition of all these images is cost intensive and requires more processing time [107]. Methods using multi-spectral information also require a registration step, which may be assumed to be an affine [107] or a deformable registration [137].

Furthermore, some studies also use a partial volume effect correction (PVEC) method to eliminate any false positive lesions detected [139, 46, 61, 36]. For instance, Guttmann et al. [46] and Kikinis et al. [61] applied a PVEC algorithm and improved their previous results. Moreover, Wei et al. [139] concluded that the PVEC algorithm eliminated only false positive errors while TDS corrected false negative misclassifications and some of the false

positive misclassifications. They also pointed out that using TDS with PVEC together showed the highest accuracy in the segmentation of white matter signal abnormalities (WMSA) [139].

2.3.1 Supervised methods

We consider those methods as supervised approaches that mainly use the image intensities of different MR images to train a classifier by using labeled tissues and manually identified lesions and those methods that use information from a template (atlas) to classify tissues and segment lesions as deviations from normal human brains. It can be seen in Table 2.1 that several techniques have been used to perform supervised classification. For instance, k-nearest neighbors (KNN), artificial neural networks (ANN), and support vector machines (SVM) are typical supervised approaches for tissue segmentations. Furthermore, Yamamoto et al. [145] have recently proposed a false positive reduction step that uses a level set method and a SVM classifier to substantially reduce the number of false MS lesion detections.

According to Udupa et al. [128], human experts usually outperform automated algorithms in the recognition task and, therefore, in their approach, brain tissues such as WM, GM and CSF are manually determined by an operator. They claim that automated algorithms conversely perform better in the delineation, hence, they used a fully automated algorithm for the delineation process from which they segmented the MS lesions based on the principle of fuzzy-connectedness [127] using the manually recognized brain tissues (WM, GM, CSF) as fuzzy connected regions. After the detection of CSF, WM and GM as 3D fuzzy objects, lesions appeared as "holes". The approach by Udupa et al. [128] can also be considered as an early multi-spectral approach, since they used both T2-w and PD-w images to classify brain tissues. They state that CSF tissue is better recognized in T2-w images whereas WM and GM tissues are better recognized in PD-w images.

Another semi-automated supervised and multi-spectral method was proposed by Ashton et al. [7]. They compared the regional-based methods (GEORG) with a directed multi-spectral segmentation (DMSS) approach, and concluded that both methods were acceptable in terms of speed and precision. They used statistical characteristics of background tissues supplied by a Bayesian classifier and target statistics supplied by the exemplar. This approach is also multispectral since they mapped the three T1-w, T2-w, and PD-w images to the red, green, and blue channel, respectively. Nevertheless, both algorithms need user interaction: a single mouse click was used to place a seed for a region growing

algorithm and a manually traced exemplar was needed for the classification method.

Warfield et al. [137] applied a TDS segmentation and a spatially varying statistical classification based on a multiple feature KNN classification process. Also based on the KNN classifier, Wu et al. [144] proposed an automatic segmentation of MS lesions into three subtypes: enhancing lesions, black holes and hyperintense lesions. An intensity-based statistical KNN classifier is combined here with an atlas segmentation to extract WM masks. Assuming that lesions are only found in WM regions, the authors discard all the lesions outside the masks. Moreover, partial volume problems (i.e. those arising from the fact that a voxel may be composed of more than one tissue type) are corrected using morphological operators. On the other hand, Wei et al. [139] and Meier and Guttman [78] included a template-driven strategy to perform the tissue class segmentation based on an expectation maximization algorithm. Meier and Guttman [78] also applied subtraction and partial volume corrections to identify lesion load changes and to eliminate false positive lesions. After the lesion segmentation, they combined space and time into the MS lesion characterization process via direct quantitative analysis of the signal intensity in the time domain obtained from serial MR images. In this way, they showed the signal dynamics of active and chronic MS lesions [78].

Zijdenbos et al. [152] proposed a supervised MS lesion segmentation method using multi-spectral information (T1-w, T2-w, and PD-w) using an artificial neural network (ANN). In particular, they used a back propagation ANN method to classify the MS lesions because of the reliability of the method under different imaging conditions. Similarly, Cerasa et al. [25] propose a technique to segment white matter lesions in MS patients by using a Cellular Neural Network (CNN) based approach. Unlike ANN, in a CNN, interconnections among cells are local, that is, each processing unit directly interacts only with its neighbouring cells located within a prescribed sphere of influence. The authors applied this CNN-based technique to automatically segment MS lesions in FLAIR images, comparing the performance of their approach with the manual segmentation provided by two expert radiologists. Moreover, Anbeek et al. [4] combined a supervised classification algorithm with a multi-spectral approach for white matter lesion (WML) detection. They used five different modalities (T1-w, T2-w, PD-w, IR and FLAIR) and applied a KNN classification technique. Likewise, Zacharaki et al. [146] recently presented a supervised WML segmentation method based on SVM. They applied an Adaboost algorithm to each of the scans. As they reported, WMLs had intensities similar to GM tissue in T1-w images, and similar to CSF in T2-w and PD-w images, so they applied a multi-spectral approach. Another multi-spectral supervised (T1-w, T2-w and FLAIR) approach was proposed by Geremia

et al. [43]. applying a discriminative random forest classification (RDF) to the MS lesion segmentation problem. They included knowledge on tissue classes and long-range spatial context in order to discriminate lesions from background.

Besides these techniques, texture analysis has also been proposed as an alternative strategy to identify active MS lesions [60]. First order-statistics (individual pixel values such as mean and variance of gray level), second order statistics (properties of pixel pairs) obtained primarily from gray level co-occurrence matrix (GLCM) and run-length matrix (RLM), and additionally, some spectral approaches (Fourier, Wavelet and Stockwell transforms) can be used for this purpose [150]. Zhang et al. [148] demonstrated that 9 particular features show larger differences when distinguishing NWM (normal white matter), NAWM (normal appearing white matter) tissues and MS plaques. They also pointed out that using a combined set of features provides a better performance than using single feature extraction approaches. Classifying a region of interest (ROI) based on texture features, KNN, ANN, Bayesian and SVM classifiers are widely used. For instance, Antel et al. [5] used texture feature maps obtained by using co-occurrence matrices together with a supervised classification based on the two-step Bayesian classifier to perform the MS lesion detection, whereas Zhang et al. [149] used KNN and ANN classifiers and, more recently, Rode et al. [102] used the SVM classifier. While Zhang et al. [149] demonstrated that ANN classified more accurately than the KNN, Rode et al. [102] found that the SVM classifier has a higher accuracy over ANN classifier. For these techniques, the selection of the ROI is crucial and is typically carried out manually by experts.

Alternatively, Shen et al. [107] identified MS lesions using their inconsistency by a defined threshold. They combined the fuzzy c-means (FCM) algorithm and TDS to create tissue probability maps. There are more examples of atlas-based approaches. The method proposed by Shiee et al. [108] segmented brain tissues in an iterative way, interleaving a fuzzy segmentation and defining topologically consistent regions. MS lesions were identified as dark holes inside the WM. The authors used multi-channel images to segment the major structures of the entire brain. Basically, their method is an atlas-based segmentation technique employing a topological and statistical atlas, together with the FCM algorithm to perform the classification. As reported by Shiee et al., the advantage of using the topological atlas is that all the segmented structures are spatially constrained, thereby allowing subsequent processing to perform cortical reconstruction and unfolding.

One of the drawbacks of the supervised segmentation is that the accuracy may depend highly on the selection of the training set and the control groups [107] used to compare

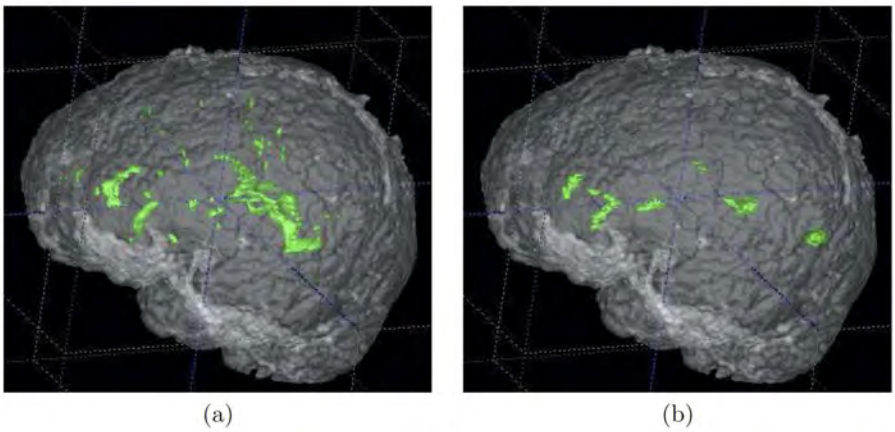


Figure 2.4: Generated 3D volume with MS lesions segmented by two different experts showing a large inter-rater variability. Note here the importance of using more than one manual annotation when evaluating the automatic algorithms [71].

individual patient images to a normal control group (model-based strategy) [114]. Gerig et al. [44] compared a clustering technique (ISODATA (Iterative Self Organizing Data Analysis Technique)) with a supervised classification (parametric maximum likelihood classification and Parzen window technique) for brain MR images and found similar estimated parameters. Furthermore, although supervised methods are more efficient for the segmentation purpose, they require some user interaction for the training steps. Besides, different users or trainings at different times with the same data may produce different results. Figure 2.4 shows an example of a MS patient volume segmented by two different experts. Thus, unsupervised methods are less subjective, completely automated and more reproducible with respect to supervised classifications.

2.3.2 Unsupervised methods

As illustrated in Table 2.1, many of the unsupervised classification methods [38, 46, 61, 140] use the expectation maximization (EM) algorithm [142]. For instance, Ettinger et al. [38] combined statistical tissue classification based on the EM algorithm and subtraction in order to detect positive and negative changes. In a similar way, Guttmann et al. [46], Kikinis et al. [61] and Weiner et al. [140] used a similar strategy to segment MS lesions based on tissue classification and expectation maximization.

Lee et al. [65] used a local threshold defined by a single observer in order to segment MS lesions. Areas of new lesions and areas of resolving lesions were defined by subtracting normalized and co-registered images. They labelled the lesion areas with color, and

subtracted two successive images. The outcome image yielded a colored subtraction map that indicated areas of new lesions and areas of resolving lesions.

More recently, Duan et al. [36] compared two different approaches called conventional image segmentation (CSEG) and segmentation of subtraction image (SSEG). The first was a supervised approach due to the use of TDS, while the second was an unsupervised approach using the intensities of the subtracted images to detect the MS lesions. However, both segmentation methods are refined by applying an automated Otsu threshold and manual editing. The authors concluded that the SSEG method provided a significantly higher measurement of reproducibility and enhanced sensitivity to cortical and subcortical lesions.

Hillary et al. [50] used an ISODATA technique consisting of a multi-parametric unsupervised segmentation method. Jacobs et al. [54, 53] applied ISODATA technique for MRI tissue characterization in clinical stroke. As a different approach to classification methods, Juang and Wu [58] applied color-based segmentation with k-means clustering based on applying a histogram based metric to produce colored images indicating tissues and lesions.

2.4 Change detection approaches

As we have already stated before, the patient's follow-up over time is crucial to determine the evolution of the disease. Therefore, change detection techniques are needed to compare the brain's evolution over time. As shown in Figure 2.2, we distinguish between three main different strategies to perform these tasks, which are described in the following subsections. Figure 2.5 shows the different flowcharts of each category.

2.4.1 Intensity-based approaches

Intensity-based approaches for change detection use voxel-to-voxel intensity comparison to distinguish evolving lesions. Therefore, a lesion without changes in the follow-up scan, i.e static lesions, cannot be detected using this strategy.

Voxel-to-voxel comparison methods usually suffer from repositioning errors due to patient movement, inconsistent objects over time such as blood and cerebrospinal fluid flow artifacts [85], noise in the images, and partial volume effects [30]. Therefore, image registration, bias field correction, intensity normalization [106], and using multisequence in-

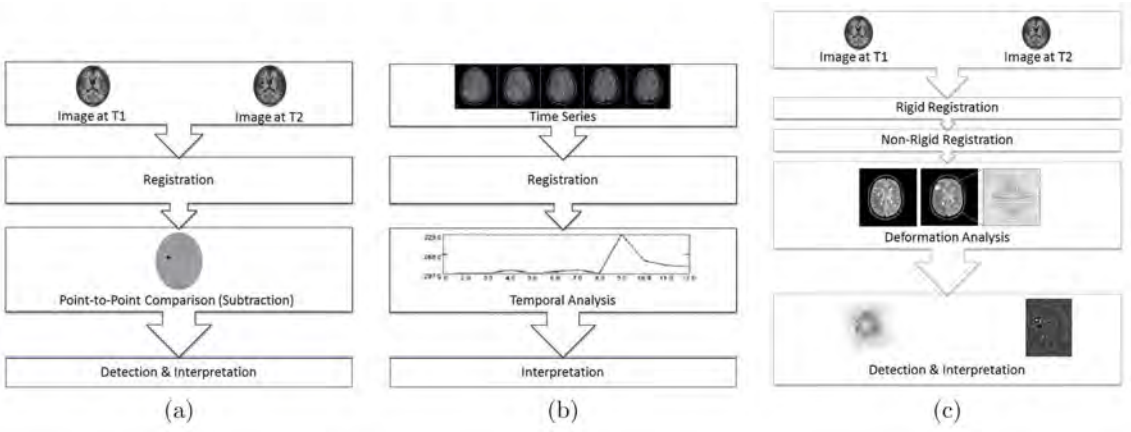


Figure 2.5: Flowchart of the main change detection categories. (a) intensity-based techniques, (b) temporal analysis, and (c) deformable approaches.

formation are necessary to compensate for these problems. Furthermore, the selection of image type (T1-w, T2-w, PD-w, FLAIR, MP-RAGE) and the interpolation method during the registration process are also important criteria for the accuracy and robustness of the subtraction methods.

Deterministic approaches

We include in this group those intensity-based approaches that are based on subtracting two successive images in order to find intensity differences due to evolving lesions. Typically, after the subtraction of two consecutive temporal images, positive activity (new or enlarging lesions) appears as hyperintense areas while negative activity (resolving or shrinking lesions) appears as hypointense areas when compared to the background [119].

The roots of the subtraction approach to detect MS lesions were made by Curati et al. [30], who investigated contrast enhancement with registered difference images. They reported that the recognition of small changes, changes at the boundaries, and tissues and fluids with very high or very low signals were more difficult to determine. Furthermore, they noted that while the use of thin slices decreased the partial volume effect, it increased the misregistration. Thus, they stated that an accurate alignment was necessary to assess change. They also claimed that using 3D scans of MP-RAGE images might increase the accuracy of the results since these types of scans have better contrast.

Tan et al. [119, 120] suggested that using only the variation in the intensity signal to determine negative or positive activity was not sufficient, since change in the intensity

signal may also be due to different conditions such as the use of a different scanner or a high level of noise. Thus, they determined regional activity by also checking if there was a change in the lesion's size or shape. They concluded that using subtracted images for lesion detection showed better agreement for positive activity than for negative. Besides, they reported that the success of this approach highly depends on the lesion's size. To detect enlarging lesions smaller than 5 mm in diameter, they must increase their size by more than 100%. On the other hand, the detection of shrinking lesions with a diameter smaller than 5 mm was not reliable.

Following a similar approach, Moraal et al. [85] concluded that subtracted images provide a sufficient measure for the quantification of positive disease activity. The authors found a good inter-observer agreement in the quantification of positive disease activity and compared their results with previous studies in terms of inter-observer agreement, concluding that their success was due to the improvements in the registration and intensity correction methods used. They also noted that results obtained for the negative activity were not as good as the results obtained for the positive disease activity. In a different study, Moraal et al [86] evaluated the performance of 2D and 3D subtraction methods, and concluded that 3D subtraction techniques, after image registration, provided greater inter-observer agreement. Furthermore, they compared several image sequences (3D DIR (double inversion recovery), 3D FLAIR, 3D T2-w, 3D MP-RAGE) and found that negative active lesions, even small ones, could be detected using the 3D MP-RAGE images, owing to good anatomical detail and clear GM-WM contrast. More recently, Battaglini et al. [11], proposed using an overestimated mask of candidate lesions obtained by applying a low-intensity threshold to the subtraction image. Specifically, the obtained hyperintense voxel clusters are filtered using a set of specific constraints for shape, size and intensity to provide the final detections.

Statistical approaches

Statistical change detection techniques for interpreting intensity differences aim at reducing the noisy results obtained by direct point-to-point subtraction [17]. This group of methods is based on building a statistical model of intensity changes between successive scans in order to detect active lesions and their evolutions. These methods rely on changes in the lesions and not on changes in individual voxels.

For instance, after the image subtraction, Lemieux et al. [68] classified each voxel as changed versus unchanged according to a threshold value, and subsequently grouped to-

gether the changed voxels. They called these grouped voxels structured differences objects, which can be caused by either biological processes or image artifacts. Afterwards, in order to quantify changes in the image difference, these structured difference objects were thresholded by applying the structure difference filtering that was used to estimate the Gaussian noise level. After the normalization, the outcome image was a map of the voxel classification indicating no signal change, signal increase, signal decrease, or outside of the brain. The authors also compared this map with the one obtained by a set of normal volunteers in order to assess the significance of the changes. By using this full scheme, they avoided the structured noise, and were able to determine real changes more correctly. However, note that this statistical method cannot directly give the total count of active lesions, although a set of statistics, such as the total genuine change voxels and total number of normal structured voxels, can be easily obtained.

On the other hand, Bosc et al. [17] presented both a single-modal and multi-spectral (FLAIR, RARE, and MP-RAGE) change detection approach. They registered the images into a common reference according to their modality, instead of choosing a baseline reference from the serial images of a patient, since the registered images undergo geometrical transformations while the reference image does not. In this sense, all the images undergo equivalent processing steps, as is done in the study of Moraal et al. [86], using the well-known half-way registration [56, 55]. Affine registration was used to register the single modality matching while affine and deformable registration was used for the multi-modality matching. Afterwards, they computed the voxels probability ratio of change, and grouped any neighboring changed voxels together. Thus, clustered voxels (also sorted in decreasing likelihood) were presented to the experts instead of individual voxels. Notice that evaluating individual voxel changes is more difficult, and also, manually delineating the lesion's evolution is more subjective. They evaluated their results with simulated lesions and found that lesions with a radius greater than 0.6 voxels could be detected. Furthermore, they found that the multi-modality detection increased the detection probability from 79% to 95% due to the richer information and avoids a lot of false positive detection.

More Recently, Elliott et al. [37] and Sweeney et al. [118] used both baseline and follow-up images together with multisequence image information in a supervised subtraction pipeline to interpret the results of subtraction images. Elliott et al. [37] used baseline and follow-up images to obtain a tissue classification and then combined it with a random forest classification considering voxel neighborhoods and incorporating lesion level features to refine new lesion candidates. On the other hand, Sweeney et al. [118] included reference

FLAIR images and subtraction images of other sequences in a logistic regression model.

Serio-temporal analysis approaches

Temporal analysis is based on the analysis of long time-series of MR images, i.e. more than two explorations. Note that in these cases, the subtraction techniques should not be employed. Hence, in temporal analysis, the intensity of each voxel is regarded as a function of time, and the aim is to see how the brightness of these voxels varies over time. This analysis is useful for both lesion segmentation [45] and characterization [78].

Gerig et al. [45] combined space and time into a 4D volume in order to track the brightness of each voxel. They first applied a supervised method to segment normal tissues based on a parametric maximum likelihood classification and parzen windows [44]. Afterwards, they distinguished active lesions by computing the mean and variance of the voxel time-series, since voxels belonging to active lesions show a higher variance compared to static tissues. Note, however, that this temporal analysis relying on voxel level comparison assumes a perfect registration among the different volumes, which cannot be true in most cases. This drawback can be minimized by taking the spatial correlation between neighboring voxels into account [143]. Therefore, the voxels' gray-value information and their surrounding tissue in all the serial scans were stored in the database, implicitly assuming that the mean spatio-temporal evolution of all the lesions in the database can be regarded as characteristic models of typical MS lesions.

Recently, Srivastava et al [114] presented a statistical segmentation method based on building a lesion specific feature map. They incorporated a template-driven segmentation of the three main tissues (CSF, WM, and GM) and then used the ratio of cortical thickness over an absolute image intensity gradient. The statistical parametric map was thresholded in order to detect lesions. They stated that their method can be applied to almost any lesion satisfying the thickening and blurring models, hence lesions with a volume smaller than $3.8cm^3$ could be detected.

2.4.2 Deformation field-based approaches

An MS lesion is generally seen as the combination of two different effects, tissue transformation and tissue deformation [122]. Tissue transformation refers to the intensity change in the lesion's tissue, while tissue deformation refers to the modification of its surrounding tissue, due to the lesion's expansion or contraction. Therefore, using only approaches

based on intensity changes between serial scans to evaluate the evolution of lesions may not give satisfactory results, since the surrounding tissue deformation due to the presence of the lesion is not taken into account. In order to consider the mass effect of lesions, deformation-based approaches should be employed.

In deformation field-based approaches, a non-linear registration is performed between successive scans, and the structural changes are determined based on the local deformation of voxels. Note, however, that due to the fact that this approach looks for the differences between successive scans, static lesions cannot be detected.

Vector displacement fields

Thirion and Calmon [122] proposed a semi-automatic approach using vector displacement fields obtained with a non-rigid registration of two successive scans to track MS lesions. They proposed using both the divergence and the norm of the displacement vector fields in order to provide sensitivity to deformation and intensity changes. 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. Moreover, they also observed that noise was characterized by high divergence and low norm, while the norm was large and the divergence low in the case of a translation. Hence, a region of interest encompassing the lesion and the surrounding tissues should be selected to perform this analysis. In their evaluation, the authors demonstrated that this method worked better than intensity-based methods when there was a mass effect without any change in enhancement, although intensity-based methods performed slightly better when there was no mass effect.

Rey et al. [99] improved Thirion and Calmon's approach [122] by using the Jacobian operator to determine local volume changes instead of using the divergence and norm of the vector fields. Furthermore, they used multi-resolution levels to avoid the influence of motion in the center of a lesion by the vectors in the boundary. By using the Jacobian operator, it is possible to distinguish the lesion's evolution. As is commonly accepted, the authors stated that a Jacobian operator larger than 1 indicates a local expansion, while smaller values indicate local shrinking. Furthermore, they can segment lesions by using a threshold defined on the Jacobian operator (for instance, a threshold of 0.3 indicates significant shrinking). Actually, in their work, they only analyze shrinking lesions, due to the richer information when looking at the shrinking field and expanding areas more greatly influenced by the spatial smoothing. Note that this is not a main drawback,

since they use both information about the deformation field from old to new images as well as from new to old images. Comparing this algorithm with image subtraction, they demonstrated that the Jacobian operator was invariant to registration errors, although the algorithm gave poor results for segmentation.

Deformation field morphometry

Recently, Pieperhoff et al. [94] applied deformation field morphometry to the detection of local volume changes in Parkinson patients, although this algorithm could also be used to detect the evolution of MS lesions. The authors considered MR images as a $3D$ set of grid points and calculated the deformation vectors related to the grid points between the images that indicate shifted voxels in the source image (a deformed image to target image). Hence, they defined the local volume ratio (LVR) as the volume of the deformed voxels in the source image divided by the volume of the non-deformed voxels in the target image. A local volume ratio greater than 1 shows a local increase and vice-versa. Subsequently, they created LVR-maps that comprised the LVR values of all the voxels. An LVR-map can be used in a ROI by adding up the LVR values of all the voxels. Furthermore, they compared LVR and the Jacobian determinant, and reported that LVR gave smoother volume measures since the latter only considers 4-6 deformation vectors, whereas LVR is computed from 27 deformation vectors. Moreover, the Jacobian operator requires the calculation of partial derivatives, which usually introduces approximation computation problems.

2.5 Classification of MS lesion quantification in serial brain MRI

As well as performing the MS lesion detection and the change detection in MR images, the quantification process is also essential for radiologists and neurologists to analyze the patient's follow-up. As already presented in Section 2.2.2, there are different ways to quantify lesion evolution: *visual inspection*, *statistical change detection*, and *volumetric approaches*.

As shown in Table 2.1, the approaches based on lesion detection typically use volumetric approaches to quantify the lesion's evolution. Metcalf's 4D connected component analysis [79], which uses a time domain on registered segmented images, may be the most common approach for this purpose (see, for instance, the following study that use this

quantification approach [38, 46, 61, 140]). A 4D connected component analysis provides the size and position of the lesions in a time line and is commonly used to identify individual lesions in a time series.

In order to perform the quantification with temporal-based approaches, the outcome of the images obtained must be interpreted. Observing Table 2.1, it is clear that point-to-point subtraction methods commonly use visual inspection to detect active lesions and interpret the lesion's evolution. For instance, Moraal et al. [86] detect positive activity by analyzing the bright area against a gray background. Furthermore, statistical intensity-based approaches use additional techniques to interpret the outcome images. For example, Lemieux et al. [68], who used a structured noise map (SNM) to identify lesion evolutions by comparing the outcome image with the SNM, and Bosc et al. [17], who used the generalized likelihood ratio test to avoid the drawbacks of a direct manual visual inspection.

Regarding the deformation-based approaches, both Thirion and Calmon [122] and Rey et al. [99] used vector fields obtained from the non-linear registration step to identify the lesion's evolution. Vector fields allow the displacement of tissues and lesions so as to be more readily visible. For instance, Rey et al. [99] showed how the displacement field emphasizes a shrinking lesion while Thirion and Calmon showed the 3D deformation field measured between two volumetric MRI's of the same patient at the level of the lesion. Moreover, Thirion and Calmon also used the volume variations measurements to validate their method's accuracy by comparing it with a conventional segmentation result. The approach by Rey et al. could also be used to segment lesions by defining a threshold. Therefore, volumetric analysis was also used to quantify the lesion's evolutions.

2.6 Experimental validation

An experimental validation of brain MRI serial methods is not an easy task. The main problem when evaluating serial brain analysis remains the difficulty of obtaining a solid ground truth. Also, some of the automated methods do not provide a final quantitative result, but a processed image that is later shown to experts who provide the final diagnosis. In these cases, the experimental results usually evaluate the performance of the radiologists with and without using the software. In what follows, we explain the main steps that researchers follow to prepare the data and evaluate their approaches.

2.6.1 Data preparation

The initial step needed to perform a validation of any algorithm is the selection of cases. Depending on the aim of the validation, a different subset of images may be necessary. For instance, if the accuracy of a segmentation algorithm is evaluated, only lesioned volumes are necessary, while lesioned volumes along with healthy controls are necessary to evaluate the performance of a lesion detection algorithm. Moreover, it is usually interesting to cluster the data according to the total lesion load in order to correlate this with the algorithms performance.

Reviewing the literature, a variety of MRI scanner machines are used, such as the 2-T Bruker [17], the 1.5-T Phillips [4], the 1.5-T Siemens [119] and the 1.5-T GE machine [82, 46, 36]. All these systems provide different fields of view (FOVs) (25.6 cm, 230 mm, 196 $\times$ 310 mm, 230 $\times$ 310 mm, etc.), different slice thicknesses (usually between 2 and 6 mm), and different sizes of the final image volume (256 $\times$ 256 $\times$ 54, 256 $\times$ 256 $\times$ 38, 162 $\times$ 256 $\times$ 20, 128 $\times$ 256 $\times$ 22, etc.). Moreover, different MRI modalities are acquired for each patient, typically T1-w, T2-w, PD-w, and FLAIR images (2D or 3D), which can be acquired from different views, usually axial or sagittal. This variety of inputs should be covered by the algorithm developed, which cannot be an easy task in terms of computational speed or amount of memory used. The most common way to deal with this data is to construct a (virtual) 3D volume. Hence, in serial analysis, where two or more volumes are analyzed at the same time, researchers use the term *4D dataset*, assuming that time is the fourth dimension.

Once the 3D volumes have been obtained, they are still not ready for direct processing. As explained in Section 2.2.3, some inherent problems of the MRI data should be addressed before tracking the lesions. Bias-field correction, spatial co-registration and intensity normalization are applied to correct for inter-scan intensity variations (due to scanner drift or other technical sources) and are usually applied to each 3D volume individually. Once these artifacts have been minimized in both volumes, the registration step between the volumes can be performed. New problems arise here such as different intensity normalization between the different volumes and issues caused by deformation artifacts that may be related to the registration itself (repositioning) or to the voxel interpolation. Note that the brain extraction is usually performed after the registration step in order to take advantage of the fact that the skull should be invariant in the different scans.

2.6.2 Ground-truth preparation

In general, there are two different ways of evaluating the approaches: with experiments using synthetic data or with experiments using real data. The use of a phantom brain (like the BrainWeb one [2, 26]) provides an excellent framework to quantitatively evaluate the algorithms. However, it is well-known that synthetic data do not reproduce all the complex factors involved in real data, and algorithms working in these environments may fail when tested on real data. In contrast, the introduction of simulated lesions into real MRI scans provides a controlled ground-truth in a more realistic environment. For example, Thirion and Calmon [122] introduced spherical lesions with blurred contours that were obtained by averaging the intensities of real lesions, while Bosc et al. [17] used cubic lesions with Gaussian profiles obtained from real lesions in all the modalities they used. In contrast to these studies that only introduce lesions in the new volumes, Rey et al. [99] suggested the addition of lesions in both old and new volumes to obtain a more realistic evaluation.

The common way of obtaining the ground-truth of real data is with an accurate manual segmentation performed by at least one expert. If more than one expert segments the images, the final ground-truth will be more reliable [138]. For instance, the ground-truth used in Anbeek et al. [4] was first segmented by an expert, and then the manual segmentations were independently reviewed and corrected by two other experts, who were blinded to the clinical symptoms of the patients. Finally, the manual segmentation was re-evaluated in a consensus meeting and considered as a gold standard. Molyneux et al. [82] also noted that the potential for any memory of the images may introduce a systematic bias. Therefore, they suggested minimizing it by randomizing the scan order and ensuring a delay of at least one week between repeated measurements of the same scan.

One of the key points usually not considered in the approaches is related to the degree of difficulty of the data, which can be measured using the coefficient of variation (COV) between the annotations. The COV is the ratio of the standard deviation of the measurements to the mean and provides a measure to indicate the reproducibility of one strategy [147]. It is common to differentiate between:

- Inter-rater COV: variation of the results between different experts.
- Intra-rater COV: variation of the results at different times with the same expert.

As an example, Zijdenbos et al. [152] presented a COV of 44% for an evaluation made by experts from seven different institutes. This value indicates that the image data they used was complicated and resulted in a large variability even among the experts.

Moraal at al.[86] also noted the necessity of training the radiologists when performing the ground truth. First, the radiologists checked the image differences in healthy patients. Subsequently, they checked the difference in brains with MS lesions (not present in the testing set). Therefore, when they provided the ground-truth using their software, they looked for lesions individually, and finally arrived at a consensus opinion.

2.6.3 Validation with the ground truth

The validation of an algorithm using a ground-truth depends on its final aim. In many computer aided diagnosis (CAD) systems [34, 35], the output is not the accurate segmentation of the lesion but the capacity of the algorithm to detect lesions. In these systems, the performance is computed using ROC and FROC analysis. ROC analysis is performed at a case-level, and is used to evaluate the capacity of the algorithm to distinguish between normal or abnormal (containing lesions) cases [59]. In contrast, FROC analysis is performed at the region-level, and plots the percentage of detected lesions against the number of false positive regions detected. This analysis is useful when evaluating the performance of the algorithm to detect lesions [59]. In this latter analysis, a region of interest should be defined. For instance, Yamamoto et al. [145] assume that a lesion is detected when a single voxel is marked inside the lesion. On the other hand, to evaluate the performance of a segmentation algorithm, the most commonly computed measurements are sensitivity, specificity, and the Dice similarity coefficient [31], all computed at voxel-level. The sensitivity measures the percentage of well-detected voxels among all the lesions in the volume, the specificity is related to the capacity of an algorithm to avoid false positive voxels, while the Dice coefficient indicates the overlap between the automated and the manually delineated lesions (this measure is also known as the similarity index [6]). Again, the COV coefficient may be used to compare both the automated and the manual results obtained. On the other hand, with respect to MICCAI challenge [1], the number of lesions correctly identified, the number false positive lesions as well as volume difference and surface distance are becoming common measures for evaluation of the segmentation algorithm.

However, in serial analysis, this quantitative evaluation is less important, since it does not quantify the effectiveness of the algorithm in tracking the lesion's evolution. In this sense, the comparison between the result of the automated algorithm and the ground-truth in terms of absolute [152] and changed lesion [82, 36] volume may provide a more realistic evaluation of the algorithm. A reliable qualitative evaluation was performed by

Bosc et al. [17], who visually evaluated their algorithm using two experts, and classified the automatically detected lesions into three different categories: valid lesion evolution, valid non-lesion evolution, or false detection. Although subjective, this evaluation provides a clear indication of how well the algorithm tracks the evolution of the lesions. On the other hand, volumetric analysis can also be done by using the number of voxels or number of true positive (TP) and false positive (FP) regions instead of comparing volume of lesions. For instance, Battaglini et al. [11] and Elliott et al. [37] use region-wise sensitivity and false discovery rate (FDR) while Sweeney et al. [118] provide both voxel-wise sensitivity and volume change in order to compare with the ground truth.

2.6.4 Validation without a ground truth

Due to the difficulty of obtaining the ground-truth in their experiments, researchers developed different ways to demonstrate the consistency of their approaches. One of the most common ways to show the robustness of an algorithm is the scan-rescan validation, where the experiments are repeated several times to show the differences in the final result, which can be done with the COV coefficient (COV for reproducibility or also known as inter-scan COV). Note that to correctly perform this validation, patients are removed from the MR room after the first scan and then repositioned in the MR machine by a different technologist [36, 46, 78].

Other ways to show the robustness of the approaches is through temporal coherency and sequence coherency, although some specific features are needed in both cases. Temporal coherency consists of checking the differences in the lesion's volume through the different explorations [122]. The idea is that the lesion's volume should not change drastically between two consecutive explorations in time (assuming there are no relapses in that time). This is analog of the evaluation of SIENAx [111], where the authors computed the error of their method for atrophy quantification in a three-times exploration by checking if the tissue loss in $T1-T2$ added to the loss in $T2-T3$ were equal to the loss in $T1-T3$. On the other hand, the sequence coherency aims to compare the results of an algorithm when detecting lesions through the different MRI sequences independently [82]. However, those algorithms that rely on the analysis of a single sequence (i.e. FLAIR) or the use of two or more sequences together cannot perform this evaluation.

2.7 Evaluation strategies for change detection

The evaluation of lesion change detection techniques is an extremely difficult task [70]. The common way to obtain a ground truth for validation purpose is an accurate manual segmentation performed by at least one expert but preferably more. However, in many cases, such a ground truth estimate is very subjective and varies highly between experts [152]. These variations can be analyzed using measures like the correlation of variation (COV), which may be computed between different experts (inter-observer COV) or between different annotations made by the same expert (intra-observer COV) [152]. Note that these measures can also be seen as measures to evaluate data complexity.

In the following, we analyze the advantages and drawbacks of the different validation methodologies used in the literature and propose an objective and quantitative evaluation of MS lesion change detection.

2.7.1 Visual inspection of outcome images

The most common approach for the detection of change in serial imaging is visual inspection, which must be performed manually by experts [93]. Tan et al. [119] validated their approach by computing the inter-observer agreement between six observers that separately identified new, enlarging, resolving, and shrinking lesions directly in the subtracted images. Although the use of different observers reduces validation errors, presenting the experts with subtracted images might intrinsically be a source of errors. The fact that the subtracted images, the only images that experts actually see, are the result of the whole lesion detection pipeline and make the introduction of information that corresponds not to valuable lesion information but to artifacts or errors introduced by the lesion detection process possible. This might induce experts to make erroneous interpretations. A similar approach was proposed and validated in terms of inter-observer agreement by Moraal et al. [85, 86]. Besides, Moraal et al. [85] compared the inter-observer agreement using the subtracted images as well as the original unregistered images. Significantly higher inter-observer agreement was observed when using subtraction. Consequently, there seems to be a higher degree of consensus among experts when using subtraction images, however, this does not prove that lesions detected in this way are correct. For example, a wrong subtraction or registration might produce a clear indication of a lesion, leading experts to a consensual, albeit wrong, decision. Thus, apart from the intra/inter observer agreements, a quantitative validation of a pipeline should also be performed on original image pairs.

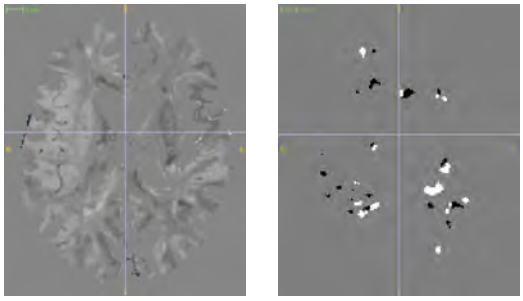
2.7.2 Validation by segmented lesions in baseline and follow-up images

A possible way to avoid biasing the experts would be to annotate images directly on the original baseline and follow-up images. Lesion evolution ground truths would then be obtained by subtracting them. Although this looks like a straightforward solution, images cannot be subtracted directly. The reason for this is the need of image registration previous to the subtraction. Consequently, the transformation matrix computed (either normal rigid or halfway registration) should be applied to ground truth images (using nearest neighborhood interpolation, since the ground truth annotations are binary images). Afterwards, the ground truth might be obtained by subtraction. However, this validation method assumes that either the registration step is perfect or the registration errors between the baseline and the follow-up images affect the ground truth in the same way. It is, however, a well-known fact that this is not the case [33].

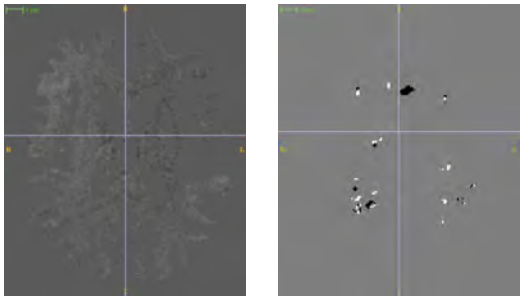
The following example shows the pitfalls of this evaluation methodology. Two subtracted images were computed between baseline and follow-up images using the proposed pipeline. The only difference in the images was due to registration. In the first case (R1), registration was stopped far before reaching a satisfactory stopping condition, while in the second (R2), convergence was reached. Therefore, the R1 registration was "worse" than the R2. To evaluate the results of the pipeline, we applied the transformation matrices to the ground truth and manually tuned the pipeline's thresholding step in order to obtain the best lesion change overlap coefficient. Surprisingly, despite the fact that registration R1 aligned the images undesirably, the overlap obtained was better. The explanation for this is that the worse registration (R1) changed the ground truth in an undesirable way, as graphically shown in Figure 2.6. Using the R1 registration, the overlap obtained consists of 3255 positively changed voxels, while when using the more accurate registration R2, this number decreased to 1853. Therefore, we seized our chance to get more true positive (TP) voxels for the R1 registration, as this error was a failure for the validation procedure.

2.7.3 Validation by segmented new lesions in follow-up images

The above example clearly shows how a different strategy needs to be used in order to obtain a more reliable ground truth. The problem above appears mainly when subtracting the same lesion manually marked in the two scans. Note that here there are two different problems. First, experts might annotate the lesion in a different way. This might happen because of different acquisition parameters (including head positioning) or because of



Registration R1



Registration R2

Figure 2.6: An example of the pitfalls of this evaluation methodology: The first column shows the results of the pipeline before manual threshold, while the second shows the ground truth obtained applying the registration transforms found during the pipeline. White areas in the subtracted ground truth images indicate positive activity (new or growing lesions) while the black areas indicate negative activity (shrinking or resolved lesions).

morphological changes (either in the lesion itself or in the brain tissues). Second, even if experts annotate the lesion in the same way, due to registration inaccuracies, the lesion will look slightly different in the two scans registered.

Therefore, in order to avoid this problem and perform a quantitative analysis of a change detection pipeline, we propose a more reliable ground truth that focuses only on the presence of new MS lesions in the follow-up scan. In this situation, we still need to register the baseline image with the follow-up one to compare them, but we are able to avoid subtraction errors in the ground truth. Hence, we focus the quantitative analysis on evaluating the appearance of new MS lesions using the manual annotations of new lesions located by the experts in the follow-up image.

Lastly, in order to compare the ground truth with the detection pipeline's results, they have to be in the same space, otherwise they must be registered as well. In this case, great care has to be taken since moving the ground truth image to another space could

also yield undesirable problems.

2.8 Analysis of the reported results

Table 2.2 summarizes the results of the lesion detection algorithms reviewed in terms of reproducibility (comparison without a ground-truth) and agreement with the experts (comparison with a ground-truth). Note that the automatic segmentation methods obtain good reproducibility results. Regarding the comparison with a ground-truth, we can see that the work of Anbeek et al. [4] provides the highest performances in terms of DSC and sensitivity (computed voxel-wise). On the other hand, Geremia et al. [43] achieved similar results than those of Anbeek et al. [3] using the datasets from the MICCAI 2008 Workshop. Notice also that most of the studies provide specificity values close to 1. This is due to the fact that this measure evaluates the ratio between the number of voxels correctly classified as healthy divided by the total number of healthy voxels. Therefore, considering that lesions are small spots within the whole volume, the specificity value always tends to be close to 1 [71]. A different way to evaluate the performance of an algorithm is to use a region-wise measure instead of voxel-wise one, as done in the work of Yamamoto et al. [145]. In this case, the sensitivity is computed as the number of detected lesions divided by the total number of lesions (81.5% in their work) and is compared with the total number of false positive lesions per volume or slice (2.9 per slice in [145]).

Looking at the results of the algorithms, clustering techniques perform better than conventional segmentation methods [36], and the use of additional strategies like PVEC or TDS [139] leads to increased accuracy. Note that these strategies are based on introducing the experience of the expert into the algorithms, and hence, supervised segmentation methods perform better than unsupervised methods. Nevertheless, it should be considered that this additional information, which either comes from a training set or from an anatomical template will bias the accuracy of the results.

On the other hand, Figure 2.7 provides a comparison of the results obtained by different subtraction methods in terms of inter-observer agreement, detailing the results for positive and negative lesion detection. Tan et al. [119] investigated the lesion evolution from 26 patients using a 2D subtraction-based approach, while Moraal et al. [85] also tested the use of a 2D subtraction-based method using 46 pairs of MR images from 40 patients. In later work, Moraal et al. [86] proposed and evaluated a 3D subtraction-based approach using controls from 14 patients. Comparing the results obtained by the 2D subtraction

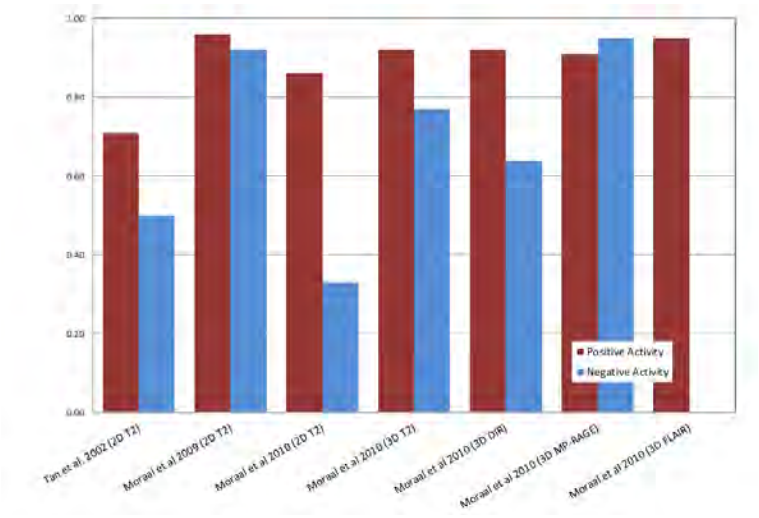


Figure 2.7: Inter-observer agreement of the subtraction-based approaches. The performance of the algorithms according to the lesion activity is shown. 2D and 3D refers to the way the subtraction is performed.

approaches, Moraal et al. [85] outperformed the results of Tan [119], thanks mainly to both the improvement in the registration algorithms and the use of an initial normalization step. However, the results of a similar strategy used with different data [85, 86] were drastically decreased. On the other hand, comparing 2D and 3D subtraction, one can see that the 3D subtraction outperforms the 2D approach, especially in the detection of negative activity. Furthermore, analyzing the results for each MRI sequence made it possible to see that the FLAIR sequence provided the best overall performance, while the use of the MP-RAGE sequence improves the detection of MS cortical lesions.

2.8.1 Improvements and further trends

Regarding the imaging modalities, the analysis of the approaches has shown that FLAIR discriminates well between lesions and healthy tissue and is used in numerous approaches to perform the automated lesion segmentation and lesion evolution analysis [87]. Recent reports have also stated that 3D FLAIR imaging reduces artifacts and provides an excellent signal-to-noise ratio compared with 2D FLAIR images. Notice that 3D FLAIR images provide 3D volume data with isotropic information and minimize the partial volume effect between small lesions and the surrounding tissue. Therefore, the use of 3D FLAIR imaging may improve the estimates of the WM and GM as well as the MS lesions.

As the MR images suffer from various image acquisition issues, pre-processing and

Table 2.2: Summary of the results obtained by different lesion detection approaches. The acronyms refer to: **MRI Sequences:** **DE** **D**ual **E**cho, **SE** **S**pin **E**cho, **GE** **G**radient **E**cho, **VE** **V**ariational **E**cho, **FSE** **F**ast **S**pin **E**cho **FFE** **F**ast **F**ield **E**cho; **Patients:** **CP** **C**hronic **P**rogressive, **FCD** **F**ocal **C**ortical **D**ysplasia, **RR** **R**elapsing **R**emitting **TPI** **T**raumatic **B**rain **I**njury; **Measures:** **CC** **C**orrelation **C**oefficient, **COV** **C**oefficient **O**f **V**ariation, **DSC** **D**ice **S**imilarity **C**oefficient, **FPR** **F**alse **P**ositive **R**ate, **FDR** **F**alse **D**iscovery **R**ate. The datasets are defined by (number of patiens) × (number of controls). If not specified, the measures are computed voxel-wise.

References	Methods	Data acquisition	Dataset	Measure	Results
[Udupa, 1997]	FCS	DE FSE T2-w/PD-w	20 MS patients	Avg. COV with FCM vs COV of 3 experts without FCM	0.9% vs 22.6%
[Guttmann, 1999]	EM+PVEC	SE/DE-SE T1-w GE Signa 1.5T	20x2 RR MS patients	Avg. LVE	$0.05cm^3$
[Kikinis, 1999]	EM+PVEC	SE/DE/LongTR T1-w GE Signa 1.5T	1 RR MS patient	COV of WML	39.5% vs 52.0%
[Wei, 2002]	EM+PVEC	DE SE PD-w/T2-w GE Signa 1.5T	11x2 CP MS 9x2 RR MS patients	Avg. inter-scan COV Zscore	7.50% −2.84
	EM+TDS			Avg. inter-scan COV Zscore	2.57% 1.84
	EM+TDS + PVEC			Avg. inter-scan COV Zscore	4.98% −0.99
[Zijdenbos, 2002]	ANN-BP	T1-w/ 2D SE T2-w/PD-w 14 Hospitals	500x3;100x4 MS patients	Avg. inter-scan COV Avg. CC with 7 rater Avg. Kappa (Dice)	0% 0.93 0.60
[Ashton, 2003]	Bayesian (DMSS)	SE VE T1-w/T2-w/PD-w	10 dataset for intra 1 dataset for inter	intra-rater COV Avg. inter-rater COV	5.1% vs 1.5% 16.5% vs 5.2%
[Ashton, 2003]	GEORG	SE VE T1-w/T2-w/PD-w	10 dataset for intra 1 dataset for inter	intra-rater COV Avg. inter-rater COV	5.1% vs 1.4% 16.5% vs 2.3%
[Antel, 2003]	Bayesian	FFE T1-w	18 MS patients with FCD	Region-wise sensitivity Voxel-wise sensitivity	0.85% 0.2%
[Anbeek, 2004]	KNN	T1-w/T2-w/PD-w/ FLAIR/IR	18 MS patients	Avg. DSC Avg. sensitivity Avg. specifity	0.81% 0.971% 0.974%
[Wu, 2006]	KNN+TDS + PVEC	DE-SE PD-w/T2-w SE /T1c-w	6 MS patients	Avg. sensitivity Avg. specifity	0.70%-0.623% 0.987%-0.997%
[Duan, 2008]	SSGE	DE PD-w/T2-w MR GE Signa 1.5T	10x2 RR MS patients	Avg. inter-scan COV Avg. LVE	0.98% 1.50%
[Duan, 2008]	CSEG + PVEC	DE PD-w/T2-w MR GE Signa 1.5T	10x2 RR MS patients	Avg. inter-scan COV Avg. LVE	8.64% 11.40%
[Shiee, 2010]	FCM	T1-MPRAGE/ FLAIR	10 MS patients	Avg. DSC Avg. sensitivity	0.633% 0.712%
[Yamamoto, 2010]	LS+SVM	T1-w FSE / T2-w / FLAIR	3x2 MS patients	Avg. DSC	0.77%
[Cerasa, 2011]	CNN	FLAIR GE Signa 1.5T	11 RR MS patients	Avg. DSC	0.64%
[Geremia, 2011]	RDF	T1-w/T2-w/FLAIR	20 MS patients MICCAI'08 Dataset	Avg. TPR Avg. FPR	0.55% 0.73%
[Elliot, 2013]	Bayesian + RFC	T1-w/T2-w/PD-w FLAIR/T1c	160x2 RR MS patients	Region-wise sensitivity Region-wise FDR	0.90% 0.23%
[Sweeney, 2013]	Logistic Regression	T1-w/T2-w/PD-w/ FLAIR	11x2 RR MS patients	sensitivity FPR	0.83% 0.001%

post-processing steps play an important role for MS diagnosis and follow-up MS patients. Therefore, bias field correction algorithms and global scaling of the images are commonly employed before registration. Besides, most of the approaches use a normalization algorithm as, for example, the recent N4 algorithm [126], particularly for MR images.

To perform a better comparison between images of different controls and particularly for the change detection algorithms, the registration is, without a doubt, the most important step. However, the registration procedure includes an interpolation process to re-sample the moving image which may affect the images and the posterior measure of the lesion's volume. Moreover, the lesions themselves, for instance, enlarged, shrunken or resolved, may negatively affect the registration accuracy. One possible way to reduce this miss-alignment caused by the lesions' evolution is to use a similarity metric robust to local differences. For instance, mutual information (MI) or normalized mutual information (NMI), which are the most commonly used measures in multi-modal registration [101, 116, 49], can be used for the serial MRI registration to reduce the effects of the lesion's evolution and other variations in the images which are caused by misalignments. The correlation ratio used for this purpose can also be a good choice for the serial MRI registration, since it can deal with intensity differences [153] and has been shown in some cases to be more robust than MI with respect to the initialization of registration [101]. In order to avoid residual artifacts caused by the registration, we have seen that some approaches also used the half-way registration method [17, 86], which is a robust way to avoid interpolation artifacts and consists of applying the same interpolation effect to both the fixed and moving images. Notice that the type of interpolation method used is also important. For instance, using a spline interpolation will provide better results than using a linear interpolation method. Some authors also suggested the *sinc* interpolation to register MR images while using a 3D pipeline [30, 93] since the frequency content of MR images is strictly band-limited [30] and therefore, is suitable for a *sinc* interpolation. However, using a high-level interpolation method drastically increases the processing time with respect to the number of iterations and resolution. Thus, a linear interpolation method may be used in the iterations of the registration, while the principal interpolation method could be used in the last iterations or just for the final re-sampling process.

By analyzing these approaches, we have seen that lesion detection and change detection techniques can be combined. In fact, this may help to carry out the diagnosis and follow-up of the patients at the same time and compensate for their inherent weaknesses. For instance, Duan et al. [36] combined a change detection algorithm based on subtraction of registered serial MR images with a detection algorithm based on a direct segmentation

of the lesions. Rey et al.[99] proposed a uniform threshold over the Jacobian operator obtained from a deformation analysis to perform the lesion segmentation. Though they provided an experimental evaluation, the results were still far from a desired segmentation. Considering the detection of new lesions, more recent techniques tend to combine change detection and subtraction imaging paradigms with single time point segmentation using multisequence information [118, 37]. The selection of one MR image sequence (i.e. T1-w, T2-w, PD-w and FLAIR) for specific purposes such as registration, detection or segmentation, or a combination of some of them will have an important effect on the results obtained. In fact, combining the advance characteristics of the different MR image types is another important factor, which was also pointed out by Mortazavi et al. [87]. Some of the approaches reviewed have already applied multi-spectral algorithms that benefit from the different signal characteristics in the MR images. Moreover, contextual features such as surrounding tissue type of the candidate lesion [37] and region-level features [11] are also used to refine the lesion detection. Making use of more information obtained from sequences, hybrid approaches can yield a better performance. Therefore, other types of information such as texture features could also be included in such a hybrid pipeline. Regarding these strategies that merge different methods, we believe that the quantification of the mass effect in vivo for MS will be a new challenge in the near future.

We want to stress also that performing an exhaustive evaluation and comparison of the existing studies is a very difficult task. The use of different data sets and evaluation measures has been a major obstacle to reviewing these methods. Ideally, approaches should be applied to a common database and compared to a single ground truth. This is, however, very difficult due to the lack of common public databases of real image scans at different time-points along with their ground truths and the fact that the methods are not publicly available. Implementation of some significant work and comparison with a common database will, without a doubt, provide a more objective comparison. However, integration of expert knowledge and a proper setting of the algorithms' parameters will be another important issue when trying to reproduce these results. As an example, Klein et al. [62] recently evaluated 14 different nonlinear deformation algorithms applied to human brain MRI registration. However, the work only focused on deformable registration, comparing a set of protocols rather than independent algorithms.

2.9 Conclusion

A review and classification of classical and to-date approaches for automatic monitoring of MS lesion evolution has been proposed and discussed in this chapter. These techniques, which have been classified according to their nature, are essential for the diagnosis and follow-up of MS patients using MR images. Assessment of MS lesion evolution involves both detection and quantification of the lesions' changes. In accordance, we have also distinguished between lesion detection and lesion change detection techniques.

The lesion detection-based methods rely on using only a patient's scan to detect lesions, and a posterior quantification method may be used to determine the lesion's evolution, which is usually carried out by using the total lesion volume between the image time-series. In this category, we have distinguished between supervised and unsupervised techniques, based on the use or not of a priori training of the algorithm. On the other hand, lesion change detection techniques make it possible to detect active lesions and interpret the lesion's evolution at the same time. However, these algorithms cannot detect static lesions since they need changed or deformed regions between the time-series. We have further sub-divided these strategies into two main categories: intensity-based and deformation-field based techniques; the former based on performing a subtraction of successive scans, while the latter can also detect the mass effect of the lesions, which is an aspect overlooked by lesion detection and intensity-based methods, and may be crucial for the MS patients.

Comparing different approaches and highlighting a single strategy is a difficult task due to the lack of a common database and a proper gold standard, which prevents making an exhaustive analysis. Furthermore, the setting of all the algorithms' parameters and the integration of expert knowledge are also important aspects to consider for a proper experimental validation. In this work, we have studied the reported results of all the automated MS lesion detection and quantification methods analyzed. We have seen that, for the lesion detection methods, the work by Anbeek et al. [4] was the most remarkable approach in terms of precision since they provided the highest DSC values and sensitivity (computed voxel-wise). Other approaches have used a different way to evaluate the performance, using region-wise measures instead of voxel-wise ones [145]. We have also seen that the precision of a proposal may be analyzed by considering the reproducibility and repeatability. In this case, the COV measure is a good way to indicate these two aspects. For instance, among the lesion detection methods, the work by Zijdenbos et al. [152] had the best reproducibility and reliability since it provided the best COV value. On the other hand, among the change detection techniques, the approach of Moraal et al. [86] provided

the highest performances with respect to inter-observer agreements. We have also seen that new techniques, particulary those aiming at the detection of new lesions, tend to combine different approaches such as including single time point information (baseline and follow-up images) in a subtraction pipeline also using feature properties of candidate lesions and multisequence information. In this sense, the work by Elliot et al [37] achieved a satisfactory performance even for smaller lesions.

Summarizing, from this analysis, we have seen that the 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 allow the mass effect of the lesions to be detected, although analyzing all the individual lesions detected is a time-consuming task and may not be necessary for expert radiologists.

Chapter 3

Temporal analysis proposal on MS lesion detection

3.1 Overview

After analysing the state-of-the art on MS lesion detection approaches in chapter 2, we concluded that statistical approaches including multi-sequence information, yield better performances. Following this fact, we propose a multi-sequence subtraction pipeline including an unsupervised auto-thresholding step in order to establish a rough detection of new lesions. Subsequently, we examine the candidate lesions with either a supervised or an unsupervised approach. PD-w and T2-w images are combined to improve the lesion detection performance in a supervised and an unsupervised manner. In order to avoid partial volume errors, we include a template driven tissue segmentation (atlas-based), particulary to obtain an accurate WM mask to apply to the subtraction images. Furthermore, we analyse the various preprocessing steps, registration methods and several other possible techniques within this pipeline. In order to assess the pipeline's performance, we used two data-sets according to the time interval between the consecutive scans, which are studies acquired with one year (12M) and four years difference (48M).

3.2 The proposed framework

Following an unsupervised thresholding strategy, we revisit the use of a simpler and fully automated subtraction pipeline based on a thresholding strategy for the detection of new MS white matter lesions (WML) in brain MRI data. The main challenges that voxel-to-voxel subtraction methods deal with are: repositioning errors (patient movement), inconsistent objects over time such as blood and cerebrospinal fluid flow artifacts, noise in the

images, and partial volume effects. Therefore, image registration, bias field correction, intensity normalization and using multi-sequence information are necessary steps to compensate for these problems [86]. After the subtraction of two consecutive temporal images, unchanging areas (normal tissue) appear as gray areas, while changing areas are darker or brighter due to either the appearance or disappearance of lesions. Typically, positive activity (new or enlarging lesions) appears as hyperintense areas while negative activity (resolving or shrinking lesions) appears as hypointense areas compared to the background. Therefore, the final step in these pipelines is a thresholding process. The selection of the threshold may be done by experts considering a trade-off between specificity and sensitivity, or in an automatic way within the pipeline. Nevertheless, in order to achieve a fully automatic pipeline, selection of these thresholds is still an open issue which will be addressed in this chapter.

The main body of the proposed pipeline in this work, schematically depicted in Figure 3.1, shares some aspects with that of Moraal et al. [86] (i.e. image registration and bias field correction), although we added some additional steps to improve the performance, specifically: the use of a WM mask obtained from single time point tissue classifications based on T1-w, T2-w and PD-w sequences [20], Gaussian filtering and the use of the mean and standard deviation of the positive activity in WM to define an automated threshold that produces the initial detection result in an unsupervised manner. Afterwards, we introduce two different postprocessing approaches based on examining the candidate lesions in the registered baseline and follow-up images and the use of multisequence information (as done by Elliott et al. [37] and Sweeney et al. [118]), combining the subtraction of PD-w and T2-w sequences in order to refine the final detection of new lesions. While Elliott et al. [37] apply a lesion level classification after defining candidate lesions by a supervised voxel-wise classification, here we propose an unsupervised approach to define candidate lesions since we believe that after white matter segmentation the lesions are outliers in the WM tissue which is usually assumed to follow a Gaussian distribution [151, 113, 20] regardless of whether or not the tissue segmentation method uses prior knowledge like Atlases [113, 20]. Hence, lesions can be taken as outliers in a Gaussian distribution and easily extracted from WM tissue without needing any prior information, improving the reproducibility of the method and reducing the complexity and computational cost. The former post-processing method is an unsupervised approach based on intensity features, particularly using the local intensity neighbor information. In the latter method, following the main framework, we implement similar postprocessing steps but in a supervised manner, studying also the use of texture features of the candidate lesions to perform the

false positive reduction step.

3.3 Validation spaces

As stated in Chapter 2 section 2.7, to assess the performance of the pipeline, a reliable ground truth and validation method is needed. For this purpose, we focus on the ground truth including only the new MS lesions determined by experts in the follow-up image (see section 2.7.3). Nevertheless, new lesions are annotated in the follow-up space. Therefore, registering by moving the follow-up image to the baseline image might cause problems since the final subtracted image is registered to the baseline space which is inconsistent with the annotation made in the follow-up space. In this case, the ground truth image has to be moved to the baseline space to perform the validation step. Moreover, we face the same problem when comparing half-way registration with standard registration. Therefore, in this section, we examine the validation spaces to evaluate whether moving the ground truth image has an affect on the validation step or not. We distinguish among the following solutions using the same transformation matrix for all cases (see also Figure 3.2):

- Direct or Normal (NW) registration. The follow-up image is used as the source of the registration (moving image) and the baseline image as the target. The transformation obtained is also applied to the ground truth of new lesions in order to perform an evaluation.

- Inverse or Reverse (RW) registration. In this case, the baseline image is the source of registration and the follow-up image is the target. This situation allows a quantitative evaluation to be carried out without having to move the ground truth annotations, which were already defined in the follow-up space.

- Halfway (HW) registration. In halfway registration, both baseline and follow-up images are registered towards a half space. In this case, the ground truth has to be moved using the same half transformation the follow-up image undergoes.

- Halfway reverse (HWR) registration. In halfway reverse registration, the output subtracted image in the half space is registered back to the follow-up space. Note that this strategy does not allow the modification of the ground truth but requires a second transformation step in order to bring the subtracted image outcome to the follow-up space.

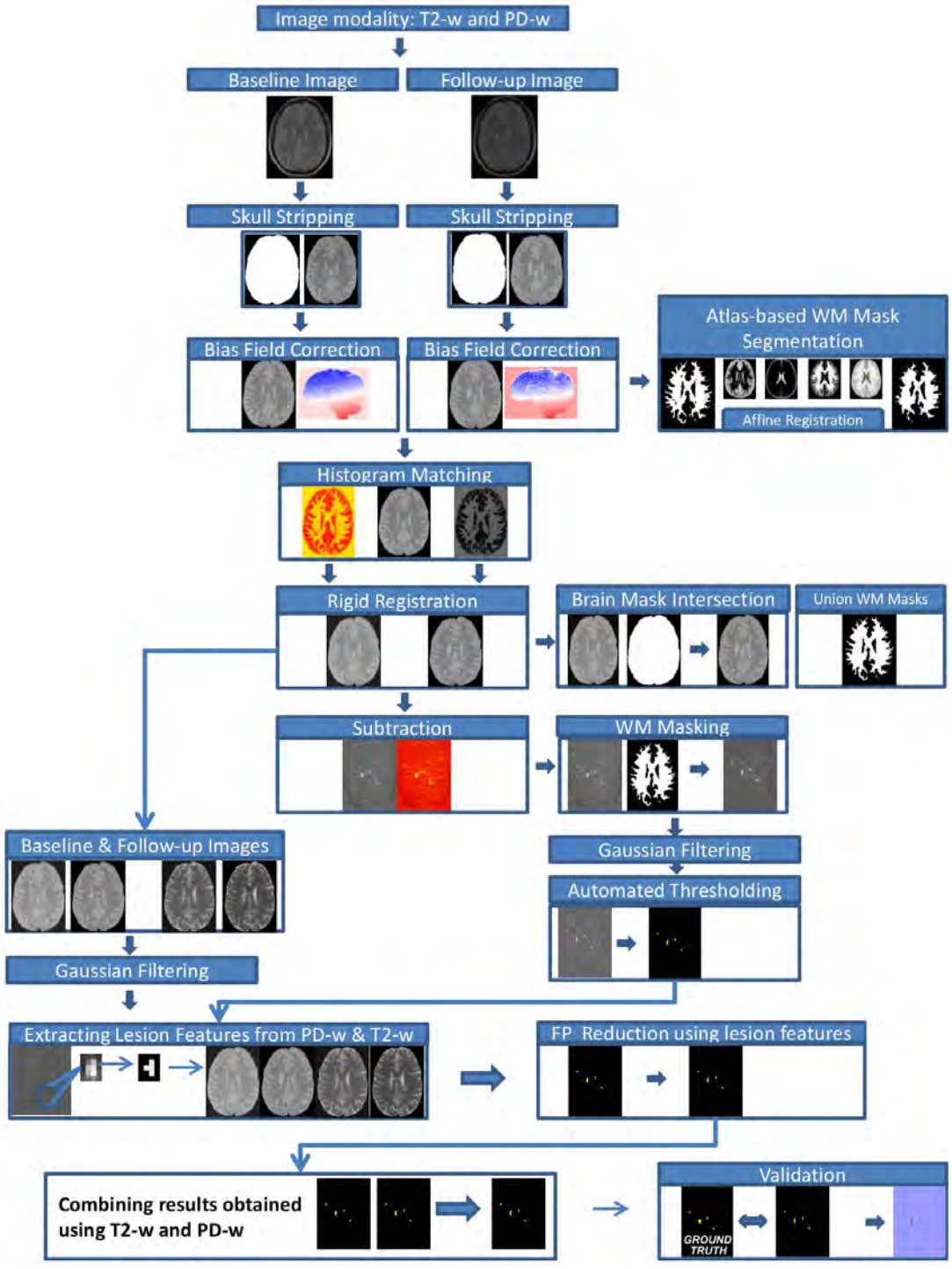


Figure 3.1: Flowchart of the unsupervised pipeline used for MS change detection.

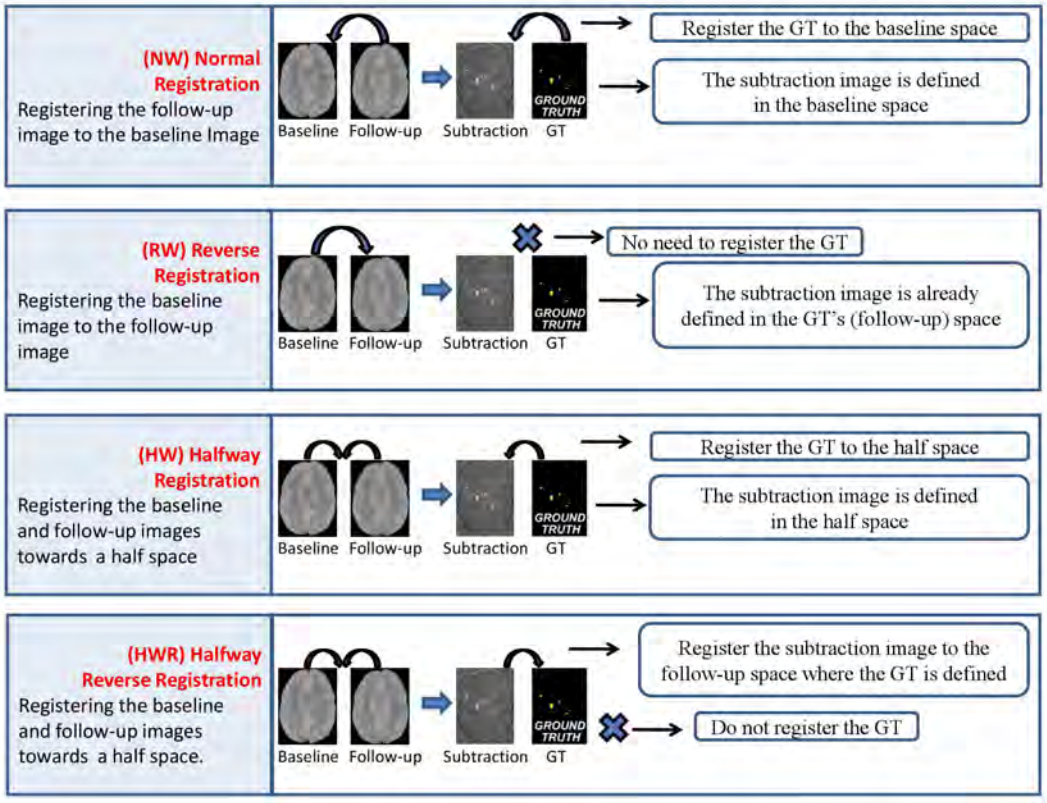


Figure 3.2: Illustration of the validation spaces.

Once a proper ground truth has been obtained, we compare the thresholded subtracted image with the ground truth. Notice how, whenever the ground truth is moved, rounding and interpolation errors can be introduced into the (displaced) ground truth mask. Considering the small size of the lesions compared to the entire image, these errors might affect all subsequent evaluation and be difficult to discern from those errors attributable to the lesion detection. We claim that the results obtained can vary when the ground truth is modified even if the same transformation is used. We back this claim in the results section by evaluating all the validation spaces described previously.

3.4 Preprocessing

Moraal et al. [85, 86] interpreted the success of their subtraction pipelines compared to the other subtraction approaches as evidence that the quality of subtraction images relies on the quality of the registration and intensity correction procedures [85]. In their subtraction approaches [85, 86], they applied a nonparametric bias field correction. Moreover, they

matched the brightness and contrast of all the follow-up images to the baseline images on the basis of the signal intensity of the intracranial cavity (ICC). Furthermore, many other studies concerning the MS lesion detection problem use similar preprocessing steps [46, 61, 17, 37, 118, 11].

Consequently, in our study, we included skull stripping, bias field correction, histogram matching steps before the subtraction process. Moreover, we analyzed these processes as well as different registration methods in detail to the point that, we computed the results when: not using normalization or histogram matching (Original), using normalization but not histogram matching (N), using histogram matching but not normalization (H), and using both normalization and histogram matching (N+H). Furthermore, we also analyzed the pipeline with different registration methods such as rigid registration, rigid half-way registration and four different non-rigid registration methods. Note that, we also analysed the performance when using scans separated by 12 months and 48 months.

3.4.1 Skull stripping

In this step, MR scans are processed in order to identify the ICC. The importance of this step is twofold. First, it allows the limiting of the search space for lesions to internal brain tissue. Second, it prevents the introduction of errors coming from the rest of the brain in subsequent steps of the pipeline. Furthermore, Johnston et al. [57] suggested applying a brain masking step before the inhomogeneity correction so that the correction would be carried out only on those voxels belonging to the internal brain tissues.

There are various methods available to perform ICC extraction. Boesen et al. [16] compared the performance of their novel brain extraction algorithm (McStrip) with three other brain extraction algorithms widely used in the neuroimaging community: Statistical Parametric Mapping (SPM), Brain Extraction Tool (BET) and Brain Surface Extractor (BSE). Using manually stripped T1-w MRI brain volumes as the "gold standard", they concluded that the McStrip outperformed SPM, BET and BSE methods. They also stated that overall the results reported for volume and boundary metrics indicate that all four algorithms provide reproducible results. On the other hand, Hartley et al. [47] compared BSE and BET software on PD-w images and concluded that neither method has a definite advantage over the other and both correlated well with the manually calculated intracranial volume (TICV), though the BET software had large positive errors (overestimated TICV) and very low negative errors compared to the BSE method.

Owing to fact that the McStrip algorithm is not publicly available and the differences

(a) (b) (c)

Figure 3.3: An example slice of 3D BET Extraction: (a) brain with skull, (b) brain mask extracted using the BET software, and (c) intracranial cavity of the brain image (ICC).

in accuracy are not significant when compared to other publicly available tools (BSE and SPM), it is also prone to low negative errors, so in this work we decided to use the BET (Brain Extraction Tool) algorithm[1], which segments the brain from the non-brain structures and also models the skull's surface. The BET algorithm defines an intensity-based estimation of the brain/non-brain threshold and lower/upper intensity values of the image from the intensity histogram to obtain a rough initial mask. With this initial mask, the center-of-gravity for the image is found. Afterwards, a triangular tessellation of a sphere's surface is initialized inside the brain and allows to deform slowly outward toward the brain's surface until the surface is well-spaced and smooth [110]. Figure 3.3 shows an example of the result of applying this algorithm.

In the BET algorithm, two parameters are user-adjustable: the fractional intensity threshold (FIT, default = 0.50) and the threshold gradient (TG, default = 0.0). Following our experimental observations, we used default parameters and, instead of applying the BET tool independently to all image sequences, applied it once to the PD-w images and masked the result to the other sequences. Moreover, manual lesion annotations have been made on this sequence. Hence, this step assumes that the various sequences are already registered to the same space. Otherwise, the images should be co-registered either using a rigid body or an affine registration [49, 91, 92, 40, 63]. Finally, since the output masks of the baseline and follow-up images may be slightly different, we intersected them in order to use only those voxels present in both.

[1]BET is part of the public FSL software.
`http://www.fmrib.ox.ac.uk/analysis/research/bet/`

3.4.2 Bias field correction

The bias field is a multiplicative smooth field that causes intensity inhomogeneities in images due to imperfections in the image acquisition process often encountered in MR imaging generally caused by the inhomogeneous RF excitation process, non-uniform reception sensitivity or electrodynamic interactions with the object [109]. This is an important issue since tissue intensity varies with its location in the image. Therefore, in order to the increase sensitivity to small changes, these intensity variations must be corrected in a preprocessing step.

Despite the fact that other similar models in the literature [135] can be used, the most common model in describing the multiplicative bias field [52] is shown:

$$v(x) = \beta u(x) + \varepsilon, \tag{3.1}$$

where, at location x, v is the measured signal, u is the true signal emitted by the tissue, β is the unknown multiplicative smooth bias field that causes intensity inhomogeneities and ε is the noise assumed to be independent of the true signal and can be approximated by a Gaussian distribution [109, 52, 135]. Note that, to simplify the computation, the noise parameter (ε) is often ignored in the equation [52].

The methods concerning retrospective bias field correction can be classified into various groups such as segmentation-based, filtering-based, surface fitting-based, histogram-based, and other specific techniques. However, none of these methods has shown to be superior to the others [52, 135]. On the other hand, a non-parametric non-uniform intensity normalization method was proposed by Sled et al [109] particularly for MR images, called the N3 method. This algorithm is based on a high-frequency maximization that assumes a simple parametric model (Gaussian) for the bias field and does not require a priori knowledge like tissue segmentation. Furthermore, the N3 method has become a "de facto" standard when other methods need comparing [135]. More recently, the N3 method was improved in 2010 by using an improved B-spline fitting and modifying the iterative optimization which improved the convergence performance and was renamed the N4 method [126]. Therefore, after the skull stripping, we used the N4 algorithm[2] for this purpose.

An example of this procedure can be seen in Figure 3.4. The first row of Figure 3.4 shows the result of the N4 algorithm for one axial slice, while the second row shows the

[2]The N4 algorithm is part of the ITK library
`http://www.itk.org/Doxygen/html/classitk_1_1N4BiasFieldCorrectionImageFilter.html`

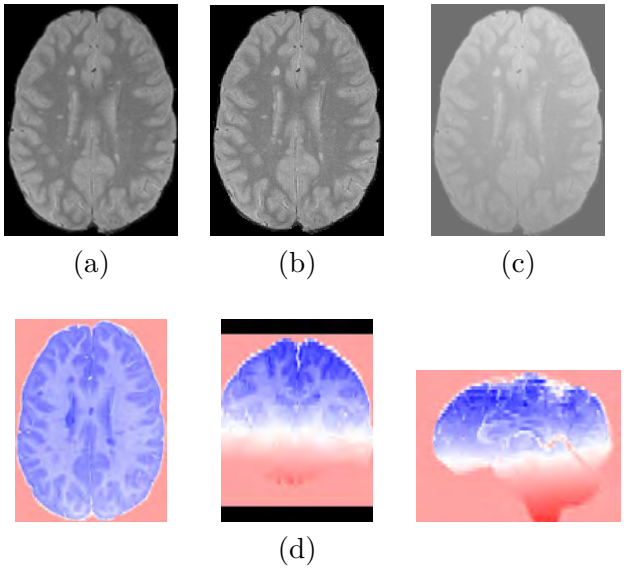


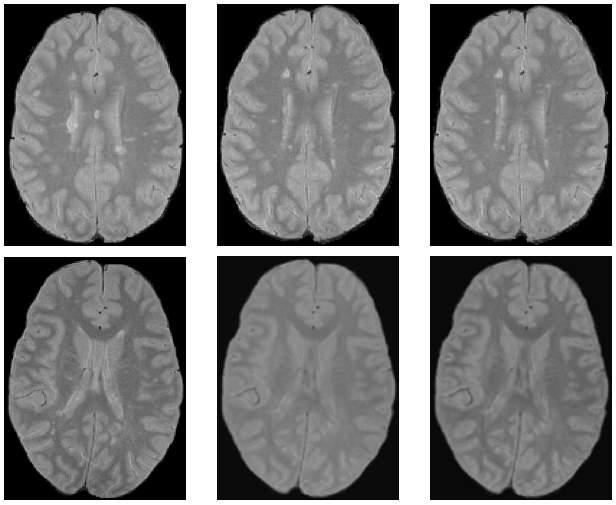
Figure 3.4: Bias field correction: (a) ICC, (b) normalized image, and (c) computed bias error of the slice. (d) shows the resulting bias error in the volume. High biased values can be seen at the top and bottom of the volume. Red areas indicate positive bias fields, blue areas indicate negative bias fields, and white areas no bias field.

result of applying it to the whole volume. Notice that the correction is different according to the different parts of the brain.

3.4.3 Histogram matching

The next step in the pipeline is histogram matching. MR images taken from the same patient at different times may appear different from each other, even if they are acquired using the same scanning machine. This presents a problem when comparing two volumes from the same patient since the actual meaning of each intensity value might vary from one to the other. Therefore, the aim of this step is to map the grayscale intensity values of the source image onto the grayscale range of the reference image. This is done by using the technique known as histogram matching [89, 136, 115]. Histogram matching aims at bringing together the intensity distribution of two images at a specified number of sample values. The histogram matching method used in our pipeline was specifically designed to normalize MR images of the same MR protocol [89]. As suggested by the authors, all pixels with grayscale values smaller than the mean were excluded in order to obtain better results[3]. Figure 3.5 shows an example of the result of applying this step.

[3]This software is part of the ITK library
`http://www.itk.org/SimpleITKDoxygen/html/classitk_1_1simple_1_1HistogramMatchingImageFilter.html`



(a) (b) (c)

Figure 3.5: Histogram Matching examples of 12M (first row) and 48M (second row). (a) N4 normalized baseline image, (b) N4 normalized follow-up image, and (c) histogram matched normalized follow-up image onto source image.

3.5 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 goal of this step is to align the two MRI volumes so corresponding voxel in both scans have the same physical spatial localization. Once this is achieved, differences in intensity between both volumes will presumably be part of an active (appearing / disappearing) lesion. Registration is done using PD-w images as this sequence obtained higher similarity metrics after registration. Subsequently, the transformation matrix obtained is applied to the other sequences as well as to the binary masks in order to move them all to the same space.

3.5.1 Rigid registration

If we assume that there is no big difference between the two successive scans, and particularly if the effect of atrophy is small, then rigid body registration is well-suited for this case. In rigid registration, images are spatially aligned using a rigid body transformation. We used a 3D versor transform[4] to define the transformation matrix between the images

[4]ITK 3D Versor Transform can be found at:
`http://www.itk.org/Doxygen/html/classitk_1_1VersorRigid3DTransform.html`

and is represented by a rigid rotation and translation in 3D space. The rotation is specified by a versor or unit quaternion, while the translation is represented by a vector. The advantage of this notation is that it includes only 6 parameters (3 for the versor components and 3 for the translation components), and therefore reduces the search space for the optimizer.

Concerning the cost function, similarity metric mutual information (MI) has been widely used in medical image registration [95] and in MR imaging [73]. It has been demonstrated that those similarity measures based on joint entropy (MI, NMI) produce better consistency and less sensitivity to the presence of extra-dural tissues such as the tissues in the brain [51]. Furthermore, MI based metrics provide acceptable accuracy in the presence of noise and RF inhomogeneity, and are the most suitable measures to determine a rigid body transformation between serial MR images of the head [51]. Therefore, the registration step in our pipeline is conducted with the MI metric proposed by Mattes et al. [76] as the cost function. This cost function is minimized by using a specialized version of the regular step gradient descent optimizer[5]. Lastly, as resampling the images requires an interpolation method, B-spline interpolators produce remarkable results as well as running fast [66]. Furthermore, B-spline interpolators are preferable, particularly for those applications in medical image processing that require high precision [67, 130]. As a consequence, we decided to resample the images at their new coordinates employing the B-spline interpolation [131, 132, 129], with the exception of the intermediate (internal) registration steps that involve linear interpolation as suggested in the study by Bosc. et al [17].

3.5.2 Rigid halfway registration

Traditional registration approaches consist of registering one volume into the other. However, when analyzing the results of a subtraction approach using the registered image and the reference one may result in many interpolation artifacts [70]. This happens because the intensities of the moving volume are interpolated while those of the reference volume are not. To avoid this issue, some authors proposed a solution known as halfway registration. In this scenario, both volumes are moved to an intermediate space, and a similar interpolation effect is applied to both the moving and reference images [86].

For this purpose, instead of moving one image to another, we simply use the same matrix obtained from the normal registration. However, this time we calculate half the

[5]ITK specialized step gradient descent optimizer for Versor Transform can be found at:
`http://www.itk.org/Doxygen/html/classitk_1_1VersorRigid3DTransformOptimizer.html`

angle radian obtained from the versor component and half the translation parameters and then, reversing them when necessary, align two images into a common space at their halfway positions as done in Moraal et al [86].

3.5.3 Non-rigid registration

Deformable registration algorithms might introduce deformations in the lesions, which for the purpose of this study, is not desirable. Nevertheless, depending on its basics, a deformable algorithm might be useful in some cases, particularly when there are larger changes between successive scans, such as when brain atrophy occurs. Therefore, if a deformable registration algorithm is able to compensate the larger changes without suppressing the lesions, this may improve the pipeline's performance. For this purpose, we also evaluate the pipeline with non-rigid registration techniques.

In the study by Diez et al. [32], the SyN and Nifty reg methods consistently outperformed other non-rigid techniques in terms of the lesion overlap Dice coefficient and obtained good values for the image similarity metrics, the sum of squared differences (SSD) and normalized mutual information (NMI). Klein et al. [62] also demonstrated that the SyN method obtained the highest rank for all tests when compared to other registration tools. On the other hand, the ITK Demons and the Dramms registration performed well in terms of NMI metric, however, though ITK demons ranked first in NMI metric, it produced worse results in terms of the lesion overlap. Therefore, four different non-rigid registration methods were especially chosen based on the study by Diez et al. [32]; Nifty registration[6], SyN registration[7], Demons[8] and Dramms[9]. All these methods were tested with rigid plus affine initialization carried out with the ITK library.

The Nifty method

Nifty Reg is a B-spline based deformable algorithm that provides faster convergence. The Nifty method includes graphics processing units (GPU) based implementation designed to reduce the computational cost of cubic B-Spline methods. They use NMI as the cost

[6] Nifty Reg download at sourceforge,
`http://sourceforge.net/projects/niftyreg/`

[7] Advanced Normalization Tools webpage,
`http://www.picsl.upenn.edu/ANTS/download.php`

[8] ITK implementation can be downloaded at
`http://www.insight-journal.org/browse/publication/154/`

[9] Dramms can be downloaded at:
`http://www.rad.upenn.edu/sbia/software/dramms/download.html`

function that is minimized by using a conjugate gradient ascent optimization [81].

The SyN method

The SyN method is a symmetric deformable image registration method that uses cross-correlation as the cost function within the space of diffeomorphic maps (topology preserving maps) and relies on Euler-Lagrange equations in the optimization process. This method particulary uses the cross-correlation formulation to provide the advantage of symmetrizing the cross-correlation Euler-Lagrange equations. Therefore, the algorithm provides inverse fields as well [8].

The Demons method

The Demons method is a non-rigid registration method based on Thirion's demons [121] and registers two images by computing the displacement field that maps the moving image onto the fixed image. A displacement is a vector whose elements behave like floating point scalars. The method encompasses techniques close to optical flow [69, 49]. Mattes mutual information is used as the cost function and is minimized by a regular step gradient descent optimizer.

The Dramms method

The Dramms method is based on the study by Ou et al. [90] and refers to a deformable registration via attribute matching and mutual-saliency weighting. As stated by the authors [90], the Dramms method bridges the gap between the traditional voxel-wise methods and landmark/featured-based methods. Note that the landmark based methods are often considered to provide better registration but require expert annotations which is time consuming and difficult to obtain. In practise, the authors try to automate determining landmarks into the image. With the 'attribute matching' term they refer to a rich set of Gabor attributes assigned to each voxel, with the 'mutual-saliency weighting' term they make reference to a novel method that aims at giving more weight in the transformation to more distinctive voxels so they can make use of distinctive voxels like the conventional landmark based methods [74] used in image aligning. Therefore, they propose a gradient descent based optimization that utilizes all the imaging voxels but with different weights [90].

3.6 Tissue segmentation and WM masking

The output of the subtraction step may yield a large number of false positives, most falling outside the white matter. On the other hand, as stated in Chapter 1, only around 5-10% of the lesions might involve gray matter [105] and approximately only 5% of the total lesion volume [23] could be comprised of gray matter. Moreover, GM lesions are more obscure and thus subtle and more difficult to detect in MR imaging [105]. As a consequence, we apply a WM masking step in order to reduce the search space to only lesions within white matter.

WM masks are computed before applying the registration step to avoid interpolation errors when segmenting the tissues. Thus, the masks are registered and re-sampled by using nearest neighbor interpolation. Finally, both WM masks defined in the baseline and follow-up images are combined (see Figure 3.6 (b)). The combination is made by applying the union operator (voxel-wise OR logical operator). Figure 3.6(c) shows the final subtraction restricted to the white matter. Notice that the white regions located at the border of the skull are now not considered as lesions. Note that, partial volume errors may arise when a single voxel contains a mixture of multiple tissue values (generally at the tissue's boundaries), and these errors may cause undesired results [36, 46, 61, 139]. Therefore, the partial volume class should be included in the clustering algorithm to avoid aggregating these voxels into the WM mask.

To obtain WM masks, we employ and test two different approaches. In the former approach, we propose an algorithm using single modalities. For this purpose, we use FSL tools[10] based on the segmentation approach in the study by Zhang et al. [151]. In the latter approach, we follow a more sophisticated strategy that combines atlas information and multisequence information obtained from T1-w, T2-w and PD-w images.

3.6.1 Tissue segmentation using single modalities

FSL segmentation tools segment 3D brain MR images based on the hidden Markov random fields (HMRF) model fitting by the expectation-maximization (EM) algorithm [151] using Gaussian estimation. The initial parameters for the EM algorithm, the mean and standard deviations for each class type, are determined after an initial estimation of the tissues using the discriminant based thresholding method proposed by Otsu [88]. The advantage of

[10]FSL (FAST: FMRIB's Automated Segmentation Tool) :
`http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FAST`

(a) (b) (c)

Figure 3.6: Subtracted image and WM masking: (a) result of the subtraction between the images 3.5(a) and 3.5(c), (b) (union) WM mask obtained by an atlas-based supervised algorithm using PD-w, T1-w, and T2-w images, and (c) WM masked subtracted image. Note that new and enlarging lesions appear as white spots in the final image.

the HMRF model is that the tissue information is encoded through the mutual influences of neighboring sites, therefore, partial volumes can also be determined. We obtain a WM mask for T1-w sequences with and without partial volumes. Additionally, we show the results for WM masks obtained from T2-w and PD-w modalities in the proposed pipeline.

3.6.2 Atlas based multi-modal tissue classification

In this approach, the WM mask is obtained using an atlas-based multi-spectral tissue segmentation algorithm that uses PD-w, T1-w, and T2-w images [20]. This algorithm, similar to the study by Souplet et al [113], uses an Expectation-Maximization algorithm to maximize the log-likelihood between real MRI data and a Gaussian model of four classes. The four classes considered are: pure tissues (WM, GM, CSF) and a partial volume class (GM / CSF). For the pure tissue classes, prior probabilities are provided by the Gaussian distributions guided by an atlas (ICBM atlas of 452 patients[11]), while for the partial volume class, a weighted atlas of CSF and GM is used. The algorithm creates a PV atlas that is used during the expectation step as a prior. The priors are redefined by a similarity map computed during registration in order to avoid a segmentation completely driven by the atlas. Finally, in order to improve the tissue parameter estimation, a threshold and a trimmed likelihood estimator are used to compute the mean and covariance matrix during the maximization step.

[11] Publicly available at
`http://www.loni.ucla.edu/ICBM/Downloads/Downloads_Atlases.shtml`

3.7 3D subtraction

Once the volumes are aligned, the 3D voxel-wise subtraction can be applied. If the follow-up image is subtracted by the baseline image, positive activity (new or enlarging lesions) appears as brighter areas while negative activity (resolving or shrinking lesions) appears as darker areas against the gray background. An example of a subtracted 2D slice is shown in Figure 3.6(a). A brighter area corresponding to a new lesion can be observed.

3.8 Thresholding and locating of WM candidate lesions

Two different types of thresholds are proposed in our proposal: intensity and size thresholds. The first one allows us to select the regions more likely to contain MS lesion changes, while the second allows us to discard small spots that may appear due to image artifacts or small registration misalignments.

We set a minimum number of voxels greater than 3 as the size threshold. This is also consistent with previous approaches, where authors also mentioned that small spots should not be considered since most of them were caused by spurious noise [119, 85]. Notice that intensity and size thresholds should be adjusted with respect to each other, so that a lower intensity threshold may require a higher threshold and vice versa. On the other hand, as the selection of intensity-based threshold can be made empirically by the experts, we aim at determining an automatic threshold providing a satisfactory trade-off between sensitivity and specificity. Thus, we propose a fully automatic unsupervised method for intensity-based thresholding.

3.8.1 Gaussian filtering

We apply a low pass Gaussian filter to the subtraction image before thresholding and to the baseline and follow-up images before postprocessing, hence, the usage of Gaussian filtering here is twofold. Firstly, the Gaussian filter enables us to reduce noise in the subtraction image and, moreover, it allows the pipeline to include neighborhood information into the lesion area when carrying out post processing steps on the baseline and follow-up images. To incorporate spatial information of the neighboring voxels, Sweeney et al. [118] proposed a Gaussian kernel with window size of 3 mm. On the other hand, notice also that shape and size of the lesions determined by the pipeline can be ill-posed depending on the threshold or method used, thus, incorporating spatial information of the neighboring

voxels might help to compensate this shortcoming. Consequently, using the information from tissue surrounding the lesion might improve sensitivity. Therefore, as done in previous studies [118], we also relax the subtraction images obtained and smooth them by applying a 3D Gaussian filter with a window size of 3 voxel radius and a σ value of 0.5. The selection of the window size and σ values will be discussed in the experimental section.

3.8.2 Unsupervised thresholding

In order to fix the threshold automatically, we propose an intensity thresholding method based on the average intensity and standard deviation of the positive activity in the union WM tissue in subtracted images. Notice that Atlas based multi-modal tissue classification [113, 20] uses an EM algorithm and a Gaussian model of four classes, while FSL segmentation tools [151] segment 3D brain MR images based on HMRF model fitting by an EM algorithm using Gaussian estimation. Therefore, both models assume that WM tissue follows a Gaussian distribution. Thus, lesions can be taken as outliers in WM tissue with respect to a Gaussian distribution, which should appear as hyperintense areas in the subtraction images. Therefore, we consider MS lesion detections as those voxels that have intensity values in the subtracted images larger than the mean and multiplication of the standard deviation by a constant (see Equation 3.2 and 3.3). The selection of the constant parameter (β) with respect to trade-off between the specificity and sensitivity will be analyzed in the experimental section.

$$\begin{aligned} \sigma &= \sqrt{\frac{1}{N_{voxels}} \sum_{i=1}^{N_{voxels}} (I(x_i) - \mu)^2} \qquad where, \\ \mu &= \frac{1}{N_{voxels}} \sum_{i=1}^{N_{voxels}} (I(x_i)) \qquad and \qquad I(x_i) > 0 \end{aligned} \tag{3.2}$$

$$threshold = \mu + \beta\sigma \,, \tag{3.3}$$

In this equations N_{voxels} is the total number of positively active voxels and $I(x_i)$ is the voxel intensity at the WM masked subtracted image.

3.8.3 Supervised thresholding

In this scenario, we propose a simple learning strategy to determine a threshold for every particular case using a pattern recognition approach. Firstly, the threshold selection is

performed by maximizing the Dice coefficient. Afterwards, for each training image, the average intensity of the positive activity (i.e. the positive values inside the white matter masked region) is computed. Following a leave-one-patient-out validation strategy, when a new case is tested, its average intensity of positive activity is computed and compared to the learnt ones using a K-Nearest Neighbor classifier. When $K = 1$, the most similar threshold is used, while for $K > 1$, the automatic threshold is computed as the average of the thresholds obtained for the K nearest cases.

3.9 Postprocessing: refining candidate lesions

After the thresholding step based on Equation 3.3, the binary image outcome may contain many false positives. In order to reduce the rate of these false positive detections we introduce two different postprocessing approaches: supervised and unsupervised intensity postprocessing steps, as well as the combination of PD-w and T2-w images in a supervised and unsupervised way.

3.9.1 Unsupervised pipeline

In order to reduce the rate of false positive detections in an unsupervised way, we introduce two simple postprocessing rules that analyze the original image intensities of candidate regions. The candidate region slices are removed if all the slices in the 3D region are determined as false positive according to the constraints.

The first constraint has the goal of discarding those regions detected in the subtracted images due to very low intensity values in the baseline image, for example, candidate regions that resulted from inaccuracies in the skull extraction process and misclassified during the WM masking. These regions turn out to be false positives even though they do not present high intensities in the follow-up image. To discard these regions, we compute the mean and standard deviation of the intensities of all the candidate regions in the original baseline image and remove those that have an intensity lower than $\mu^{Baseline}_{AllROIs}$ - $2\sigma^{Baseline}_{AllROIs}$ ($BZS < -2$, equation 3.4).

$$Baseline\ ZScore\ =\ BZS\ =\ \frac{\mu^{Baseline}_{ROI} - \mu^{Baseline}_{AllROIs}}{\sigma^{Baseline}_{AllROIs}} \tag{3.4}$$

On the other hand, the second postprocessing step aims at including the local intensity neighbor information for each candidate region in both the baseline and follow-up images.

This is done in order to remove any detected false regions created by using the global automated thresholding process. Notice that, when experts try to locate a lesion, they analyze the image both globally and locally since the white matter tissue is not smooth all the time and the global hypointensity areas may not necessarily refer to a lesion when looking locally, especially those voxels that have relatively low signals. In this case, we enlarged the candidate regions by applying a dilation operation to account for the local neighborhood. For each dilated candidate region we analyse, the coherence of the mean intensity of the detected area with the neighboring voxels. For the baseline image, we remove all the regions with a ratio smaller than 0.95 ($BNR < 0.95$, Equation 3.5). Notice that if the candidate region is surrounded by WM, the ratio should be close to 1. Moreover, we applied a similar process to the follow-up images. In this case, the mean intensity of the candidate lesion area should not be smaller than that of its neighbors. Therefore, if the mean intensity of the lesion area is smaller than the mean of its neighbors ($FNR > 1.0$, Equation 3.6) the region should be removed (see Figure 3.12).

$$Baseline\, Neighborhood\, Ratio \;=\; BNR \;=\; \frac{\mu_{ROI}^{Baseline}}{\mu_{ROINeighbor}^{Baseline}}\,, \tag{3.5}$$

and

$$Follow-up\, Neighborhood\, Ratio \;=\; FNR \;=\; \frac{\mu_{ROI}^{Follow-up}}{\mu_{ROINeighbor}^{Follow-up}}\,, \tag{3.6}$$

where μ_{ROI} is the mean intensity of the lesion area and $\mu_{ROINeighbor}$ is the mean intensity of the tissue surrounding the lesion.

Two different types of false positives are shown in Figures 3.7 and 3.8. In Figure 3.7 a false positive that is caused by intensity differences between CSF tissues on baseline and follow-up images is shown. Notice that the ROI on the follow-up image is not a lesion and BNR and BZS values are low. On the other hand, in Figure 3.8 the ROI turned out to be a false positive due to the low signal on the baseline image although it does not present high intensities in the follow-up image, thus, the FNR is greater than 1.0. See also Figures 3.9 and 3.10 for an FP (non-brain tissue) caused by an inaccuracy in the skull extraction process and notice that the ROI has a low BZS value in the T2-w sequence and the same ROI has low BNR and BZS values in the PD-w sequence. Note that, the constraints are applied to both PD-w and T2-w images separately.

The postprocessing thresholds were defined empirically within the cross-validation process. The 12M dataset was used to perform the parameter optimization with a 2-fold cross-validation scheme where the data was divided in two groups of five patients each.

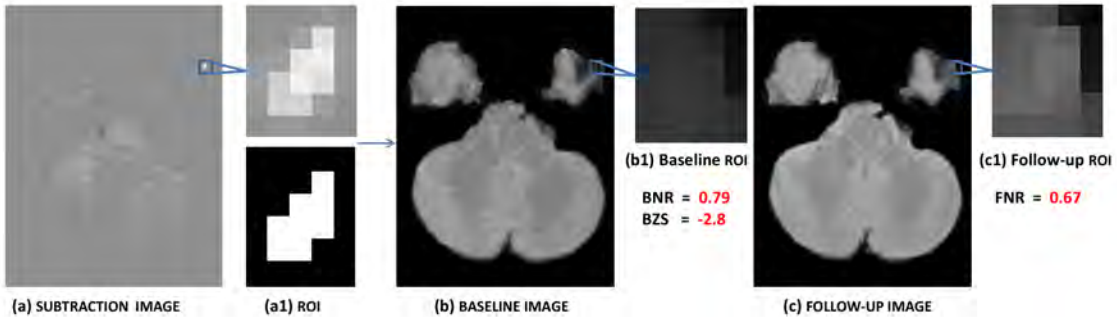


Figure 3.7: BNR, FNR and BZS features on a FP caused by CSF. **(a)** Subtraction image, **(b)** Baseline image, **(c)** Follow-up Image, **(a1)** Candidate lesion area and Region of interest (ROI) after thresholding, **(b1)** ROI in the baseline image **(c1)** ROI in the follow-up image.

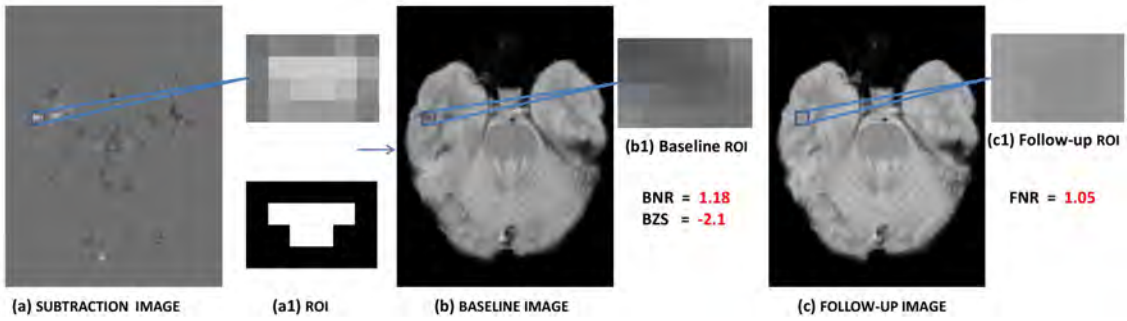


Figure 3.8: BNR, FNR and BZS features on a FP caused by low signal on baseline image. **(a)** Subtraction image, **(b)** Baseline image, **(c)** Follow-up Image, **(a1)** Candidate lesion area and Region of interest (ROI) after thresholding, **(b1)** ROI in the baseline image **(c1)** ROI in the follow-up image.

One group was retained as the testing data and the other was used for training. We repeated the process changing the fold used for training and testing, using therefore all the cases once as validation data. The DSCR evaluation measure was used to analyze the results and optimize the different thresholds. Afterwards, this 12M threshold configuration was applied to all the 48M dataset (not used for parameter optimization purposes) to provide the results shown in this work.

Combination of PD-w and T2-w images

Initially, we analyzed the unsupervised pipeline using the available image modalities in our dataset individually (T1-w, T2-w and PD images). Due to the fact that PD-w and

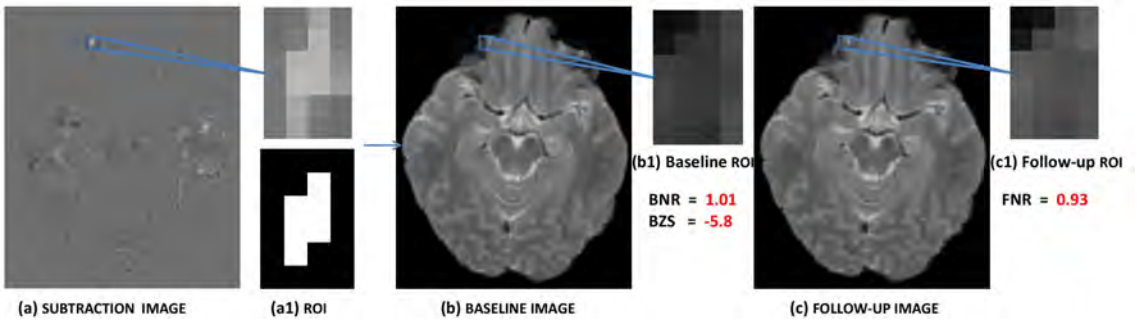


Figure 3.9: BNR, FNR and BZS features on a FP caused by an inaccuracy in the skull extraction process (non-brain tissue in T2-w). **(a)** Subtraction image, **(b)** Baseline image, **(c)** Follow-up Image, **(a1)** Candidate lesion area and Region of interest (ROI) after thresholding, **(b1)** ROI in the baseline image **(c1)** ROI in the follow-up image.

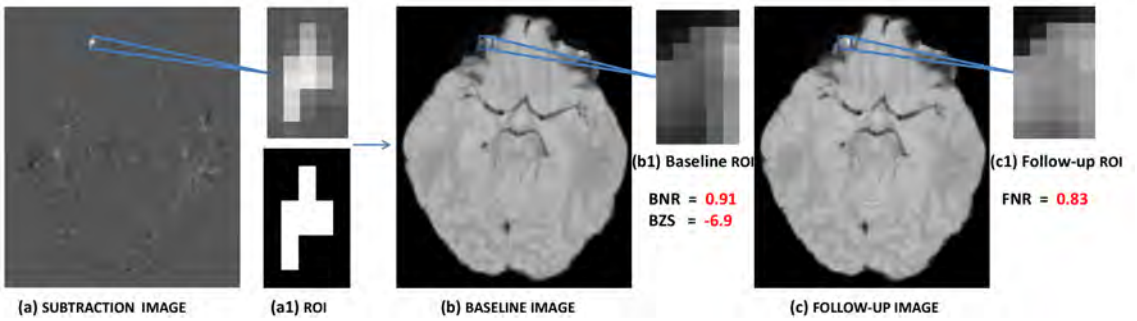


Figure 3.10: BNR, FNR and BZS features on a FP caused by an inaccuracy in the skull extraction process (non-brain tissue in PD-w). **(a)** Subtraction image, **(b)** Baseline image, **(c)** Follow-up Image, **(a1)** Candidate lesion area and Region of interest (ROI) after thresholding, **(b1)** ROI in the baseline image **(c1)** ROI in the follow-up image.

T2-w images are actually T2-w sequences, MS lesions show hyperintense areas (reflecting their increased water content) in these modalities, while false positive regions can vary depending on the sequence type. Therefore, most of the false positives caused by the scanner depending on the modality used can be removed by combining them. On the other hand, MS lesions were not always visible in T1-w images. Consequently, we decided to combine the use of PD-w and T2-w sequences to eliminate any FP detections and also to refine the MS lesion candidates.

After the unsupervised postprocessing of the sequences, in order to eliminate individual artifacts in the PD-w and T2-w images, we used a straightforward solution which is to join both results obtained in the PD-w and T2-w images using the intersection of the subtracted images. By doing this, only those lesions that appear in both PD-w and T2-

w images are considered as new lesions. Notice that the main purpose of this step is to reduce FP detections rather than increasing sensitivity since this type of subtraction pipeline suffers from a high number of FP detections in individual sequences [70].

One should also notice that owing to the fact that PD-w and T2-w images are acquired simultaneously, we can strongly assume that both images are already in the same space and well-suited for combining, also avoiding multi-modal misregistration errors.

3.9.2 Supervised pipeline

As stated in Chapter 2 section 2.3.1, first order-statistics (individual pixel values such as mean and variance of the gray level) and second order statistics (the properties of pixel pairs) can be used in a supervised classification for lesion detection [5, 149, 148, 60, 150, 102]. Using the leave-one-patient-out strategy, the SVM and KNN classifiers are employed to perform the classification of the candidate regions. The main body of the supervised pipeline in this work, schematically depicted in Figure 3.11.

Unlike the unsupervised approach, here we let a classifier determine the false positive regions using multisequence information instead of directly intersecting the binary subtracted PD-w and T2-w images. For this purpose, we collect the PD-w and T2-w information of the ROI in the baseline and follow-up images in a supervised algorithm where we have a feature vector for each ROI. Consequently, we combine features from both PD-w and T2-w modalities concatenating the feature vectors.

Supervised pipeline using intensity features

In the first place, we analyze the previous unsupervised pipeline in a supervised manner using the same features (FNR, BNR and BZS) obtained for the postprocessing step. However, we use here a supervised algorithm to remove the false positive regions instead of using fixed FNR, BNR (see also Figures 3.12(b2) and 3.12(c2)) and BZS constraints. By doing this, we also propose evaluating the pipeline without using fixed constraints. To compare this pipeline with the previous unsupervised pipeline properly, we trained 12M and 48M datasets separately.

Supervised pipeline using GLCM features

Secondly, using second order statistics, we include texture features of the candidate lesions obtained with a gray level co-occurrence matrix (GLCM) [150]. Although the selection of the ROI is typically carried out manually by experts, for this purpose, we make use of the candidate lesions obtained by our pipeline using enlarged hypothetical rectangles (see Figures 3.12(b3) and 3.12(c3)). In this sense, the energy, contrast and homogeneity properties of the ROI in the baseline and follow-up images are included in the supervised classifier. We re-scaled the ROI matrices by using 8 gray levels (see Figures 3.12(b4) and 3.12(c4)). Normalizing the intensity values, we used the mean intensity of the baseline and follow-up ROIs as the lower limit intensity value and upper limit intensity value respectively. Furthermore, in order to carry out an isotropic (bi-directional) approach, we obtained the mean energy, contrast and homogeneity values by including all the neighbors of the voxel of interest (8 directions). We have illustrated a sample analysis of a new lesion in Figure 3.12. Notice that, concerning a new lesion candidate, the ROI in the follow-up image (new lesion) involves a stronger contrast but lower energy and homogeneity compared to the same ROI in the baseline image. As a consequence, GLCM features are used in the supervised classifier to determine lesions. We trained the data using all the patients applying the leave-one-patient-out strategy.

3.10 Summary

In this chapter, we have proposed a change detection approach to detect new lesions based on a subtraction method. First of all, we have presented a reliable ground truth and validation method mentioned in Chapter 2. The effect of the validation spaces will be analyzed in the next chapter. We have also presented a framework including an initial preprocessing pipeline that comprises several steps to reduce image artifacts and MRI issues introduced in Chapters 1 and 2. Furthermore, we have presented several different methods in every step of the pipeline to improve the performance such as various registration and WM masking methods which will also be evaluated in the next chapter. A novel strategy for automating the threshold of the subtraction image has also been addressed in this chapter. We proposed using the mean and standard deviation of the active positive activity in the subtraction, and, have presented a fully automated unsupervised framework and then re-established it in a supervised manner. Afterwards, we presented another supervised framework including texture features of the candidate lesions. In this

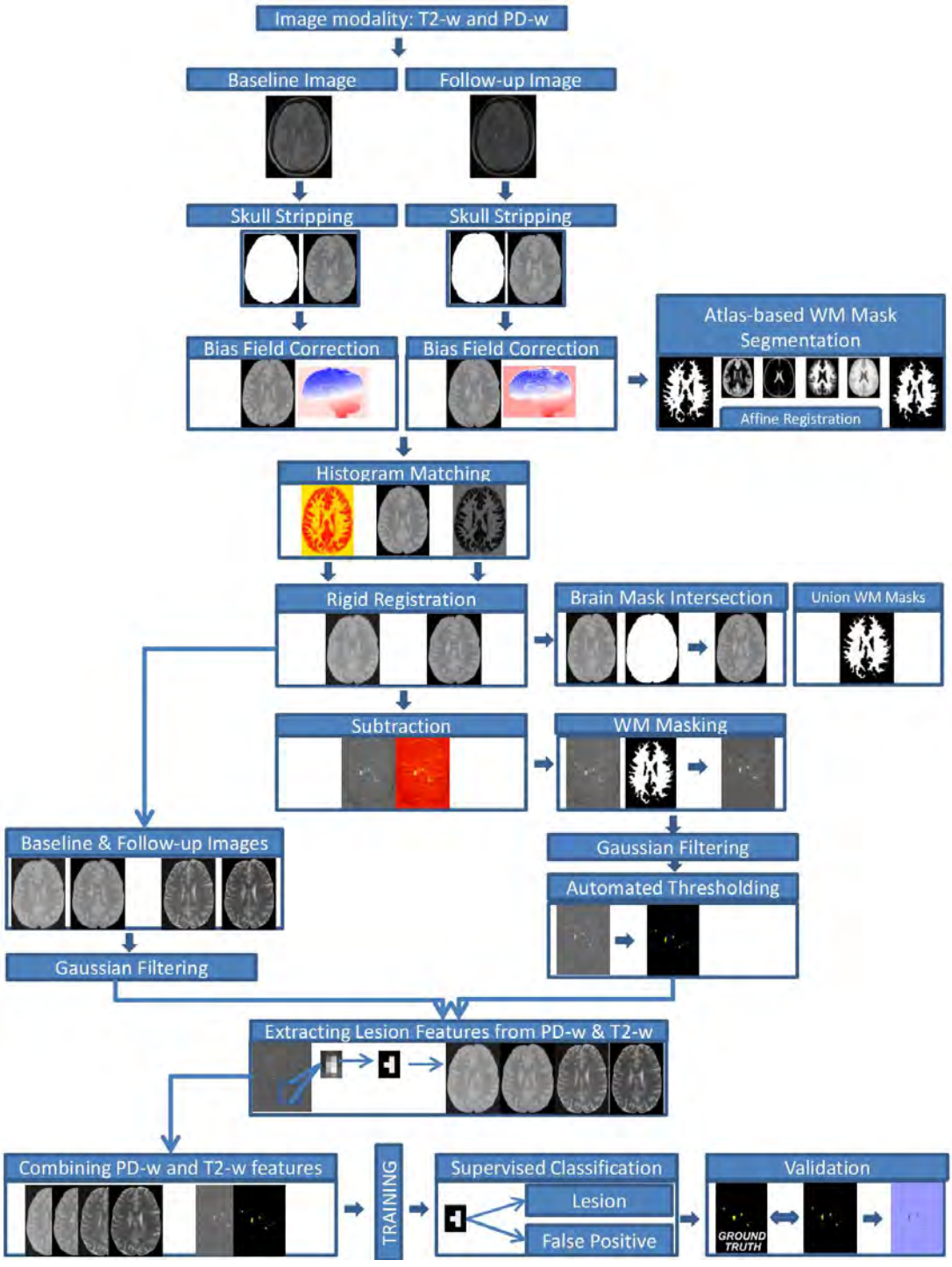


Figure 3.11: Flowchart of the supervised pipeline used for MS change detection.

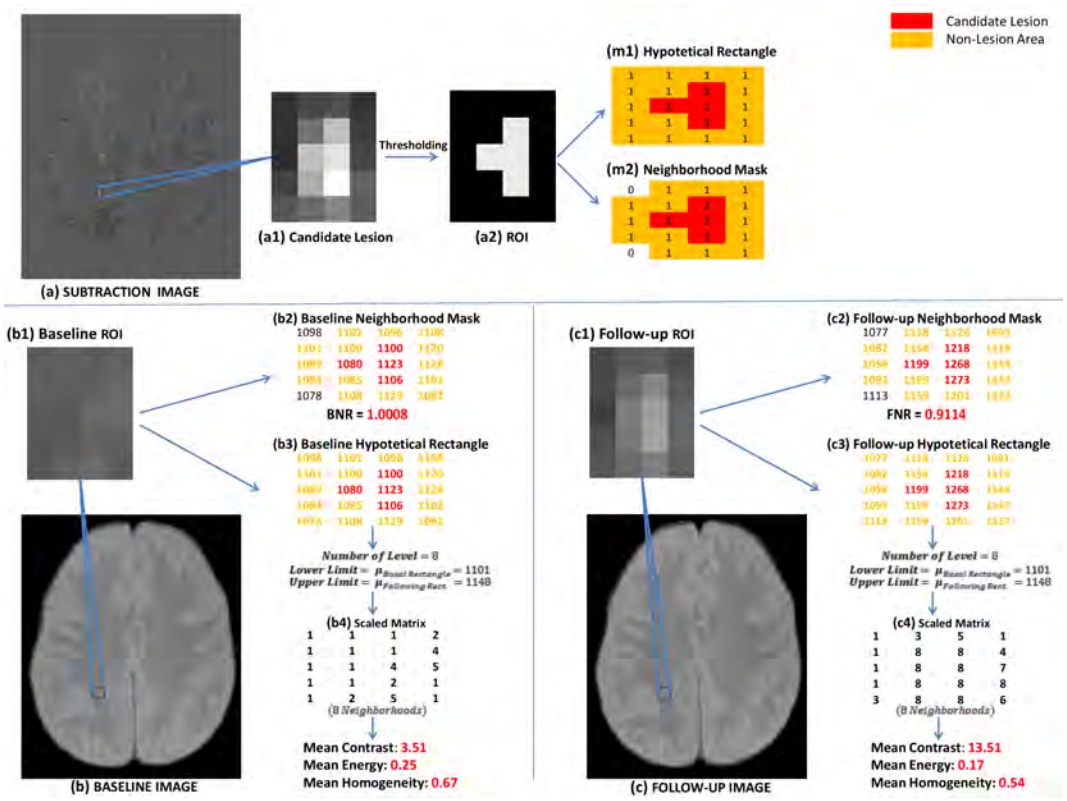


Figure 3.12: An example of MS lesion detection postprocessing: **(a)** Subtraction image, **(b)** Baseline image, **(c)** Follow-up Image, **(a1)** Candidate lesion area, **(a2)** Region of interest (ROI), **(m1)** Hypothetical rectangle to obtain texture features of the lesion's ROI, **(m2)** Neighborhood mask to obtain the ratio between the lesion and its surrounding tissue, **(b1)** ROI in the baseline image **(c1)** ROI in the follow-up image, **(b2)** The intensities in the neighborhood mask in the baseline image to obtain the BNR value, **(c2)** The intensities in the neighborhood mask in the follow-up image to obtain the FNR value, **(b3-b4)** Texture information in the baseline image, **(c3-c4)** Texture information in the follow-up image.

section, we will summarize the pipelines to be evaluated in the next chapter and highlight our contributions.

3.10.1 Preprocesing pipeline

Before the detection of new lesions, different preprocessing steps must be applied to prepare the images for further analysis. The first step includes removal of the skull which is applied by means of the publicly available BET tool [110], and is a part of the FSL toolbox. Afterwards, we applied a bias field correction based on the N4 method, which is an improved version of the N3 method from the study by Tustison et al. [126]. In the

following step, we aimed at bringing together the intensity distribution of the two images based on histogram matching [89, 136, 115] that maps the grayscale intensity values in the source image into the grayscale range of the reference image.

3.10.2 Registration of the images

Registration of the images is the most important step in a change detection approach. Hence, here we have proposed two different approaches; the rigid registration and non-rigid (deformable) registration. We considered the non-rigid registration methods to be able to deal with the brain atrophy changes that usually occur if the time interval between images is long (in our case 48M images). For this purpose, we studied 4 non-rigid methods that will be evaluated in the experimental section. On the other hand, concerning the rigid registration, we additionally presented halfway registration to remove interpolation artifacts.

3.10.3 White matter segmentation

We applied a WM masking step in order to reduce the search space for lesions only within white matter. We employed and tested two different approaches in the pipeline. In the first approach, we used a method based on single modalities with and without partial volume correction. We will analyze this strategy comparing the WM masks obtained from PD-w, T1-w and T2-w images separately. In the second approach, we included an atlas and the multisequence information on the PD-w, T1-w, T2-w images based on the study done by Souplet et al [113].

3.10.4 Thresholding of the subtraction image

Although determining the threshold could be done empirically by experts, here we presented two different automatic thresholding methods. In the unsupervised approach, we proposed an automated threshold providing a convenient trade-off between sensitivity and specificity. We considered MS lesion detections as those voxels that have intensity values in the subtracted images larger than the mean and multiplication of the standard deviation by a constant that will be evaluated in the next chapter. Afterwards, we compared this method with a supervised approach that relies on training the average intensity of positive activity and maximizing the Dice coefficient.

3.10.5 Postprocessing

Since the binary image outcome may introduce many false positives, we tried to reduce the rate of false positive detections by applying postprocessing methods. For this purpose, we removed the low intensity values in the baseline image and used the local intensity neighbor information for each candidate region in both the baseline and follow-up images to incorporate single time point segmentation in addition to a subtracted image and to avoid the errors caused mainly by the global thresholding. Afterwards, we also re-established this framework using supervised algorithms. Further, we examined the candidate lesions by using texture information that will also be evaluated in the experimental section. In the unsupervised approach, we joined both results obtained from PD-w and T2-w images using the intersection of the subtracted images after the postprocessing step. In contrast, in the supervised approaches, we used classifiers to determine lesions, including the information obtained from the ROIs in both the PD-w and T2-w sequences.

Chapter 4

Experimental results

4.1 Evaluation

As introduced in Chapter 2, the use of different data sets and evaluation measures has been a major obstacle to reviewing lesion detection methods. Ideally, approaches should be applied to a common database and compared to a ground truth. This is, however, very difficult due to the lack of common public databases of real images along with several controls and their ground truths and the fact that the methods are not publicly available.

For the study carried out in this work, we follow the strategy of validation with the ground truth of new lesions in follow-up images as described in Chapter 2 section 2.7.3. In this sense, only the annotations on the follow-up images were used to evaluate the pipeline. All annotations were made on PD-w images and semiautomatically delineated using JIM© software[1]. This software allows experts to manually define the contours of the MS lesions in the MR images.

In this chapter, after introducing the database we used to evaluate our proposals described in Chapter 3, we will present the evaluation measures that will be used in the analysis of the pipeline. Finally, we will present a comparison with the state of the art methods and a discussion of the results obtained, pointing out the important aspects of the proposed contributions.

4.2 Study population

Our database consists of data from 4 healthy controls and 20 different patients with clinically confirmed MS. Each patient underwent MR imaging using the same protocol

[1]Xinapse Systems, JIM software webpage, `http://www.xinapse.com/home.php`.

(T1-w, T2-w and PD-w). The scanner used was a 1.5T Siemens Simphony Quantum scanner, with 2D conventional spin-echo T1-w (TR 450 ms, TE 17 ms), and dual echo PD T2-w (TR 3750 ms, TE 14 / 86 ms). The field of view of the scans was $256 \times 256 \times 46$ (192×256 for two patients at 12M and 244×320 for the control image of two patients in the 48M set), resulting in a roughly 1×1mm in-plane pixel size. The section thickness was 3mm for all the sequences.

Two different MRI datasets acquired from MS patients can be distinguished according to the elapsed time between the patients' explorations. The first set (12M) is composed of 10 patients who underwent two studies acquired one year apart. On the other hand, the second set (48M) is composed of 10 different patients' studies acquired four years apart. Lesions in the baseline images were fully annotated by a trained technician and confirmed by expert radiologists. In contrast, the follow-up images, which were annotated by the same expert technician, included only the annotation of new lesions. In order to validate the new lesions, only the annotations on the follow-up images determined by experts were used to evaluate the pipeline. The healthy control dataset is composed of patients' studies acquired 1 years apart.

The 12M data contains a total of 177 lesions distributed as 53.7% small (3-10 voxels of which 62.1% had 3-6 voxels), 22.0% small-medium (11-20 voxels), 17.0% medium (21-50 voxels), 5.6% large (51-100 voxels), and 1.7% very-large lesions (101+ voxels). On the other hand, the 48M data contains a total of 152 lesions, 26.3% small (of which 62.5% had 3-6 voxels), 24.3% small-medium, 28.9% medium, 10.5% large and 9.9% very-large lesions. The aim of this grouping is to analyze the performance of the pipeline according to the different lesion sizes. A more detailed distribution of the 12M and 48M datasets can be found in Tables 4.1 and 4.2, respectively. Note that, 1 voxel refers to $0,003$cc in a 1×1mm in-plane pixel size with 3mm thickness.

The purpose of each set is different. With the first set (12M), we aim at evaluating the pipeline in a situation that is frequent in clinical practice. The main challenges of this scenario are: on the one hand, natural changes in brain tissues need to be discriminated from lesion growth, while, on the other hand, this lesion growth is, in this test set, very small, hence the algorithm needs to be able to detect small pathological changes. The second set of cases (48M) allows us to evaluate the pipeline in a much more extreme situation. In this scenario the long time elapsed between studies produces greater differences due to natural changes in the brain's morphology (i.e. atrophy). These changes are difficult to compensate for when registering the images and could compromise the performance of

Table 4.1: Study population for the 12M dataset. **N:** new ground truth lesions, **Range:** range for each of the lesion categories, **Median:** median number of lesions for each of the lesion categories. All lesion sizes defined in voxels: small (3-10); small-medium (11-20); medium (21-50); large (51-100); very-large (101+).

	Size					
	3-6	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	59	36	39	30	10	3
Range	0..19	0..18	0..9	0..9	0..4	0..1
Median	3.5	1.5	3	2	1	0
Patient1	1	0	0	2	0	0
Patient2	3	0	5	1	1	0
Patient3	11	7	3	2	1	0
Patient4	0	1	2	0	0	0
Patient5	0	0	3	4	0	0
Patient6	4	3	6	7	2	1
Patient7	1	1	1	0	0	0
Patient8	19	18	9	4	1	0
Patient9	13	2	7	9	4	1
Patient10	7	4	3	1	1	1

the pipeline. Therefore, it poses an even more challenging situation for the subtraction pipelines. In addition, further tests of the methodology were performed using an MRI dataset of 4 healthy controls with paired MRI scans.

4.3 Evaluation measures

In Chapter 2, we described different measures that have been used in the literature. From this analysis it was clear that the average inter-scan and the inter- and intra-rater coefficient of variation (COV) measures are widely used when the evaluation of the pipeline is carried out by visual inspection. On the other hand, there is a clear tendency to use sensitivity (true positive rate (TPR)), the false discovery rate (FDR) and Dice overlap coefficient (DSC) measures, particulary for a quantitative analysis of detection. Therefore, we evaluated our pipelines using these measures in a quantitative validation computed as follows:

$$Sensitivity = \frac{TP}{TP + FN}\,, \tag{4.1}$$

and

$$False\ Discovery\ Rate\ (FDR) = \frac{FP}{FP + TP}\,, \tag{4.2}$$

where TP, FN and FP refer to true positive, false negative and false positive detections,

Table 4.2: Study population for the 48M dataset. **N:** new ground truth lesions, **Range:** range for each of the lesion categories, **Median:** median number of lesions for each of the lesion categories. All lesion sizes defined in voxels: small (3-10); small-medium (11-20); medium (21-50); large (51-100); very-large (101+).

	Size					
	3-6	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	25	15	37	44	16	15
Range	0..4	0..5	0..7	1..13	0..6	0..7
Median	2.5	1	3.5	3.5	1	1
Patient11	4	2	7	2	1	1
Patient12	2	2	0	1	1	0
Patient13	1	3	6	6	0	0
Patient14	2	5	5	13	6	7
Patient15	4	0	2	5	1	1
Patient16	1	1	3	3	0	1
Patient17	0	0	4	6	2	3
Patient18	3	1	1	1	1	1
Patient19	4	1	2	4	1	1
Patient20	4	0	7	3	3	0

respectively. We considered a new lesion as detected correctly (TP) when there is a region detected by the pipeline sharing at least 1 voxel with the corresponding ground truth lesion. Likewise, a detection was considered an FP when there was no intersection between the detected lesion and the ground truth lesions by the automatic pipeline. Only lesions detected inside the WM mask were taken into account when computing the sensitivity and the Dice coefficient. A DSC, which is a common overlap metric between two binary masks, is computed as

$$DSC = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2 \times TP}{2 \times TP + FP + FN}\,, \tag{4.3}$$

where A represents the automatic segmentation mask and B the manual segmentation done by the expert. The Dice coefficient can be computed in different ways:

- Voxel-wise DSC (DSCV). This evaluation is performed by making a voxel-to-voxel comparison for lesion segmentation. The maximum value (1.0) is only reached when both images are equal (the automated detected lesions are totally matched with the manually annotated lesions). However, this maximum value is difficult to reach in practice, also obtaining low values for this measure is usual due to the inherent difficulty of the delineation process. Moreover, even if two lesions are perfectly matched but shifted by only a few voxels due to registration errors, the overlap will drastically decrease providing a very low value even though the lesion has been

detected and even had the right volume.

- Region-wise DSC (DSCR). The evaluation in this case is made at a region level: a TP occurs whenever the overlap between real and automatically located lesions exceeds a prefixed minimum. This measure is useful to evaluate the detection, but it is not a reliable segmentation measure (lesion delineation), since the size of the region is not considered.

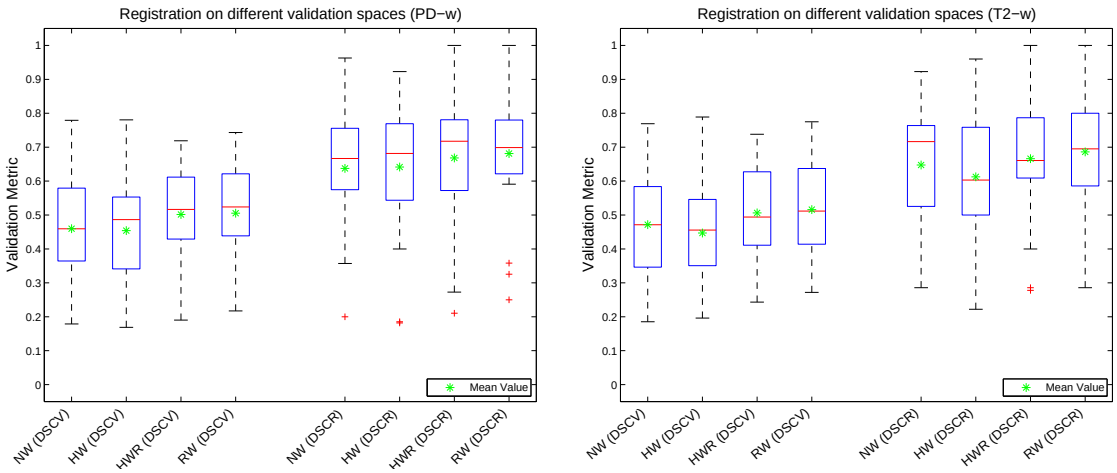
Furthermore, on account of a more reliable validation, similar to Elliott et al. [37], we also studied the performance of our approach according to different lesion sizes. We used a similar size division, except for the small size set, where we distinguish between lesions of 3-6 voxels and lesions of 7-10 voxels. This additional division is necessary since we have a large proportion of small lesions and, hence, we provide more details on these challenging detections.

4.4 Evaluating the validation space

We first evaluate the effect of the space where the evaluation is performed. In order to do this in an automatic way, the threshold selection is performed by maximizing the DSCV and DSCR coefficients individually for each case (the best possible threshold). We used this strategy based on the existence of a ground-truth image in order to emulate an expert's supervision. Figure 4.1 shows the algorithm's performance in terms of DSCV and DSCR detailing, for each of these measures, the result when using the validation space computed by the four registration strategies mentioned in Section 3.3. The figure shows how both reverse and halfway reverse registration performed similarly and obtained better results than the other two strategies. As for direct strategies (NW and HW), worse results were obtained with normal (NW) registration.

Hypothesis tests proved to be inconclusive, however, reverse strategies clearly outperformed direct ones. For example, RW obtained better results than NW in terms of DSC values in almost all cases. In some cases, these differences reached a perceptual difference. The reason why this behavior is not clearly supported by hypothesis tests is that the variability inside the data is very high, as can be seen in Figure 4.1. Consequently, although a clear behavior can be observed in the data, the diverse nature of our data base prevents us from using classical statistical inference.

Two main sources for errors in terms of validation can be identified. The first are



(a) (b)

Figure 4.1: Comparison of the registration methods on different validation spaces. (a) methods on PD-w sequences, (b) methods on T2-w sequences, Normal or direct registration (NW): the follow-up image is rigidly transformed to match the baseline study. The same transformation is applied to ground truth data, Inverse or Reverse registration (RW): the baseline image undergoes rigid transformation towards the follow-up image. The ground truth need not be transformed, Half-way registration (HW): the half-way registration where both the baseline and follow-up images undergo rigid transformation towards an intermediate space. The ground truth image is transformed using the same transformation as the follow-up image and Half-way Reverse registration (HWR): the same as the previous one except that the subtraction performed in the half-way space is rigidly transformed to the follow-up space so no transformation of the ground truth images is necessary.

interpolation errors and are introduced every time an image is moved. Consequently, all validation spaces contain them, although halfway approaches (HW and HWR) are expected to diminish their importance. The second type of error appears whenever we resample the ground truth images. Only inverse approaches (RW and HWR) are free from these errors. Looking at the results, an interesting fact that stands out is that the RW and HWR validation obtained the best results. For RW, this happens despite the fact that it does not reduce the interpolation effects in the way that HW and HWR do. This leads us to think that the negative effects of transforming ground truth data dominate tthe usual interpolation effects. This becomes clearer when comparing HW and HWR. Despite the fact that both use the same transformation matrix for registration, the HWR space has slightly better results (Figure 4.1).

Additionally, in some cases, HWR outperformed even RW and obtained the best result in all the validation spaces. However, it is still difficult to get a clear picture of the reduction of interpolation errors brought on by halfway approaches. The reason for this is

that halfway approaches only reduce and do not completely eliminate interpolation errors and no approach exists, in the fashion that happens with ground truth deformation errors that is free of them.

In summary, we claim that RW and HWR obtained better results because they allow a voxel-to-voxel validation without registering (modifying) the ground truth image. Therefore, in the following subsections, we present the results of reverse registration in order to elaborate on the performance of our pipeline. Similar trends were observed for HWR.

4.5 Evaluating the preprocessing steps

In order not to bias the impact of the preprocessing steps, we evaluated the bias field correction (normalization) and the histogram matching steps without applying the proposed automated thresholding, the PD-w T2-w combination, Gaussian filtering or any postprocessing. By this means we examined the impact of the preprocessing steps on PD-w and T2-w images separately based on the best possible threshold obtained by maximizing the DSCV and DSCR values using the ground truth on separate sequences.

Figure 4.2 illustrates the results obtained by the pipeline with and without normalization (bias field correction) and histogram matching. The results showed a distinct behavior between the two data sets. For instance, for the 12M data set, the differences regarding the use of the preprocessing steps were low. In terms of the DSCV and DSCR values, the use of the normalization step slightly improved the pipeline's performance, while histogram matching did not provide any significant difference. This was to be expected since the baseline and follow-up scans analyzed were acquired with the same scanning machine and protocol producing similar intensities for both scans.

In contrast, for the 48M data set, the results showed that both the normalization and histogram matching steps improved the performance particulary for PD-w images. Basically, the results obtained without any preprocessing and even after normalization were very bad. This can also be explained by scanning issues: images acquired by the scanners may differ more in terms of intensity after long time periods. This might happen because of a different performance of the scanner due to wear-and-tear, in other words extensive usage. Therefore, a histogram matching step between the two scans is both necessary and provides a greater improvement than with the 12M case. Once this step has been applied, results improve greatly and reach values similar to those obtained for 12M patients. Even further improvement is obtained when applying both histogram matching and normaliza-

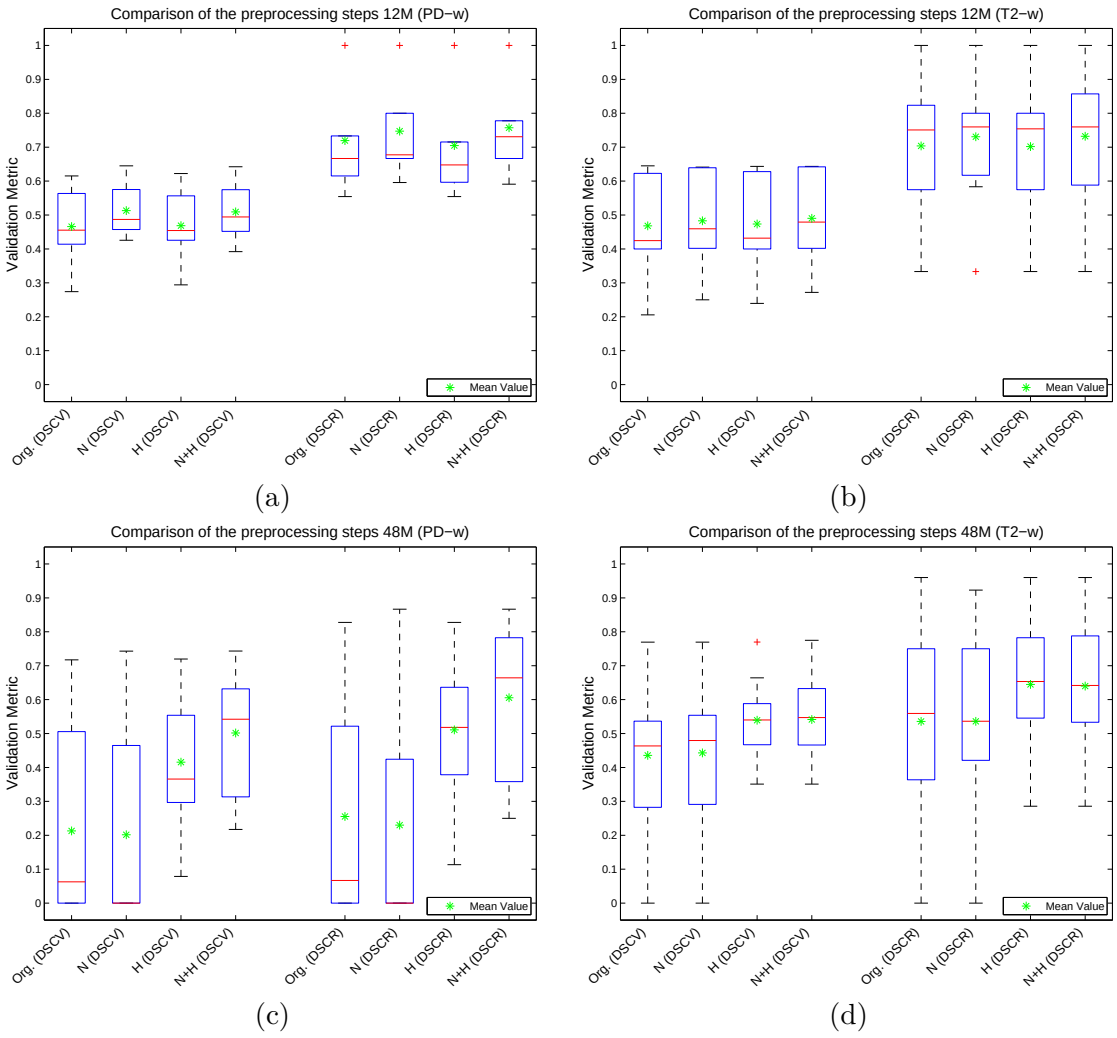


Figure 4.2: Analysis of preprocessing steps' performance in the pipeline. PD-w images from scans separated by (a) 12 months and (c) 48 months, and T2-w images from scans separated by (b) 12 months and (d) 48 months were used. Original (Org.): the pipeline without a bias field correction or histogram matching; Normalized (N): the pipeline with a bias field correction but without histogram matching; Histogram Matched (H): the pipeline with histogram matching but without a bias field correction; pipeline (N + H): the full pipeline including both histogram matching and a bias field correction.

tion. Figure 4.2 shows how the the pipeline with all the preprocessing steps reaches similar results for 48M and 12M patients.

In general, the performance of the pipeline was slightly higher for 12M images and significantly higher for 48M images when using both a bias field correction and the histogram matching steps (N+H) together, in particular for PD-w images while the bias field correction had a relatively small effect on T2-w images.

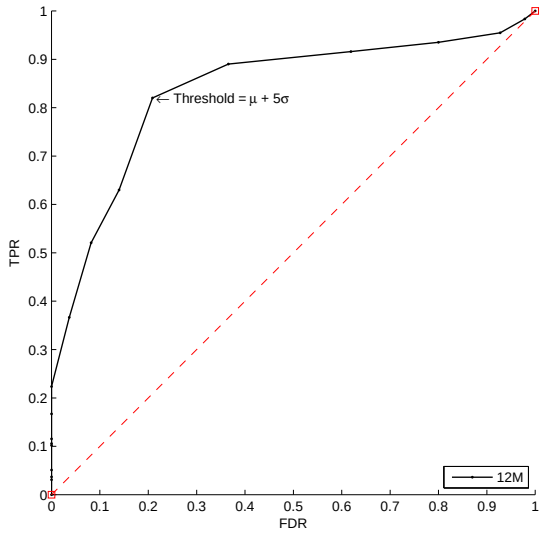
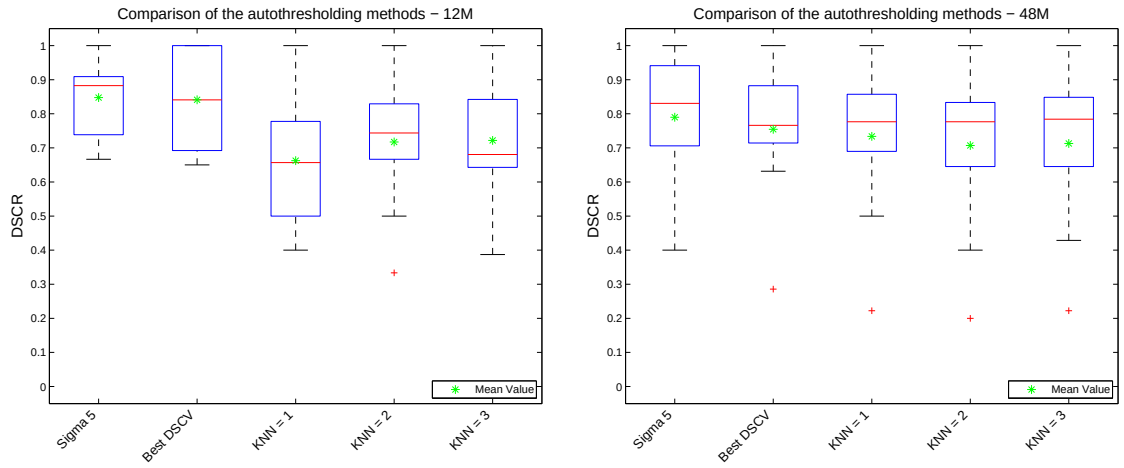


Figure 4.3: Lesion Sensitivity vs. False Discovery Rate for different operating points (12M).

4.6 Automated thresholding

4.6.1 Unsupervised thresholding

Before analyzing the results obtained for both 12M and 48M datasets, we present the procedure used to define the best operating point (parameter setting) of our approach. The 12M dataset was used to perform the training of a 2-fold cross-validation scheme where the data was divided into 2 groups of 5 patients each. One group was retained as the testing data and the other was used for training. We repeated the process changing the fold used for training and testing, using, therefore, all the cases once as validation data. The DSCR evaluation measure was used to analyze the results and optimize the various thresholds. Figure 4.3 shows the detection performance at different operating points. With these experiments, we observed that the threshold defined by $\mu + 5\sigma$ provided the best operating point for our approach in all tests. The postprocessing thresholds were also defined empirically in the cross-validation process. Afterwards, this 12M threshold configuration was applied to the entire 48M dataset (not used for training purposes) to provide the results shown in this work.



(a) (b)

Figure 4.4: Comparison of the automated thresholding methods. Automated thresholding results from the scans separated by (a) 12 months and (b) 48 months: sigma 5: results using the thresholds defined by $\mu + 5\sigma$, Best DSCV: results using the thresholds obtained by computing the best DSCV values, K = 1, K = 2, K = 3 : Results using the thresholds obtained by KNN training based on the best DSCV values.

4.6.2 Supervised thresholding

We have computed the correlations between the thresholds obtained by maximizing the DSCV values and the average intensity of the positive activity in WM masked subtracted images. These correlations reached 0.9043 and 0.9375 ($p < 0.0001$) for PD-w and T2-w images respectively. This would mean that the mean intensity of the outcome subtraction image could be used for fixing the threshold.

Figure 4.4 shows the detection performance with respect to the different automated thresholding methods. Analyzing the results, we observed that there were no significant differences for the various K values. However, when comparing the results, the KNN method performed better for the 48M images. This happened mainly due to fact that the 12M images contained more small lesions. Hence, after the combination of PD-w and T2-w images, there is more of a chance that small lesions could be lost since KNN thresholds are computed on PD-w and T2-w images separately. This suggests that one should use an underestimated threshold before combining the PD-w and T2-w images if intersection is used for the combining. This claim is also supported by comparing the unsupervised thresholding ($\mu + 5\sigma$) with the best possible thresholds and Figure 4.4 clearly shows that after the combining, using $\mu + 5\sigma$ performed better than the best possible thresholds computed on PD-w and T2-w images separately.

4.7 Evaluating the registration methods

Fig 4.5 shows the curves for different registration methods including rigid registration (RW), rigid halfway registration (HWR), non-rigid Nifty, the SyN and the Demons and Dramms methods. The black curve corresponds to the rigid registration. Observing the curves, we do not see a significant difference between the RW and halfway registration (HWR). On the other hand, the Demons and Dramms methods failed to reach desirable results for both the 12M and 48M datasets while the Nifty and SyN methods performed similarly to that of the rigid registration, though Nifty performed better with 12M images and the SyN method performed better on 48M images in terms of area under the curve. These results can be considered as reasonable since both the Nifty and SyN methods are topology preserved shape deformation algorithms whereas the Demons and Dramms methods do not.

Observing the curves, one can say that Nifty performed better than rigid registration on 12M T2-w images and the SyN method performed better than rigid registration on 48M T2-w images in terms of area under the curve. This might suggest that the Nifty and SyN methods could be useful under some circumstances when rigid registration fails. Nevertheless, the Nifty and Syn methods have a high algorithm complexity and more importantly after combination of PD-w and T2-w images, they perform similarly, in fact rigid registration is slightly better on optimum operating point ($\mu + 5\sigma$) despite the fact that the Nifty and SyN methods' performance look more consistent in terms of area under the curve. Moreover, we have seen that the deformation algorithms are more vulnerable to the modification of or removing small lesions. Consequently, we recommend using rigid registration on optimum operating point unless it fails to register.

4.8 Evaluating the white matter masking methods

We computed the performance of the different WM masks in the pipeline. The performances of the white matter masking methods using the proposed threshold is shown in Figure 4.7(b) in terms of DSCR values. The WM masks outcome can be seen visually in Figure 4.6. According to the results, only the atlas based WM method (the work done by Souplet et al [113] and Cabezas et al. [20]) and the T1-w WM method (by FSL segmentation tools (FAST) [151]) produced desirable results. The area under the curve can be observed in Figure 4.7(a), which clearly shows that the multi sequence atlas based method

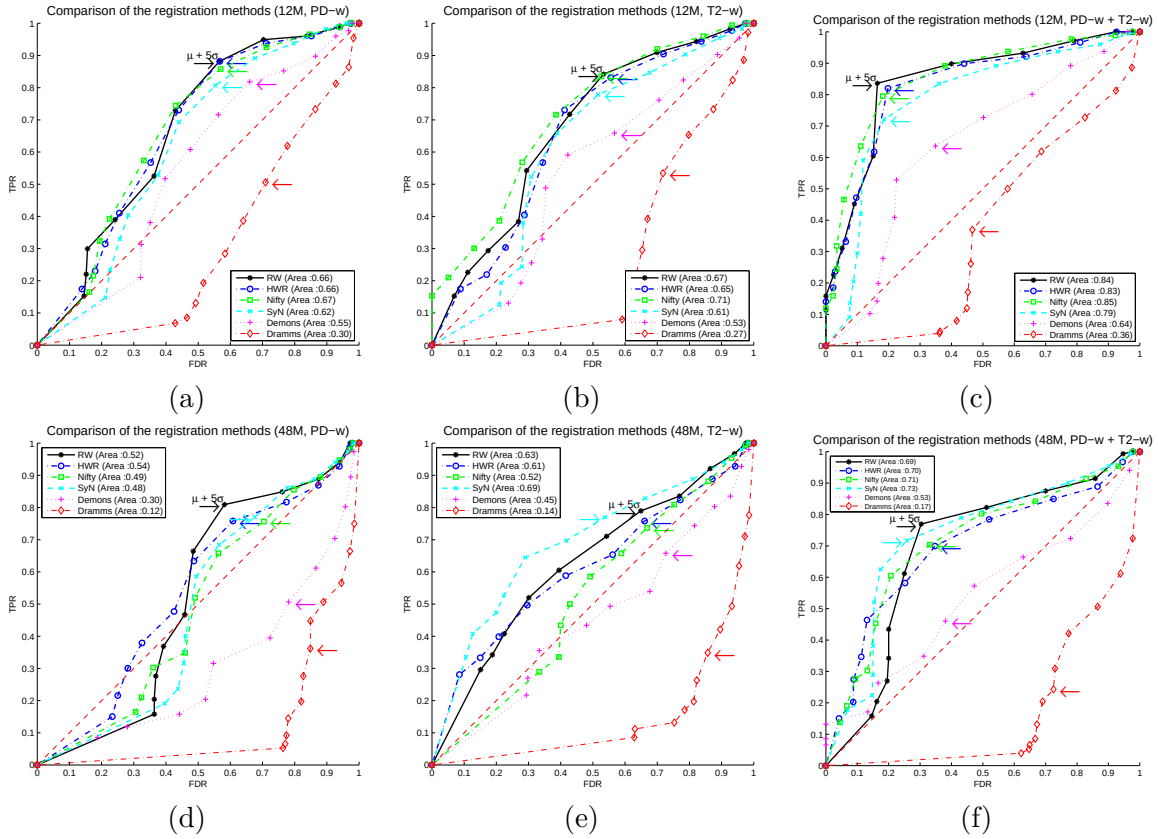


Figure 4.5: Analysis of the different registration methods for different operating points. PD-w images from the scans separated by (a) 12 months and (c) 48 months, and T2-w images from the scans separated by (b) 12 months and (d) 48 months, and Combination of the PD-w and T2-w images from the scans separated by (e) 12 months and (f) 48 months were used. RW: Rigid registration; HWR: Rigid Half-way registration; Nifty : Nifty non-rigid registration; SyN: SyN non-rigid registration; Demons: Demons non-rigid registration; Dramms: Dramms non-rigid registration.

outperformed the WM method obtained from a single T1-w modality.

On the other hand, the results also demonstrate that the curves follow the same tendency and that the thresholds defined between 4σ and 6σ produce desirable results. Furthermore, the threshold defined by $\mu + 5\sigma$ provided the best operating point regardless of the WM masking method used.

4.9 Evaluating the pipeline with different Gaussian filters

We tested the pipeline with different Gaussian filters. Figure 4.8(a) shows that using $\sigma = 0.5$ produced the best results in terms of both area under the curve and at the

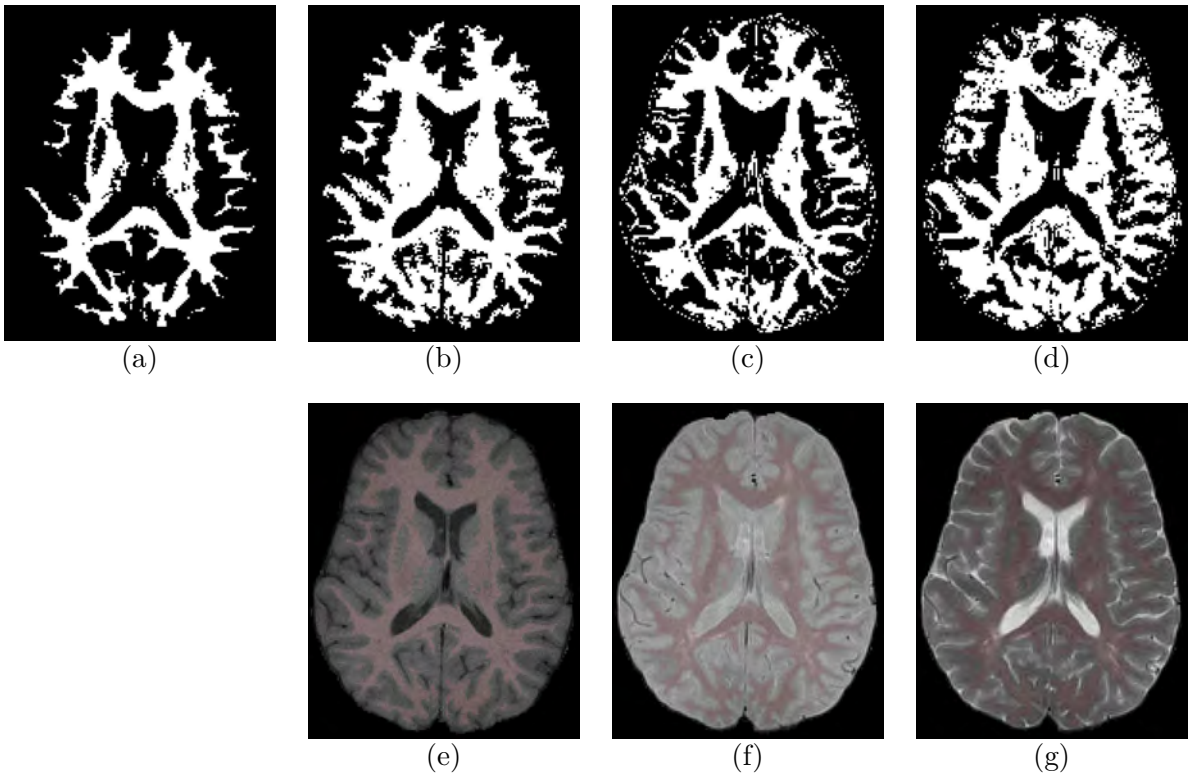
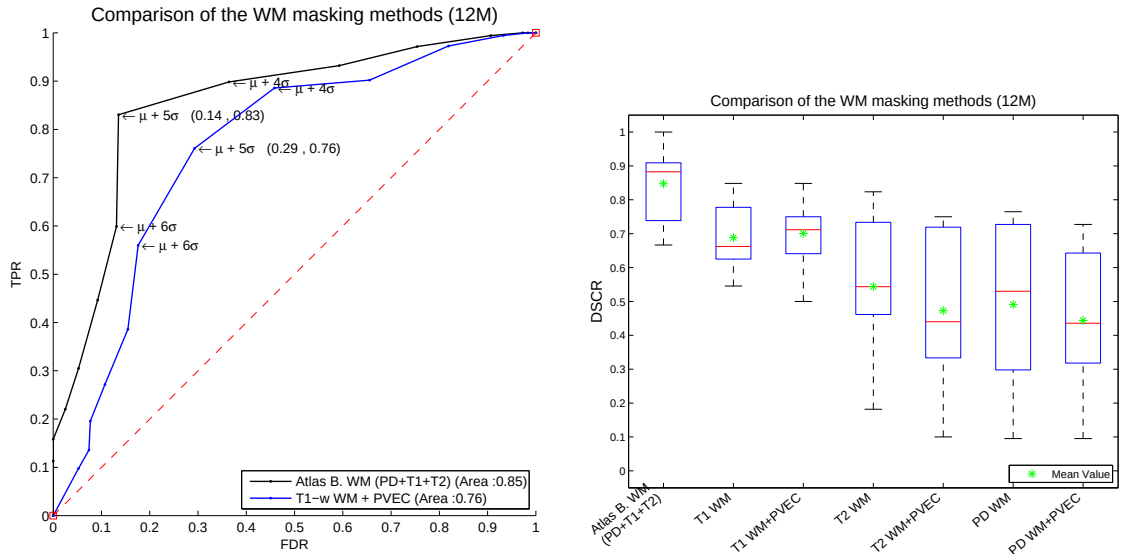


Figure 4.6: Comparison of the white matter masks. (a) Atlas based (PD + T1 + T2) WM, (b) T1-w WM, (c) PD-w WM, (d) T2-w WM, (e) Atlas based WM on PD-w image, (f) Atlas based WM on T2-w image, (g) Atlas based WM on T1-w image

threshold obtained by $\mu + 5\sigma$, and leads us to believe it could be used as the optimum threshold. On the other hand, all the Gaussian filters performed slightly better than the pipeline without gaussian filtering. Nevertheless, after the $\sigma = 0.5$, the area under the curve is prone to decline.

A comparison of the Gaussian filters with different voxel radius is shown in Figure 4.8(b). The results show that the Gaussian filters obtained by a 3 voxel radius perform better than those with a 5 voxel radius with the same σ values for 12M data. On the other hand, for 48M data, a 5 voxel radius performed better than 3 voxel radius in terms of area under the curve. However, if the curves are analyzed carefully, between the thresholds obtained by $\sigma = 4$ and $\sigma = 6$, the Gaussian filter produced by $\sigma = 0.5$ performed better than the 5 voxel radius. As a consequence, we suggest using a Gaussian filter obtained by a 3 voxel radius at $\sigma = 0.5$.



(a) (b)

Figure 4.7: (a) Lesion Sensitivity vs. False Discovery Rate for the detection of new lesions using different white matter masking methods: The T1-w white matter masking method corresponds to the blue curve and the multi sequence (PD-w + T1-w + T2-w) atlas-based white matter masking method corresponds to the black curve. (b) Performances of all the white matter masking methods on the proposed threshold in terms of DSCR values (PVEC: partial volume effect correction).

4.10 Detailed evaluation of the unsupervised pipeline for 12M and 48M datasets

We have analyzed the performance of the presented pipeline when using rigid registration and automated unsupervised thresholding. The threshold is defined in the subtracted images of PD-w and T2-w individually by using the mean and standard deviation of the positive changes. As mentioned in section 3.8, we empirically define a common threshold, which in turn was mean + 5σ of the positive activity, for all the cases and experiments presented here. Afterwards, the candidate lesions in the PD-w and T2-w images were also combined to reach the final detection of new lesions.

Table 4.3 summarizes the results obtained when using the proposed pipeline for the 12M and 48M datasets respectively. We have also indicated the performance according to different lesion sizes. Analyzing the results for the 12M dataset, we observe that, for all the groups with a lesion size greater than 10 voxels, the sensitivity reached was high (> 0.90), also obtaining a very low FDR. Notice that the large and very-large groups of

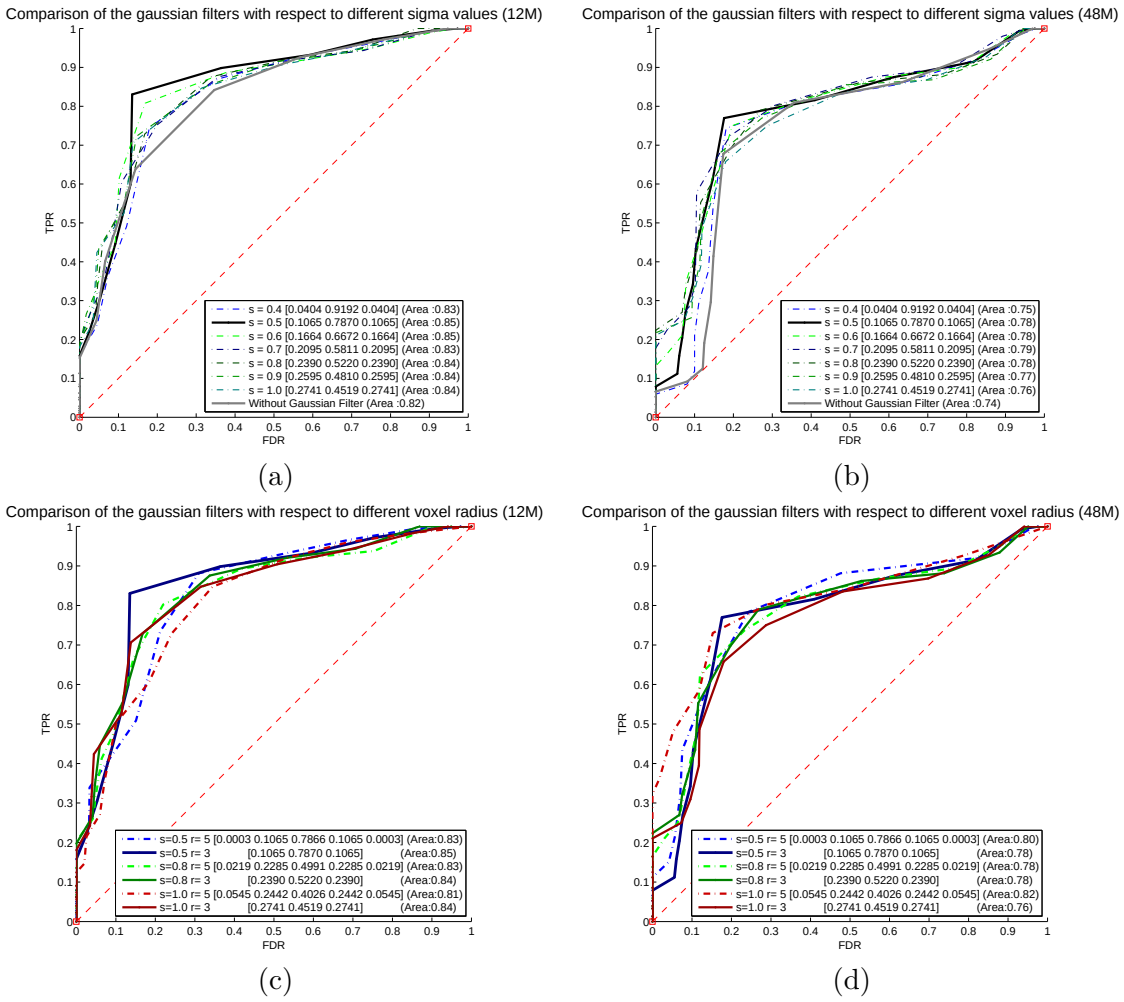


Figure 4.8: Analysis of Gaussian filter performance in the pipeline. Lesion Sensitivity vs. False Discovery Rate for the (a) 12M data and (b) 48M data under different Gaussian filters using 3 voxel radius (r=3) and different σ values (for $s = \sigma = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$), and the (c) 12M data and (d) 48M data under different gaussian filters using 3 and 5 voxel radius (r=3 and r =5) with the same σ values (for $s = \sigma = 0.5, 0.8, 1.0$).

lesions both obtained 100% TP detections at 0 FDR. However, when considering all the lesion sizes together, the overall performance was lower. This happened mainly due to the fact that our dataset is highly challenging, containing a high number of small lesions (53% of all the ground truth lesions were in the small size group). This can be appreciated with the performance obtained with lesions of 3-6 voxels. We also observed that 92% of the falsely detected regions also corresponded to this small size group (3-10 voxels).

Analyzing the results obtained per patient (see Table 4.4), we observed variability for each individual patient. This happens mainly due to the difference in number and size of

Table 4.3: Performance of the unsupervised pipeline by lesion size. **N:** new ground truth lesions,**TP:** number of true positive lesions, **FP:** number of false detected lesions, **SENS:** region-wise sensitivity, **FDR:** false discovery rate, **DSCR:** region-wise Dice coefficient. All the lesion sizes are defined in voxels: small (3-10); small-medium (11-20); medium (21-50); large (51-100); and very-large (101+).

			Size					
		Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
12M	N	177	59	36	39	30	10	3
	TP	147	39	33	34	28	10	3
	SENS	0.83	0.66	0.92	0.87	0.93	1.00	1.00
	FP	23	15	6	1	1	0	0
	FDR	0.14	0.28	0.15	0.03	0.03	0.00	0.00
	DSCR	0.85	0.69	0.88	0.92	0.95	1.00	1.00
48M	N	152	25	15	37	44	16	15
	TP	117	16	10	29	34	13	15
	SENS	0.77	0.64	0.67	0.78	0.77	0.81	1.00
	FP	25	14	4	3	2	2	0
	FDR	0.18	0.47	0.29	0.09	0.06	0.13	0.00
	DSCR	0.80	0.58	0.69	0.84	0.85	0.84	1.00

the lesions from one patient to another. Patients with larger lesions obtained better results than those with smaller lesions. Small lesions are more sensitive to misclassification and vulnerable in the PD-w and T2-w combination process. Regarding the region-wise DSC coefficient, we obtained a mean DSCR for the 10 patients of 12M of 0.85 ± 0.12, having also a mean TP detection rate of 0.91 ± 0.12 and a mean FDR of 0.17 ± 0.19. On the other hand, the result of the DSCV was 0.61 ± 0.08. In this dataset, we had five patients with perfect TP detection with two of them also having 0 FDR. The worst result corresponded to a patient with a DSC of 0.67 and an FDR of 0.50. Figure 4.9 illustrates some qualitative results obtained with the proposed approach, showing original images, subtracted images, manual segmentations provided by experts and automated detections.

We also evaluated the performance of the proposed pipeline for patients whose studies were taken 48 months apart (see Table 4.3). Except for very-large sized lesion set, the results were lower than those obtained for the 12M dataset. Please note that this dataset presented a particularly challenging situation for subtraction pipelines for two reasons: larger image intensity changes due to mostly brain atrophy changes that may cause registration errors and greater changes in lesion activity, which may produce more false positive detections since changes caused by growing lesions could turn out to be false positive detections. The overall sensitivity was 0.77 at an FDR of 0.18. Notice that, in this 48M

Table 4.4: Performance of the unsupervised pipeline per patient. **SENS.:** region-wise sensitivity, **FDR:** false discovery rate, **DSCR:** region-wise dice coefficient, **DSCV:** voxel-wise dice coefficient, **Avg:** average validation measure of all patients in the dataset.

	Evaluation measures - 12M					**Evaluation measures - 48M**			
	SENS.	**FDR**	**DSCR**	**DSCV**		**SENS.**	**FDR**	**DSCR**	**DSCV**
Avg	**0.91**	**0.17**	**0.85**	**0.61**	**Avg**	**0.80**	**0.20**	**0.79**	**0.60**
P1	1.00	0.50	0.67	0.50	P11	0.76	0.00	0.87	0.54
P2	1.00	0.17	0.91	0.65	P12	0.50	0.67	0.40	0.39
P3	0.79	0.00	0.88	0.63	P13	1.00	0.11	0.94	0.58
P4	1.00	0.50	0.67	0.60	P14	0.58	0.08	0.71	0.60
P5	1.00	0.00	1.00	0.72	P15	0.77	0.09	0.83	0.65
P6	0.91	0.22	0.84	0.57	P16	1.00	0.10	0.95	0.68
P7	1.00	0.00	1.00	0.49	P17	0.80	0.14	0.83	0.78
P8	0.80	0.02	0.88	0.71	P18	0.88	0.46	0.67	0.44
P9	0.67	0.17	0.74	0.54	P19	1.00	0.00	1.00	0.80
P10	0.94	0.16	0.89	0.66	P20	0.71	0.29	0.71	0.54

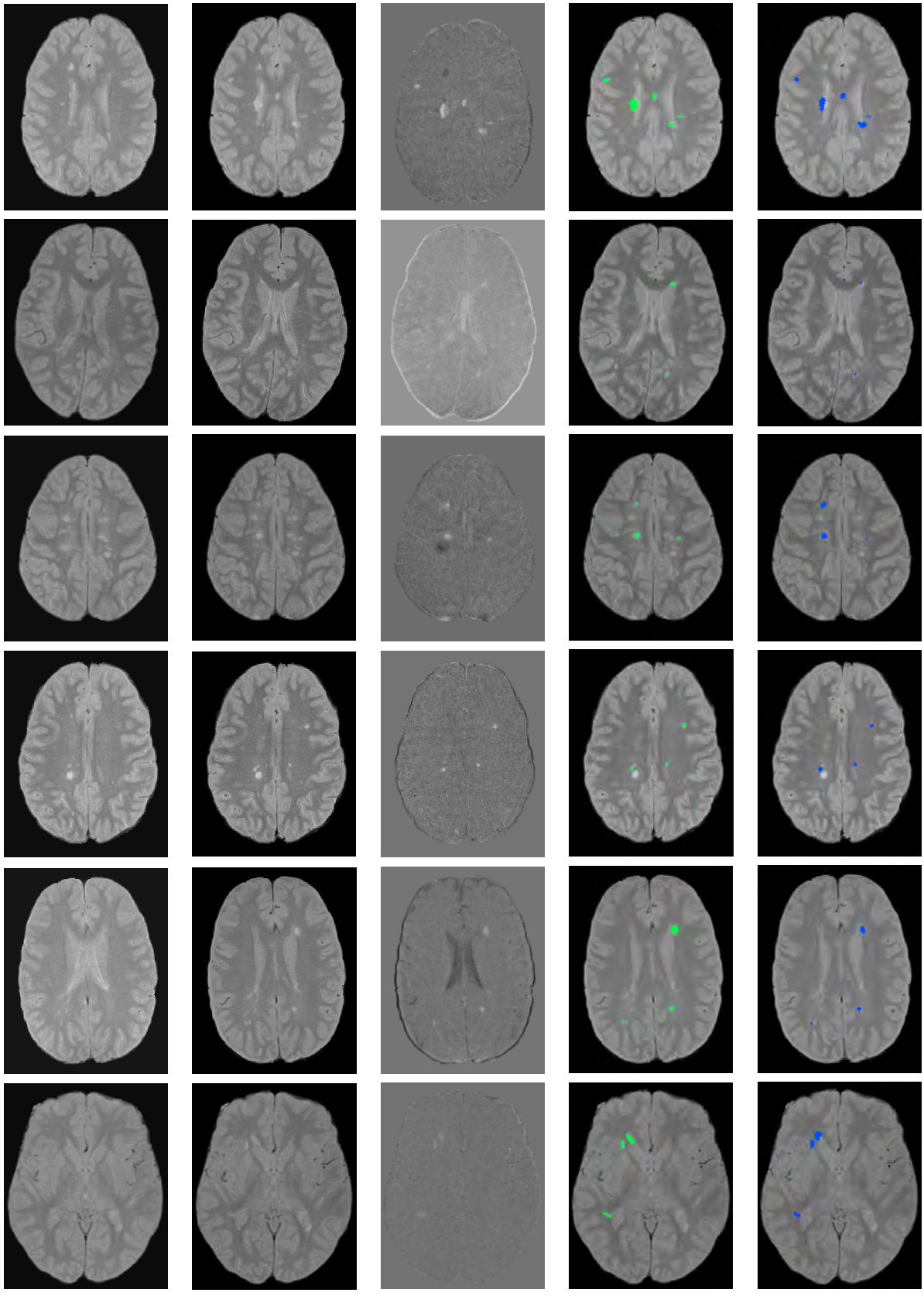
set we had 26% of the lesions belonging to the small lesion size group that had the lowest sensitivity and the higher FDR.

Regarding the results obtained per patient (see Table 4.4), we had a mean DSCR for the 10 patients of 0.79 ± 0.18. The mean TP detection was 0.80 ± 0.17, while the mean FDR and DSCV was 0.20 ± 0.22 and 0.60 ± 0.13 respectively. In this dataset, we had three patients with a perfect TP detection, one of them also having a 0 FDR. The lowest result was obtained for a patient with a DSCR of 0.40 and an FDR of 0.67.

Finally, the mean number of voxel for overall amount of lesions detected per healthy control subject was 20.75 voxels, which was 313.9 and 462.5 voxels in the 12M and 48M datasets, respectively. On the other hand, in two patients which have the minimum number of MS voxels in the 12M data, the pipeline detected 27 and 82 voxels against 30 and 42 voxels determined by experts in the ground truth. Consequently, MS patients were significantly found to have a higher number of detected voxels than healthy controls in our unsupervised pipeline.

4.10.1 Impact of image combination, postprocessing and WM masking

In this section, the impact of the different pipeline steps is evaluated. Table 4.5 summarizes the results obtained when using for instance a single image as input or when using combination (PD+T2), when using or not using the white matter masking (WMM) procedure, and when using or not using the postprocessing steps.



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

Figure 4.9: Visual examples of automated detections of new lesions in unsupervised pipeline. (a) baseline image, (b) follow-up image, (c) subtraction image, (d) manual segmentation, (e) automated detection.

When including the postprocessing steps but using only a single modality as input, PD-w or T2-w images without combination, the results obtained showed a higher number of false positives. For instance, in the 12M dataset when using only T2-w images the detection produced an overall FDR of 0.48 at a sensitivity of 0.84, and while when using PD-w images the overall FDR was 0.50 at a sensitivity of 0.88 (see Table 4.5 for both). Even though different σ values were also tested when using individual sequences, the results were worse, specially for small lesions. Notice that, when combining PD-w and T2-w sequences, the false positive regions were significantly reduced at the cost of decreasing a small number of true positive detections. This TPR was slightly higher when using a single modality. On the other hand, the improvement in terms of FP was significant when introducing an image combination. The results presented in Table 4.3 (12M) and Table 4.5 show that the use of both PD-w and T2-w images keeps true positive lesions stable while considerably reducing the false positive detections. After the combination, most of these missed lesions, 92% for 12M, were small (less than 10 voxels). This is not surprising since the combination might provide smaller detected regions, thus increasing the chance of eliminating those regions by the lesion size threshold.

Similarly, we analyzed the performance of the pipeline when not including the postprocessing steps but combining the results from PD-w and T2-w images (see Table 4.5). In this case, the combination step was also able to remove many FP detections providing an overall sensitivity of 0.83 and 0.77, and an FDR of 0.16 and 0.30 in the 12M and 48M datasets, respectively. Note that the sensitivity was similar to that of the full pipeline but with a higher FDR.

A detailed postprocessing analysis under different thresholds can be seen in Figure 4.10. Observing the curves, one should notice that the postprocessing step performs similarly regardless of the threshold, which always reduces false positives without loosing the true positive lesions. The postprocessing performance can be observed for 12M PD-w images in Figure 4.10(a), T2-w images in Figure 4.10(b) and after the combination in Figure 4.10(c). The improvement is relatively lower after the combination (see Figure 4.10(c)) since many false positives detected by the postprocessing step are also removed by combining the PD-w and T2-w images. On the other hand, for the 48M images (see Figure 4.10(a) for PD-w, Figure 4.10(b) for T2-w and Figure 4.10(c) after the combination) the postprocessing step removed many false positives that could not be removed by the combination since the 48M data mainly suffers from false positives due to the low signal areas in the baseline image caused by misalignments. This kind of false positive appears in both PD-w and T2-w images and cannot be directly removed by only combining them.

Table 4.5: Impact of image combination, postprocessing and WM masking. **SENS.:** region-wise sensitivity, **FDR:** false discovery rate, **DSCR:** region-wise dice coefficient, **DSCV:** voxel-wise dice coefficient, **WMM:** white matter masking

		SENS.	FDR	DSCR	DSCV
12M	PD+WMM+PostProcessing	0.88	0.50	0.64	0.51
	T2+WMM+PostProcessing	0.84	0.48	0.64	0.48
	PD+T2+WMM+PostProcessing	***0.83***	***0.14***	***0.85***	***0.61***
	PD+T2+WMM	0.83	0.16	0.84	0.61
	PD+T2	0.41	0.78	0.29	0.27
48M	PD+WMM+PostProcessing	0.81	0.48	0.63	0.58
	T2+WMM+PostProcessing	0.79	0.60	0.53	0.56
	PD+T2+WMM+PostProcessing	**0.77**	**0.18**	**0.80**	**0.63**
	PD+T2+WMM	0.77	0.30	0.73	0.62
	PD+T2	0.15	0.92	0.10	0.09

Analyzing in more detail the effect of postprocessing at the proposed threshold, we identified that the source of most of the FP regions appearing in the 12M dataset was caused by intensity changes due to growing lesions (36%), regions in ventricle zones (18%) and regions produced by image artifacts (18%). Additionally, for the 48M dataset and due to misalignment errors, some FP caused by misclassified WM and CSF tissue appeared (26%). The rest of the FP regions for both the 12M and 48M images were caused by GM tissue, lesion displacements due to brain atrophy, and low signal areas in the baseline image. After applying the postprocessing steps, we were able to reduce 21% and 51% of the total amount of FP in the 12M and 48M respectively. In the 12M, the neighborhood information of the follow-up image allowed for the removal of 17% of FP regions while losing only 1 TP. In the 48M dataset, the two intensity constraints applied to the baseline image allowed to correctly remove 47% of the FP regions without losing any TP.

Regarding the use of the WM mask, we also observed that the proposed approach missed some lesions due to their being outside the WM mask computed. In particular, in the 12M dataset, 10 lesions were missed (9 small and 1 small-medium size) while in the 48M dataset, 5 lesions were outside the WM (3 small, 1 small-medium and 1 medium size). In order to show the importance of the WM masking procedure, we repeated the experimental tests using the pipeline without introducing the WM mask, therefore, intensity changes outside WM were taken into account when applying the automatic thresholding step. However, this performance was far from the one achieved with the complete pipeline (see Table 4.5). For instance, in the 12M dataset, the overall sensitivity was 0.41 while the FDR was 0.78.

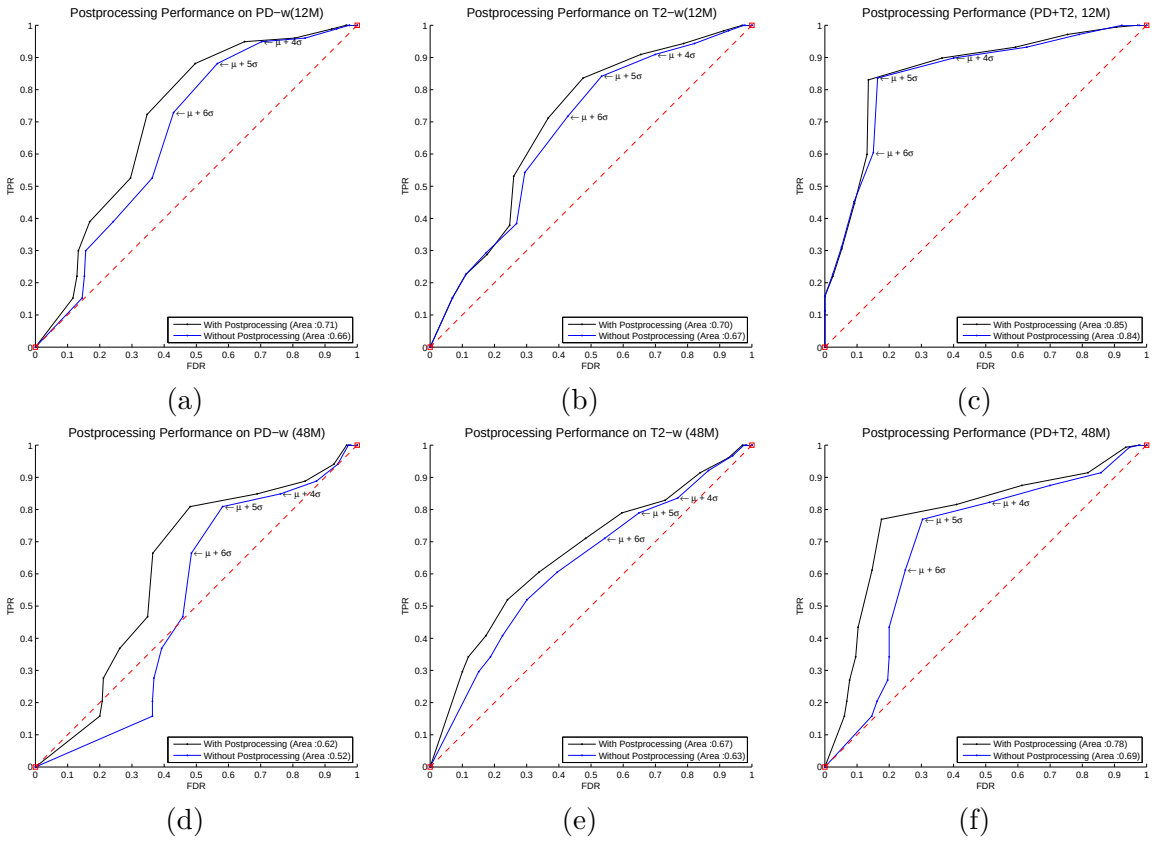


Figure 4.10: Analysis of the postprocessing steps. PD-w images from the scans separated by (a) 12 months and (d) 48 months, and T2-w images from the scans separated by (b) 12 months and (e) 48 months, and Combinations of the PD-w and T2-w images from the scans separated by (c) 12 months and (f) 48 months were used.

4.11 Comparison with state of the art methods

Even though there is no a common database for a quantitative comparison among different approaches, in this section we present a qualitative analysis with respect to the most recent proposals in the field. When comparing our results with those reported in the work by Elliott et al. [37], we observed that the performance of our pipeline is comparable and in some situations slightly better. For instance, in the 12M dataset, we obtained better results in terms of sensitivity and an FDR for the lesion groups small-medium, medium, large and very large, where they reported sensitivities of 0.74, 0.86, 0.98 and 1.00 with FDR of 0.17, 0.13, 0.02 and 0.02 respectively, in a dataset with a total of 336 new lesions. At the same time, our pipeline provided lower results in terms of FDR for the small lesion size, having a 0.23 FDR when considering both 3-6 and 7-10 voxel lesions in contrast to the 0.08 reported by Elliott et al. [37]. Also, in this case, our sensitivity was 0.76 against

their 0.61 value.

We also compared our detection results with those shown in the work by Battaglini et al. [11]. 19 MS patients with only 36 weeks of follow-up were used to report a sensitivity of 0.91 (116 detected lesions over 127 lesions manually annotated by experts) with a FDR of 0.21 (31 FP detections), also obtaining 11 FN lesions. In contrast, we obtained lower sensitivity (0.83) and better FDR (0.14). However, it is important to note that a description of the lesion size was not provided in Battaglini et al. [11], and therefore, a direct comparison with the quantitative results is a difficult task, since, as shown in our experimental tests, the performance of the automated detection depends highly on the size of the lesions.

Another recent example of automatic MS lesion detection in longitudinal analysis is the work by Sweeney et al. [118], where the authors evaluated their approach using 11 longitudinal studies each with a mean time between scans of 3 months. These cases had a total of 55 new or enlarging lesions annotated by experts. In this work, results were provided in terms of 3D volume voxel segmentation. Our pipeline provided an FPR of 0.00005 and specificity of 0.99995 at a sensitivity of 0.65 and 0.59 for 12M and 48M datasets respectively, approximately within the same FPR (0.00025) and specificity (0.99995) their approach provided with a lower sensitivity of 0.54 (see their detailed results in [118]).

4.12 Evaluating the pipeline when using supervised classifiers

Table 4.6 summarizes the results obtained when using the proposed pipeline with different classifiers. Although we found no significant differences between the classifiers, the SVM with polynomial function performed slightly better than the others. In this case, we obtained an overall sensitivity of 0.80 against a FDR of 0.12 for the 12M dataset, along with a DSCR of 0.83. On the other hand, the overall sensitivity was 0.76 at an FDR of 0.17 for the 48M dataset, which is clearly better than the direct intersection of the PD-w and T2-w images (see Table 4.5 48M PD+T2+WMM results). Figure 4.11 illustrates some qualitative results obtained with the supervised approach, showing original images, subtracted images, manual segmentations provided by experts and automated detections.

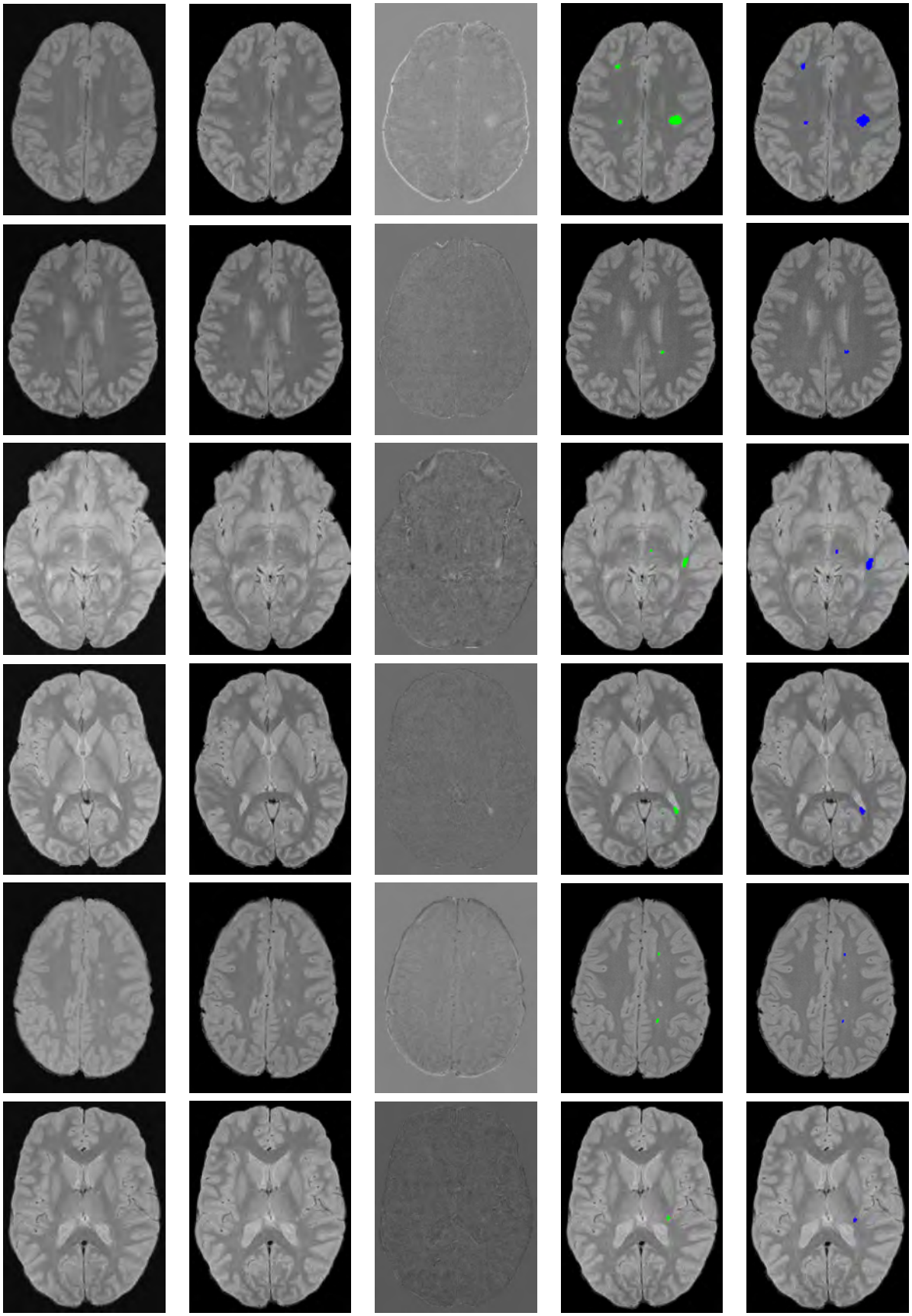


Figure 4.11: Visual examples of automated detections of new lesions in supervised pipeline. (a) baseline image, (b) follow-up image, (c) subtraction image, (d) manual segmentation, (e) automated detection.

Table 4.6: Supervised pipeline using BNR, FNR and BZS features **SVM:** Support vector machine, **KNN:** K-nearest neighborhood (K=1,3,5) **RBF:** Radial basis function, **SENS.:** region-wise sensitivity, **FDR:** false discovery rate, and **DSCR:** Region-wise dice coefficient

		SENS.	FDR	DSCR
12M	SVM Linear	0.79	0.19	0.79
	SVM Quadratic	0.79	0.13	0.83
	SVM Polynomial	0.80	0.12	0.84
	SVM RBF	0.78	0.15	0.81
	KNN (K=1)	0.72	0.20	0.76
	KNN (K=3)	0.75	0.19	0.78
	KNN (K=5)	0.77	0.18	0.79
48M	SVM Linear	0.78	0.22	0.78
	SVM Quadratic	0.76	0.21	0.77
	SVM Polynomial	0.76	0.17	0.79
	SVM RBF	0.77	0.17	0.80
	KNN (K=1)	0.70	0.24	0.73
	KNN (K=3)	0.72	0.19	0.76
	KNN (K=5)	0.73	0.19	0.77

4.13 Evaluating the supervised pipeline when using GLCM features

Among the classifiers, in general SVM classifiers outperformed the KNN classifier. The SVM classifier with linear function obtained slightly better results than the others, though we found no significant difference between any of them. Analyzing the obtained results, we have seen that the supervised pipeline with GLCM features obtained better results in terms of FDR, however, the sensitivity was lower, so that, for 12M, the linear SVM classifier reached 0.10 FDR against 0.77 of sensitivity. On the other hand, the result was 0.77 sensitivity against 0.20 FDR for 48M, which is better than the direct intersection of the PD-w and T2-w images (see Table 4.5 48M (PD+T2+WMM) results).

4.14 Comparing the unsupervised and supervised pipelines

A comparison of the supervised and unsupervised pipelines is shown in Table 4.8. In the first place, all the pipelines performed better than the pipeline without postprocessing and also obtained better results than the pipeline using only the intersection of PD-w and T2-w images.

The results demonstrated that the unsupervised pipeline performed better in terms of sensitivity whereas the supervised pipeline reached a better performance in terms of

Table 4.7: Supervised pipeline using GLCM features; Contrast, Energy and Homogeneity. **SVM:** support vector machine **KNN:** K-nearest neighborhood (K=1,3,5) **RBF:** radial basis function **SENS.:** region-wise sensitivity, **FDR:** false discovery rate, **DSCR:** region-wise dice coefficient

		SENS.	FDR	DSCR
12M	SVM Linear	0.77	0.10	0.83
	SVM Quadratic	0.77	0.12	0.82
	SVM Polynomial	0.67	0.21	0.72
	SVM RBF	0.74	0.15	0.79
	KNN (K=1)	0.73	0.20	0.76
	KNN (K=3)	0.77	0.16	0.81
	KNN (K=5)	0.79	0.19	0.80
48M	SVM Linear	0.77	0.20	0.79
	SVM Quadratic	0.75	0.20	0.78
	SVM Polynomial	0.71	0.19	0.76
	SVM RBF	0.72	0.16	0.77
	KNN (K=1)	0.70	0.29	0.70
	KNN (K=3)	0.70	0.26	0.72
	KNN (K=5)	0.75	0.23	0.76

Table 4.8: Comparison of the supervised and unsupervised pipelines **SENS.:** region-wise sensitivity, **FDR:** false discovery rate, and **DSCR:** region-wise dice coefficient

		SENS.	FDR	DSCR
12M	The unsupervised pipeline using BNR, FNR and BZS features.	0.83	0.14	0.85
	The supervised pipeline using BNR, FNR and BZS features.	0.80	0.12	0.84
	The supervised pipeline using GLCM features.	0.77	0.10	0.83
48M	The unsupervised pipeline using BNR, FNR and BZS features.	0.77	0.18	0.80
	The supervised pipeline using BNR, FNR and BZS features.	0.76	0.17	0.79
	The supervised pipeline using GLCM features.	0.77	0.20	0.79

FDR for both 12M and 48M data. For instance, for 12M images, the sensitivity of the unsupervised pipeline was 0.83 against 0.80 sensitivity with the unsupervised pipeline, however, their false discovery rates were 0.12 to 0.14 in favour of the supervised approach. On the other hand, mean number of voxels detected in the healthy control subjects is almost same. Like the unsupervised pipeline, the supervised pipeline detected only a mean number of 23 voxels for overall amount of lesions per healthy control subject.

On the other hand, using GLCM features, we obtained better FDR at the expense of decreasing the sensitivity. So that, for 12M, the pipeline reached 0.10 FDR and 0.77 sensitivity with a similar DSCR to the other pipelines (0.83).

4.15 Discussion

In this work we have implemented a well-known approach for MS lesion detection in serial analysis by adding additional steps to a subtraction pipeline. The problem being addressed in this work is very challenging since the size of the lesions being dealt with is very small in absolute terms.

As discussed in the validation section, the decisions made in terms of registration may affect the validation accuracy (Figure 4.1). These decisions are mainly the choice of source and target images and whether or not ground truth images need to be transformed. According to our experience, reverse registration is the best approach in order to obtain an accurate, repeatable one-to-one voxel-wise comparison. Besides, according to this strategy, the obtained ground truth remains free of any transformation that could compromise its integrity by introducing interpolation errors.

We also studied the effect that preprocessing steps have on the pipeline results (Figure 4.2). Although the pipeline including image normalization (but not histogram matching) slightly outperformed original images, and one can notice that the differences are small, particularly for the T2-w images. Thus, concerning practical considerations, bias field correction may not be necessary for T2-w images. On the other hand, when images are acquired in very separated time or even from different scanners, not only are the bias fields different but the mean intensity of the images might also have changed. For this reason, bias field correction itself is not sufficient, in particular when the gray value intensity scales are different. In this case, the histogram matching step is also necessary.

Concerning the thresholding of the subtraction image, we demonstrated that the average intensity of the positive activity can be used to automatically define the threshold. Using the average intensity of the positive activity, we presented two different approaches. In the first one, we followed an unsupervised strategy including the standard deviation of the positive activity. We concluded that using $\mu + 5\sigma$ produced the best results. However, thresholds defined between 4σ and 6σ also produced desirable results and can be chosen by experts concerning the trade-off between the sensitivity and specificity (Figure 4.3). In the second approach, we followed a supervised strategy using the correlations between the average intensity of the positive activity and the thresholds determined by maximizing the DSCV value in the sequences. The results showed that the unsupervised strategy slightly outperformed the supervised one and moreover, provided various thresholds using a standard deviation (Figure 4.4). However, one should also consider that in supervised

thresholding, we trained the thresholds using only 20 patients with a leave-one-patient-out strategy. Hence, using a database with more patients for training could provide better results.

The pipeline presented in this study has some potential limitations. So that, the performance of the pipeline depends mostly on the accuracy of the WM masking method and registration between the images. Both the WM masking and automated thresholding steps rely on the assumption that the white matter tissue follows a gaussian distribution. Hence, more accurate WM masks will lead to a better performance. Testing various WM masks (Figure 4.6), we concluded that the atlas-based WM method generated from PD-w, T2-w and T1w sequences produced the best results. However, a T1-w WM mask can also be considered concerning the algorithm complexity and computing time. One should also consider that 2 lesions were outside of WM mask when using T1-w WMM method whereas atlas-based WM method missed 15 lesions. Nevertheless, considering the performance of the methods, it can be negligible since the lesions missed by the atlas-based method have very small sizes. More importantly, regardless of what WM masking method is used, the pipeline showed the same tendency that thresholds defined between 4σ and 6σ provided a good trade-off between the sensitivity and specificity. Moreover, as shown in our experimental results, this step reduced the number of false detected regions while increasing the sensitivity individually in both PD-w and T2-w images in the WM tissue. However, the use of this WM masking process prevented the detection of gray matter lesions. Moreover, the FP and the FN detected by our pipeline could also be attributed to small errors in the results of the tissue segmentation. Indeed, the analysis of the results highlighted that most false positive lesions could be easily identified in the brain's regions by experts. We believe that better tissue segmentation results may be obtained with the use of high-quality 3D images (i.e. 3T) that will help improve both the sensitivity and FDR.

Furthermore, we have shown that the inclusion of simple postprocessing steps as well as the combination of both PD-w and T2-w images eliminated individual artifacts and reduced FP while keeping the detected lesions stable (Table 4.5 and Figure 4.10). We believe that PD-w and T2-w images are well-suited for combination considering also that they are routinely acquired in clinical practise.

The pipeline including rigid registration provided a mean DSCR of 0.85 ± 0.12 and 0.79 ± 0.18 for the 10 patients of 12M and 48M respectively. However, we have seen how the sensitivity and FDR depend highly on the distribution of the lesions' sizes, small lesions being more difficult to detect. On the other hand, 48M results showed that the pipeline

is still vulnerable to anatomical changes in the brain. In this sense, we studied several deformable registration algorithms to provide a solution to this problem (Figure 4.5). We have demonstrated that, while the non-rigid algorithms that do not provide topology preservation failed to produce desirable results, the topology preserved shape deformation algorithms (Nifty and SyN) outperformed the rigid registration in some cases. Although the deformation algorithms are prone to remove small lesions (particularly lesions less than 10 voxels), they can compensate for errors, something rigid registration cannot do. As a consequence, these algorithms can be included as an option with a potential application that could be used in clinical practice.

Additionally, we have demonstrated that the postprocessing methods we used in the unsupervised pipeline can be carried out by using supervised algorithms. Using the same features (BNR, FNR and BZS values), the pipeline reached similar results as the unsupervised pipeline. Additionally, we have also examined and tested the inclusion of texture features of the candidate lesions obtained from GLCM. In this case, more false positive lesions were removed, at the expense of losing some true positives. However, the supervised method used in this pipeline could be improved if the intensities across the datasets are normalized. Furthermore, it is often beneficial to scale all features to a common range, however, standardization is not appropriate when the data is sparse since it may destroy the sparsity [12]. Consequently, the features obtained by both the first and second order statistics could be useful for experts when determining new lesions. Having seen that, our pipeline is well-suited to providing candidate lesions to be further analysed in detail using these features. One should also consider including more features obtained from GLCM, run-length matrix (RLM), and additionally, some spectral approaches (Fourier, Wavelet and Stockwell transforms), that could also be used for this purpose [150].

Finally, we analyzed the images from healthy controls in order to identify residual registration and flow artifacts. Despite the fact that some false positive lesions were identified on the subtraction images of healthy controls, the mean number of voxel detected significantly differs between patients and healthy controls, in fact, the mean number of voxel is very low in healthy subjects when compared to MS patients. Consequently, the results showed that both supervised and unsupervised pipelines are proved to be robust in preventing detection of false positive changes arising from potential confounds such as registration errors and flow artifacts. See Section A.1 for the performance of the pipeline according to the different lesion sizes per patient.

Chapter 5

Conclusions

5.1 Summary of the thesis

The goal of the research presented in this thesis was to propose a new pipeline capable of detecting new MS lesions in magnetic resonance imaging. Starting with an initial study of the state-of-the-art of MS lesion detection, we classified the detection approaches in two primary categories concerning the quantification of MS lesion progression; those based on determining a lesion's volume difference after detecting MS lesions at a single time MR volume of a patient and those change detection methods relying on analyzing the differences between successive MRI controls. Following a change detection strategy, a fully automated and improved subtraction pipeline, developed for detecting new MS lesions in brain MRI data, has been described in Chapter 3 and exhaustively validated in Chapter 4.

In the first place, considering the fact that the quality of subtraction images relies on the quality of the registration and intensity correction procedures, we included skull stripping, bias field correction and histogram matching steps before the subtraction and analyzed them in detail in order to determine quantitatively the impact of the these steps on the pipeline. On the other hand, we evaluated 2 rigid and 4 non-rigid registration methods in the pipeline concluding that the rigid body registration is well-suited for this purpose, whereas the deformation algorithms that preserve the topology could outperform the rigid registration in some cases, thus, topology preserved shape deformation algorithms should also be considered as an option with a potential application that could be used in clinical practice.

Furthermore, we applied a WM masking step in order to reduce the lesion search space only in the white matter. In order to achieve our goal, we presented two different WM masking methods and compared them. In the first approach, we used FSL segmentation

tools which segment 3D brain MR images based on the hidden markov random fields (HMRF) model fitting by the EM algorithm using a Gaussian estimation. We concluded that the WM masks obtained from T1-w images including partial volume correction yielded a better performance. In a second approach, we used an atlas-based segmentation algorithm including PD-w, T1-w, and T2-w images, which was developed by our group and implemented in the subtraction pipeline. The results demonstrated that the multi sequence atlas based method outperformed the WM method obtained from a single T1-w modality.

Analyzing the literature, we have seen that the selection of the threshold is still an open issue. Hence, we focused on this issue and presented a novel thresholding method demonstrating that the average intensity and standard deviation of the positive activity can be used to define the threshold. Moreover, we presented a supervised thresholding method using the correlations between the average intensity of the positive activity and the thresholds determined by maximizing the DSCV values.

After a thorough study of the state-of-the art on MS lesion detection approaches, we concluded that statistical-based approaches including multi-sequence information yield a better performance. As a consequence, we combined PD-w and T2-w images and also presented some postprocessing methods based on the statistical analysis of the intensity features including baseline and follow-up images in order to reduce false positives. Firstly, we presented an unsupervised method. Basically, we removed the low intensity values in the baseline image and used the local intensity neighbor information for each candidate region determined by the automated threshold in both the baseline and follow-up images using fixed constraints. We compared this pipeline with state of the art methods.

Afterwards, we demonstrated that the postprocessing methods used in the first pipeline can be carried out by using supervised algorithms instead of fixed constraints and direct intersection of multisequence information. The supervised pipeline reached similar results to those obtained by the first pipeline. Furthermore, to improve this idea, we included texture features from the candidate lesions. In this case, more false positive lesions were removed at the expense of losing some true positives. As a consequence, we demonstrated that our pipeline is well-suited to providing candidate lesions to be further analyzed in detail using several features in order to remove more false positive detections.

To evaluate these approaches, we presented a reliable validation method based on the new lesions that are annotated by experts in the follow-up space. Furthermore, we discussed and evaluated the validation spaces concluding that the validation must be carried

out without registering (modifying) the ground truth image. The pipeline was tested in two major scenarios: the first with patients whose studies were taken 12 months apart stood for typical everyday clinical use. The second with studies taken 48 months apart presented a much more challenging situation due to natural changes in the brain's morphology (i.e. atrophy). The results obtained showed the validity of this approach, obtaining comparable and even better performances than those of recent state of the art methods.

Additionally, our first approach is a simple, automatic and unsupervised method that does not depend on an expert who optimizes a training dataset thus avoiding the variability between experts when making the annotations. On the other hand, this pipeline can be improved by using supervised classification algorithms including more features such as texture and region properties from the candidate lesions. Additional features can help radiologists to determine new lesions more accurately. Furthermore, the proposed pipeline could also be adjusted by experts per each case in order to increase the number of lesions detected at the price of obtaining more false positives. This pipeline flexibility makes it suitable for both automatic and semi-automatic operation. We believe that it could easily be adapted to monitor other brain pathologies such as volumetric changes in patients with vascular disease or tumors.

A prototype of the pipeline implemented by using C++.NET is shown at Figure 5.1. Note that, the postprocessing steps in the pipeline have been carried out using Matlab, which needs implementing into C++.NET platform.

5.1.1 Contributions

The goal of this thesis is to aid radiologists in their day-to-day practise by assisting them in the challenging task of detecting new MS lesions. Idealistically, our proposal should accurately detect and segment all the new lesions of any given patient. However, a more realistic expectation is to allow experts to process a batch of patients off-line in order to accurately detect a majority of the new lesions, reducing the experts interaction with the images to the correction of some segmentations or the detection of a small number of missing lesions. By reducing the interaction, we are also reducing the inter- and intra-observer variability. From this point of view, the main contributions of this thesis to both the scientific and medical communities are:

- A comprehensive survey of MS lesion detection algorithms and a classification of the approaches for automatic monitoring of MS lesion evolution. Analyzing the studies,

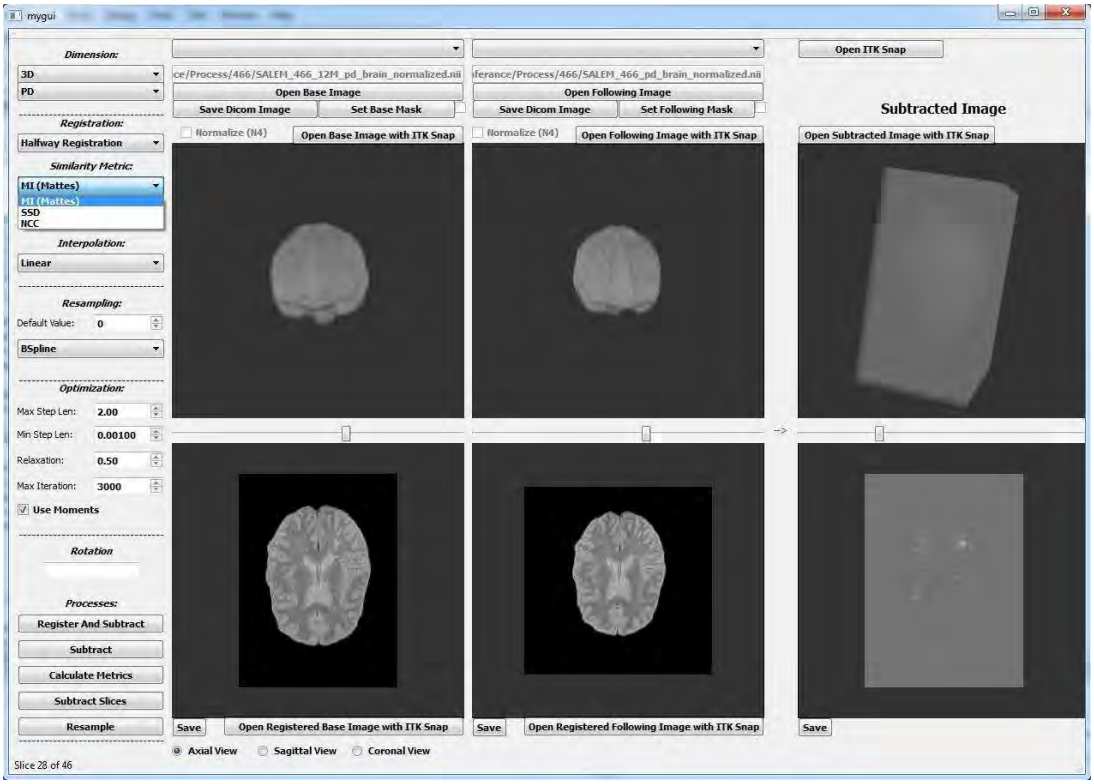


Figure 5.1: A prototype of the pipeline implemented by C++.NET.

we concluded that change detection algorithms suffer from image aligning problems and various artifacts that can be compensated for by using robust registration algorithms and preprocessing methods as well as combining different approaches such as using multi-sequence information and statistical analysis of the features attributed to the candidate lesions.

- A novel database for the SALEM project with 1.5T imaging data for 20 cases with different lesion loads. Two different datasets have been prepared: studies acquired one year apart (12M) to experience situations frequently occurring in clinical practice, and studies acquired four years apart (48M) to evaluate the pipeline in much more challenging situations. The annotations were done by experts to obtain ground truth of new lesions.

- A reliable validation strategy quantitatively demonstrating the pitfalls of the current validation methods.

- A quantitative analysis of preprocessing, rigid registration, non-rigid registration, WM segmentation and postprocessing methods.

- A novel automated thresholding strategy to provide a fully automated pipeline.
- A new, fully automatic and unsupervised pipeline to detect new lesions including lesion features.
- An alternative pipeline using supervised postprocessing of lesion features and texture properties.
- An experimental quantitative comparison between two proposed pipelines and a promising state-of-the-art approach in terms of both detection and segmentation.
- A prototype to detect and segment new lesions to be tested in hospitals. This tool has been implemented in C++.NET and is currently being implemented at Hospital Vall d'Hebron as a set of console functions and scripts.

5.2 Future work

5.2.1 Short term future improvements

The pipeline presented in this thesis has been exhaustively tested with 20 patients in real cases. Our first future work is to expand the current SALEM database to validate our proposals for various cases. Furthermore, our database lacked FLAIR images, so, we expect to improve the performance of the pipeline using FLAIR images. On the other hand, the studies described in this thesis were performed at 1.5T. In the near future, the study is to be tested for 3T, which has been found to provide better visualization and could lead to an increased detection of new lesions in the pipeline as well.

False positive detections could be further reduced by including more features into the pipeline. In our second approach, we demonstrated that these features can be well used in a supervised pipeline to determine new lesions more accurately. However, we only included some texture properties obtained from GLCM in addition to neighbor and intensity properties from the lesions. Therefore, the algorithm needs to be tested with more intensity, shape and texture features that could be obtained from run-length matrix or some spectral approaches like the Fourier, Wavelet and Stockwell transforms discussed in Chapter 2. Furthermore, using an EM algorithm likelihoods of the candidate lesion areas can be obtained to be included along with the other features. These features could be included in the postprocessing step of the supervised or unsupervised pipelines and could also provide additional information to experts to help them to determine new lesions more accurately,

reducing the intra-observer and inter-observer variability. Furthermore, the supervised method used in this pipeline can be improved by normalizing the intensities across the datasets. Implementing other supervised techniques would be other future work.

Other improvements to the pipeline would be carried out by improving the preprocessing steps and registration methods. The eyes or other non-brain tissues from the images of some patients are likely to be misclassified as white matter tissue, which may decrease the detection accuracy. Therefore, more study should be carried out in this regard and alternative skull stripping techniques should also be considered.

Other future work would be to examine candidate lesions according to their positions in the brain, classifying them into periventricular, juxtacortical or infratentorial lesions as presented in Chapter 1. For instance, most of the misclassified regions were in the periventricular due to the ventricles. An atlas could be easily adapted into the pipeline to provide prior information on the spatial properties of the lesion. This could also aid the MS diagnosis in those patients with a first clinical episode according to the McDonald criteria.

5.2.2 Future research lines

In the long term, there are several new research lines departing from this thesis that could be studied by the group. In this thesis, we have demonstrated that the topology preserved shape deformation algorithms (such as the Nifty and SyN methods) could outperform rigid registration in a subtraction pipeline, whereas other, more flexible techniques like the Dramms and Demons methods can not. However, a non-rigid registration method can also be used to display disease activity in an alternative way, as we pointed out in Chapter 2 under the title of deformation field-based approaches, by using vector displacement fields or deformation field morphometry derived from the deformation algorithm applied to register the images. In this case, that a non-rigid method removes the lesions by deforming them is not a disadvantage anymore. On the contrary, it becomes an advantage of the method since the information on the deformed lesion is stored in the deformation fields. Therefore, extracting this information from the non-rigid method would be a source of information that could help in the detection of new lesions. For this purpose, The Demons method would be a good option since it produces a displacement field that maps the moving image onto the fixed image.

On a separate note, the methods and concepts presented here could also be applied to other diseases that share similar properties as MS lesions, such as, lupus lesions appearing

in WM, which are hyperintense and can appear near the ventricles; stroke lesions can also appear as hyperintense lesions of variable sizes in T2-w images; and tumors usually appear as large hyperintense areas that deform the tissues surrounding them.

Appendix A

Detailed evaluation results

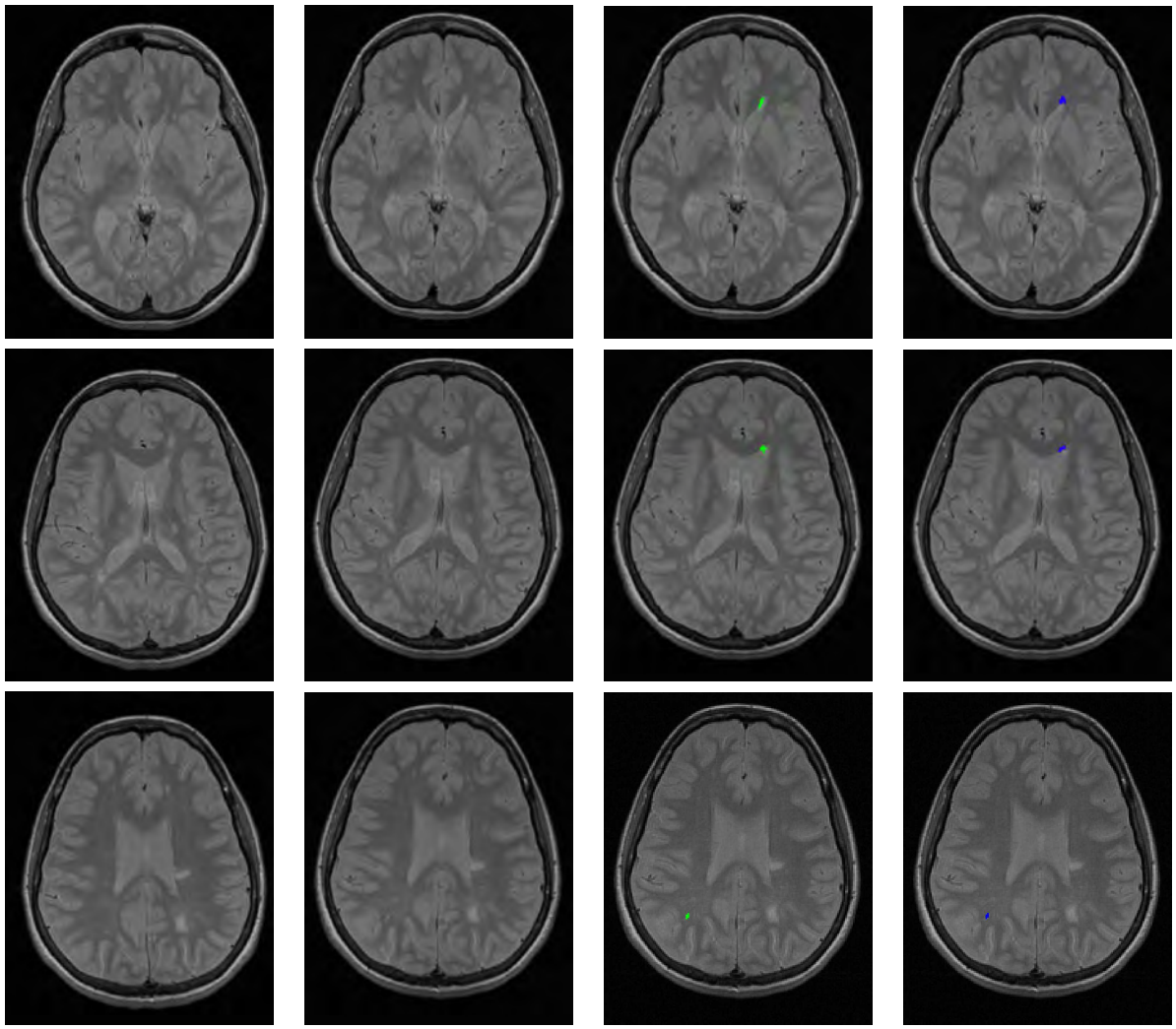
The aim of this appendix is to show the detailed results obtained by our approach for different lesion sizes and for all the studied patients (12M and 48M). Although for simplicity we show here only the results obtained from the unsupervised pipeline, apart from the number of false positive and true positive detections, the segmentation of the detected lesion are the same in both unsupervised and supervised pipelines due to the use of the same automated thresholding process.

A.1 Detailed performance per patient and lesion size

Tables A.1 to A.20 summarize the quantitative results obtained for all patients. On the other hand, Figures A.1 to A.20 show some visual examples of automated detections of new lesions for each patient. Note that the false positives and false negatives in the sample images may not necessarily indicate a false positive or negative detection with respect to the lesion detection accuracy since a lesion can appear in more than one slice. Thus, part of the lesion in other slices could have been detected by the pipeline.

Table A.1: Patient1 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	3	1	0	0	2	0	0
TP	3	1	-	-	2	-	-
SENS	1.00	1.00	-	-	1.00	-	-
FP	3	2	1	0	0	0	0
FDR	0.50	0.67	-	-	0.00	-	-
DSCR	0.67	0.50	-	-	1.00	-	-

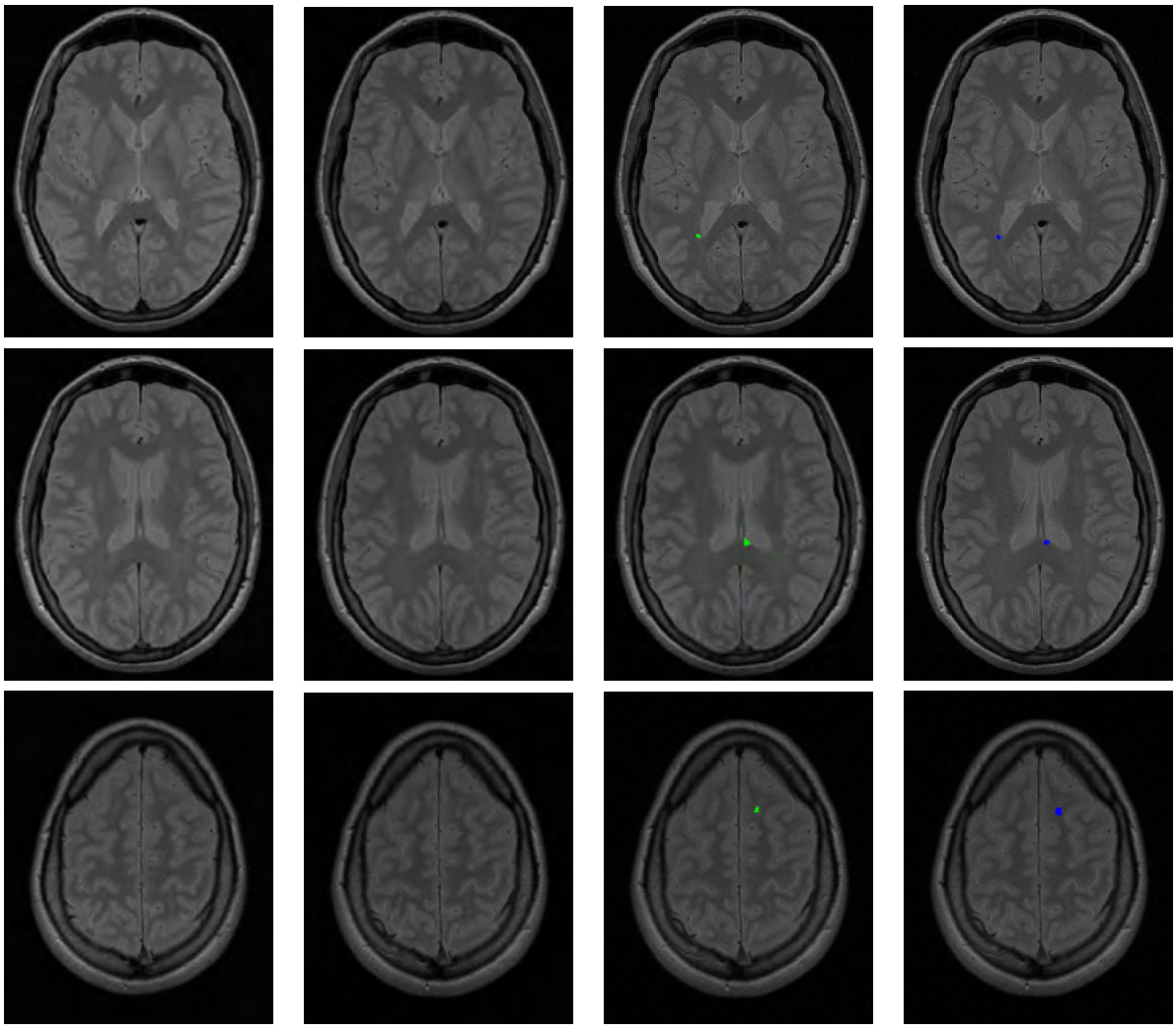


(a) (b) (c) (d)

Figure A.1: Patient1 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.2: Patient2 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	10	3	0	5	1	1	0
TP	10	3	-	5	1	1	-
SENS	1.00	1.00	-	1.00	1.00	1.00	-
FP	2	2	0	0	0	0	0
FDR	0.17	0.40	-	0.00	0.00	0.00	-
DSCR	0.91	0.75	-	1.00	1.00	1.00	-

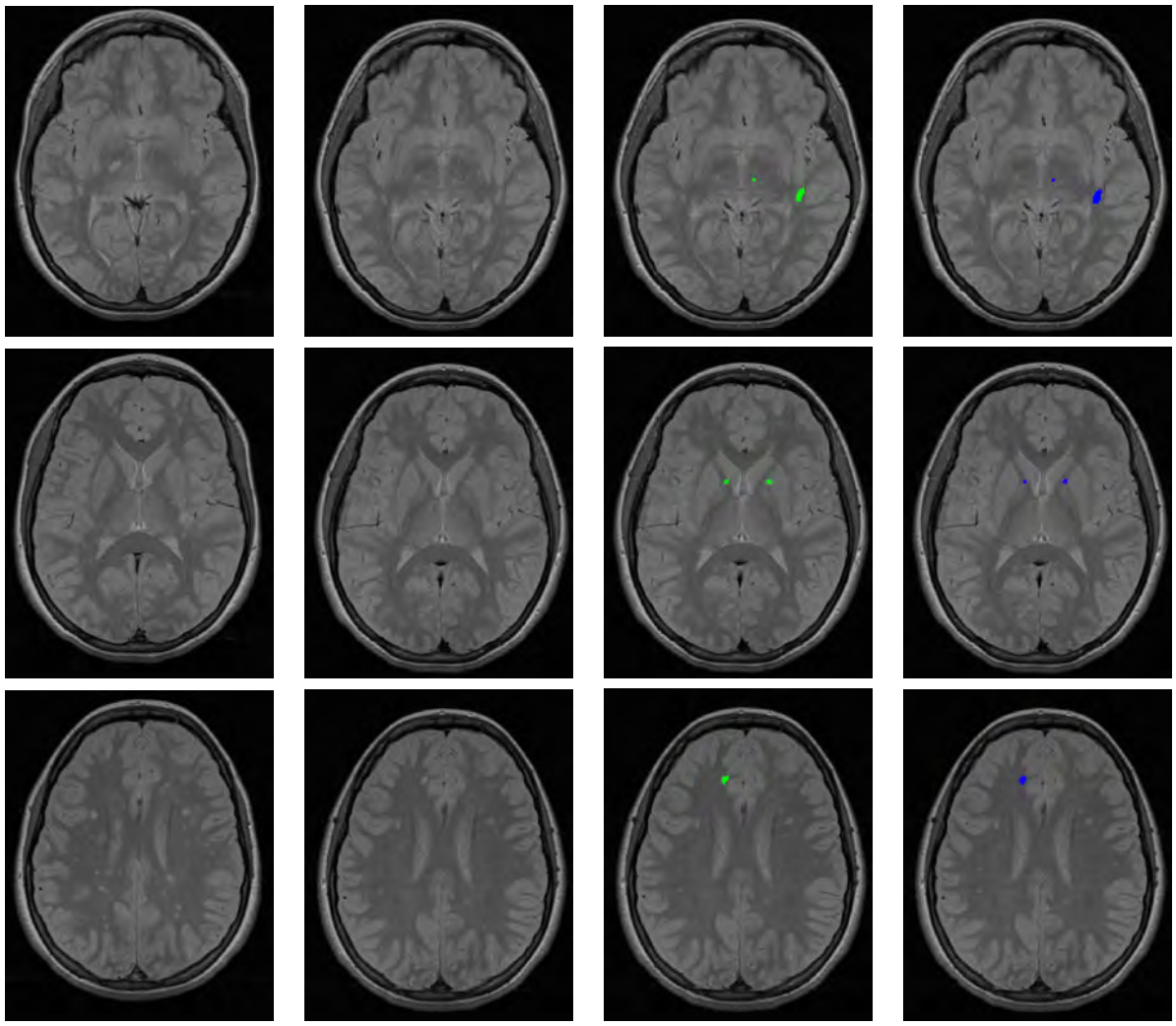


(a) (b) (c) (d)

Figure A.2: Patient2 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.3: Patient3 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	24	11	7	3	2	1	0
TP	19	7	7	2	2	1	-
SENS	0.79	0.64	1.00	0.67	1.00	1.00	-
FP	0	0	0	0	0	0	0
FDR	0.00	0.00	0.00	0.00	0.00	0.00	-
DSCR	0.88	0.78	1.00	0.80	1.00	1.00	-

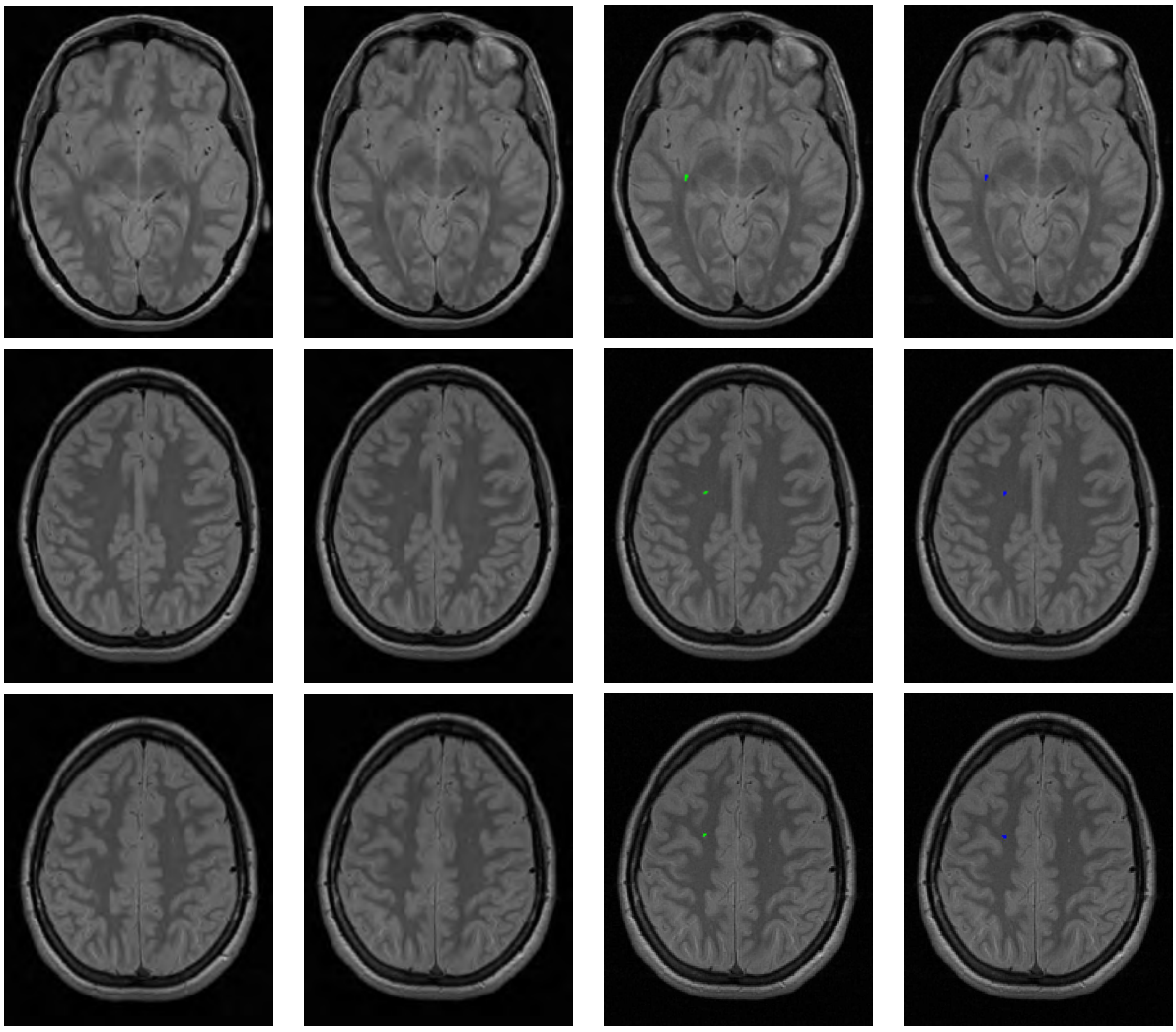


(a) (b) (c) (d)

Figure A.3: Patient3 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.4: Patient4 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	3	0	1	2	0	0	0
TP	3	-	1	2	-	-	-
SENS	1.00	-	1.00	1.00	-	-	-
FP	3	1	1	0	1	0	0
FDR	0.50	-	0.50	0.00	-	-	-
DSCR	0.67	-	0.67	1.00	-	-	-

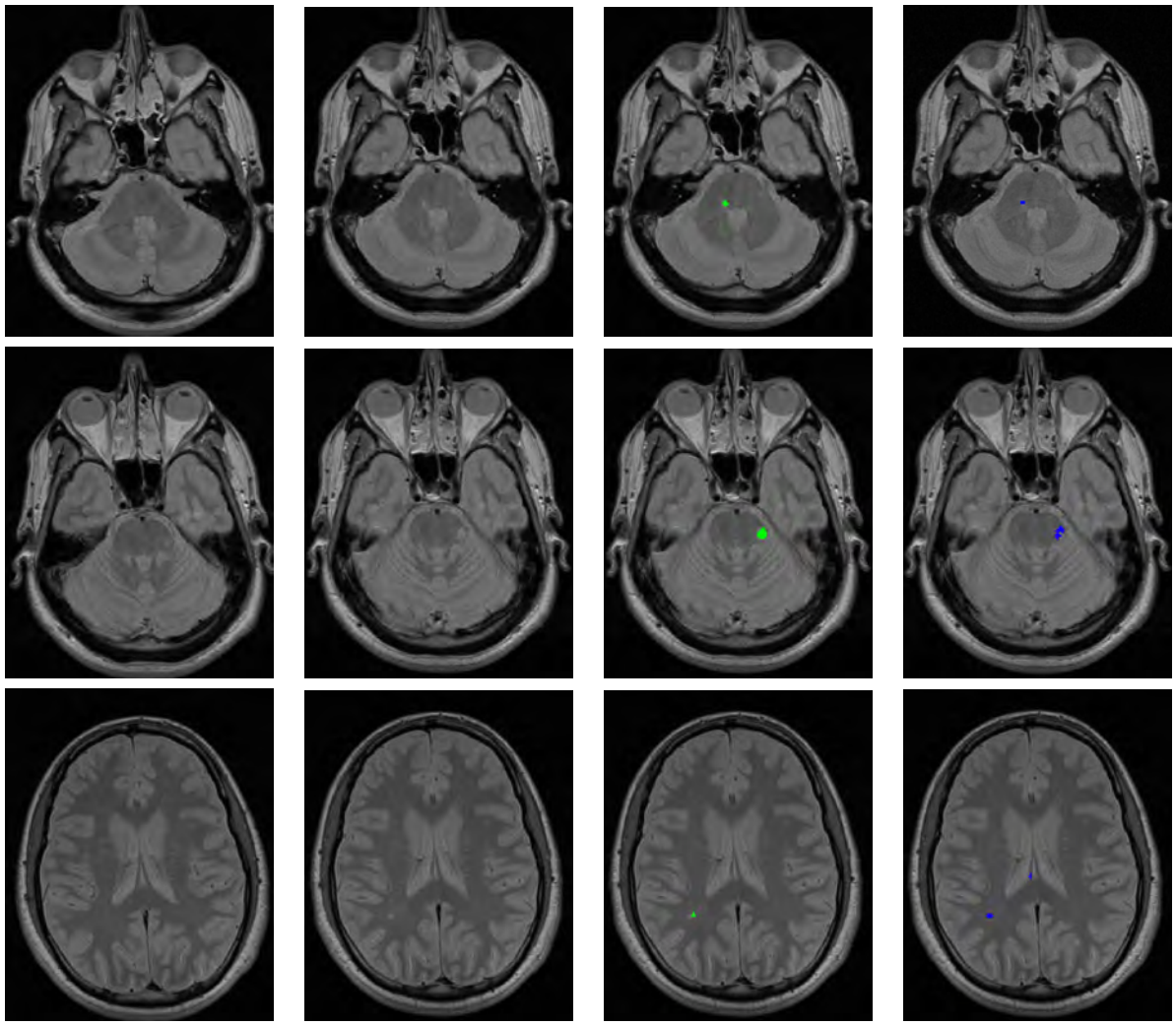


(a) (b) (c) (d)

Figure A.4: Patient4 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.5: Patient5 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	7	0	0	3	4	0	0
TP	7	-	-	3	4	-	-
SENS	1.00	-	-	1.00	1.00	-	-
FP	0	0	0	0	0	0	0
FDR	0.00	-	-	0.00	0.00	-	-
DSCR	1.00	-	-	1.00	1.00	-	-

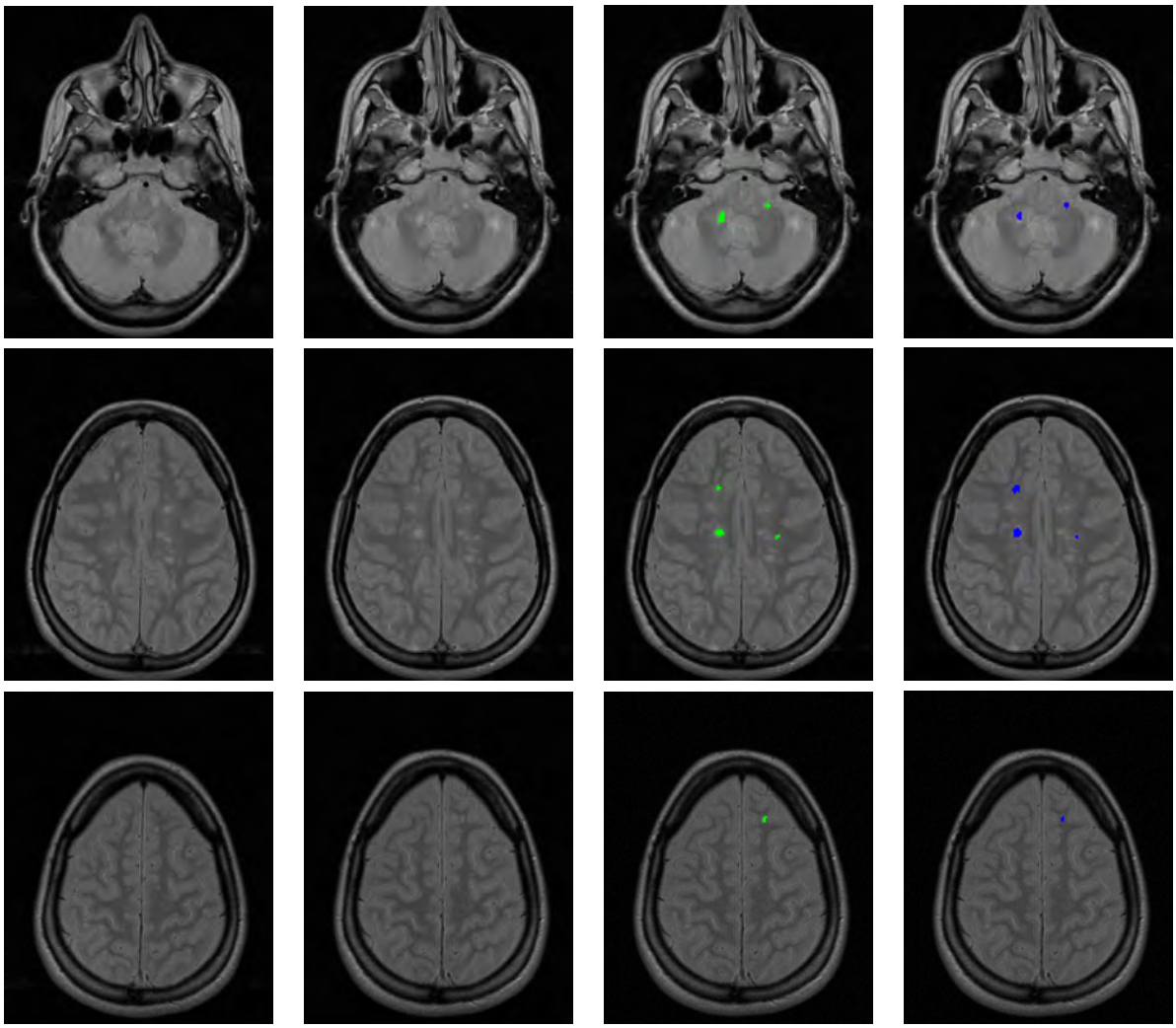


(a) (b) (c) (d)

Figure A.5: Patient5 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.6: Patient6 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	23	4	3	6	7	2	1
TP	21	4	3	5	6	2	1
SENS	0.91	1.00	1.00	0.83	0.86	1.00	1.00
FP	6	6	0	0	0	0	0
FDR	0.22	0.60	0.00	0.00	0.00	0.00	0.00
DSCR	0.84	0.57	1.00	0.91	0.92	1.00	1.00

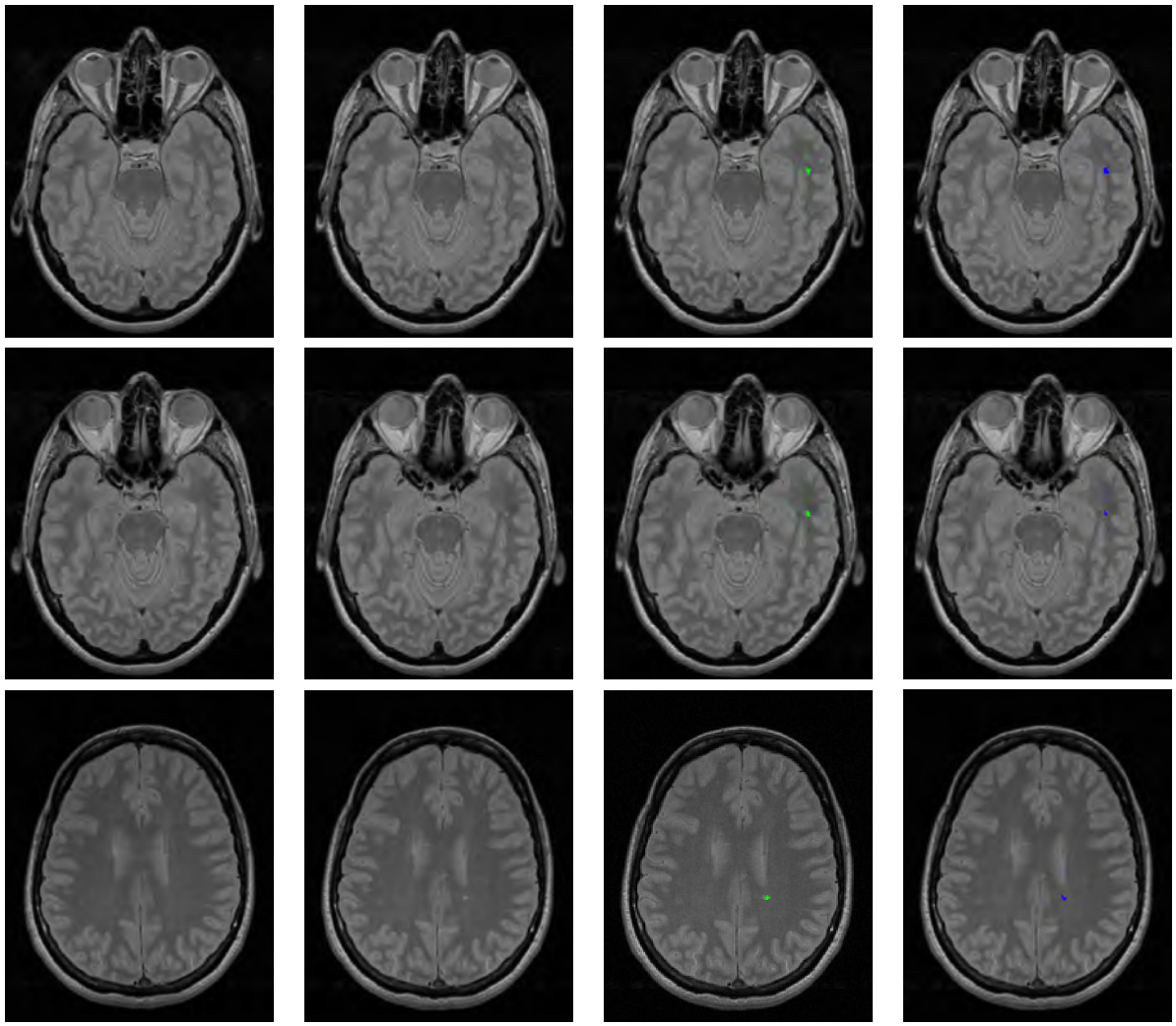


(a) (b) (c) (d)

Figure A.6: Patient6 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.7: Patient7 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	3	1	1	1	0	0	0
TP	3	1	1	1	-	-	-
SENS	1.00	1.00	1.00	1.00	-	-	-
FP	0	0	0	0	0	0	0
FDR	0.00	0.00	0.00	0.00	-	-	-
DSCR	1.00	1.00	1.00	1.00	-	-	-

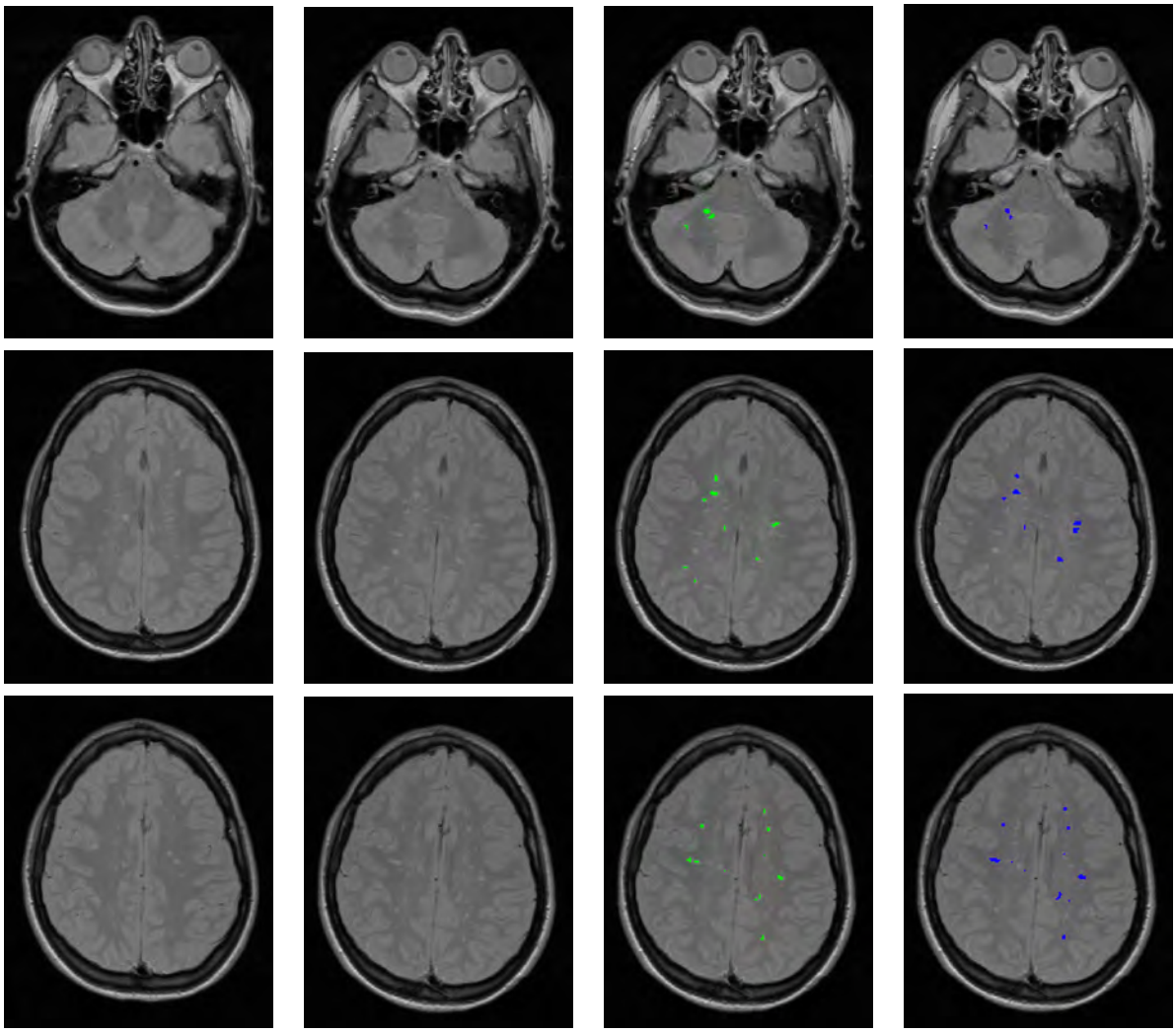


(a) (b) (c) (d)

Figure A.7: Patient7 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.8: Patient8 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	51	19	18	9	4	1	0
TP	41	13	15	8	4	1	-
SENS	0.80	0.68	0.83	0.89	1.00	1.00	-
FP	1	0	1	0	0	0	0
FDR	0.02	0.00	0.06	0.00	0.00	0.00	-
DSCR	0.88	0.81	0.88	0.94	1.00	1.00	-

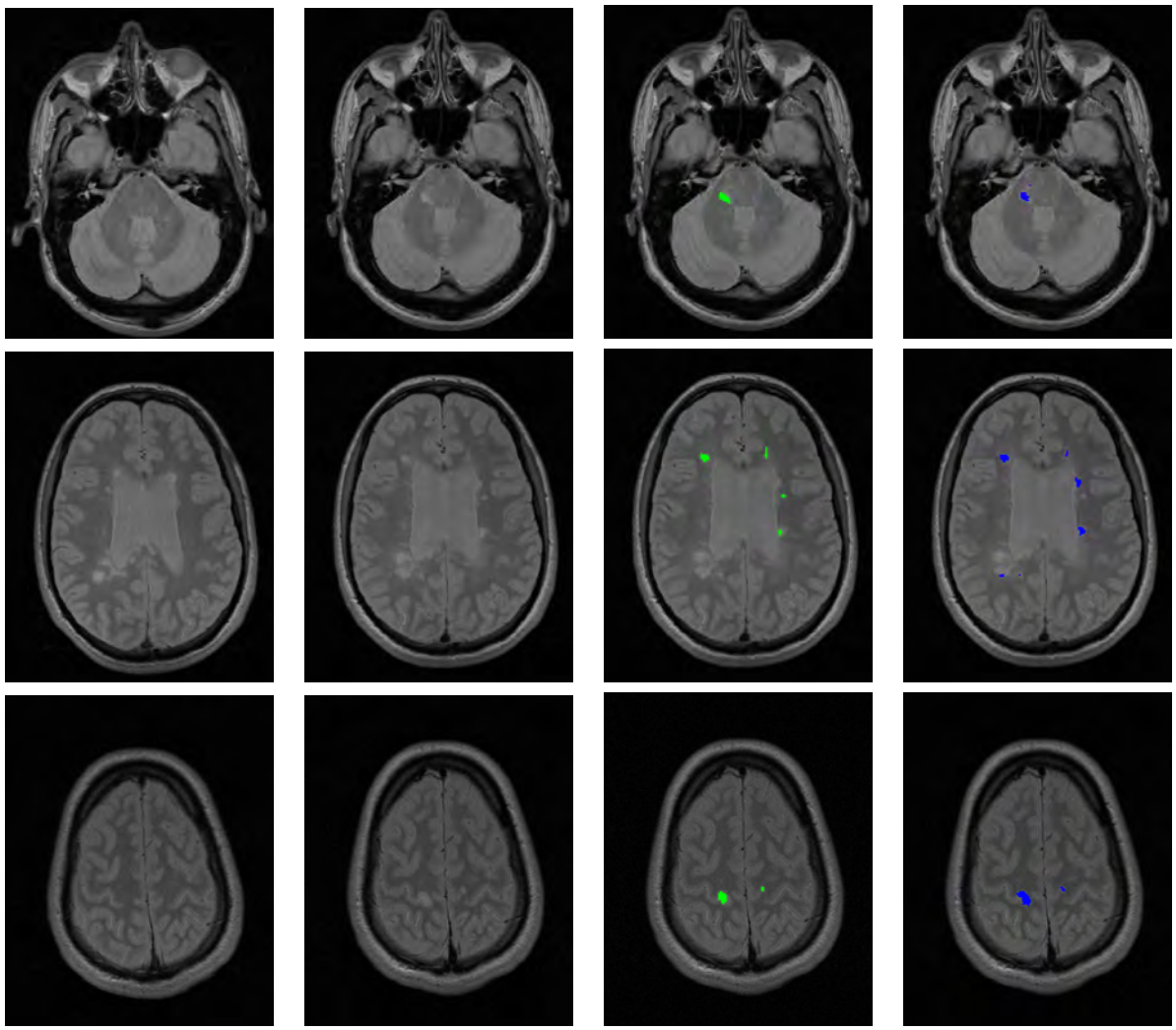


(a) (b) (c) (d)

Figure A.8: Patient8 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.9: Patient9 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	36	13	2	7	9	4	1
TP	24	4	2	5	8	4	1
SENS	0.67	0.31	1.00	0.71	0.89	1.00	1.00
FP	5	3	1	1	0	0	0
FDR	0.17	0.43	0.33	0.17	0.00	0.00	0.00
DSCR	0.74	0.40	0.80	0.77	0.94	1.00	1.00

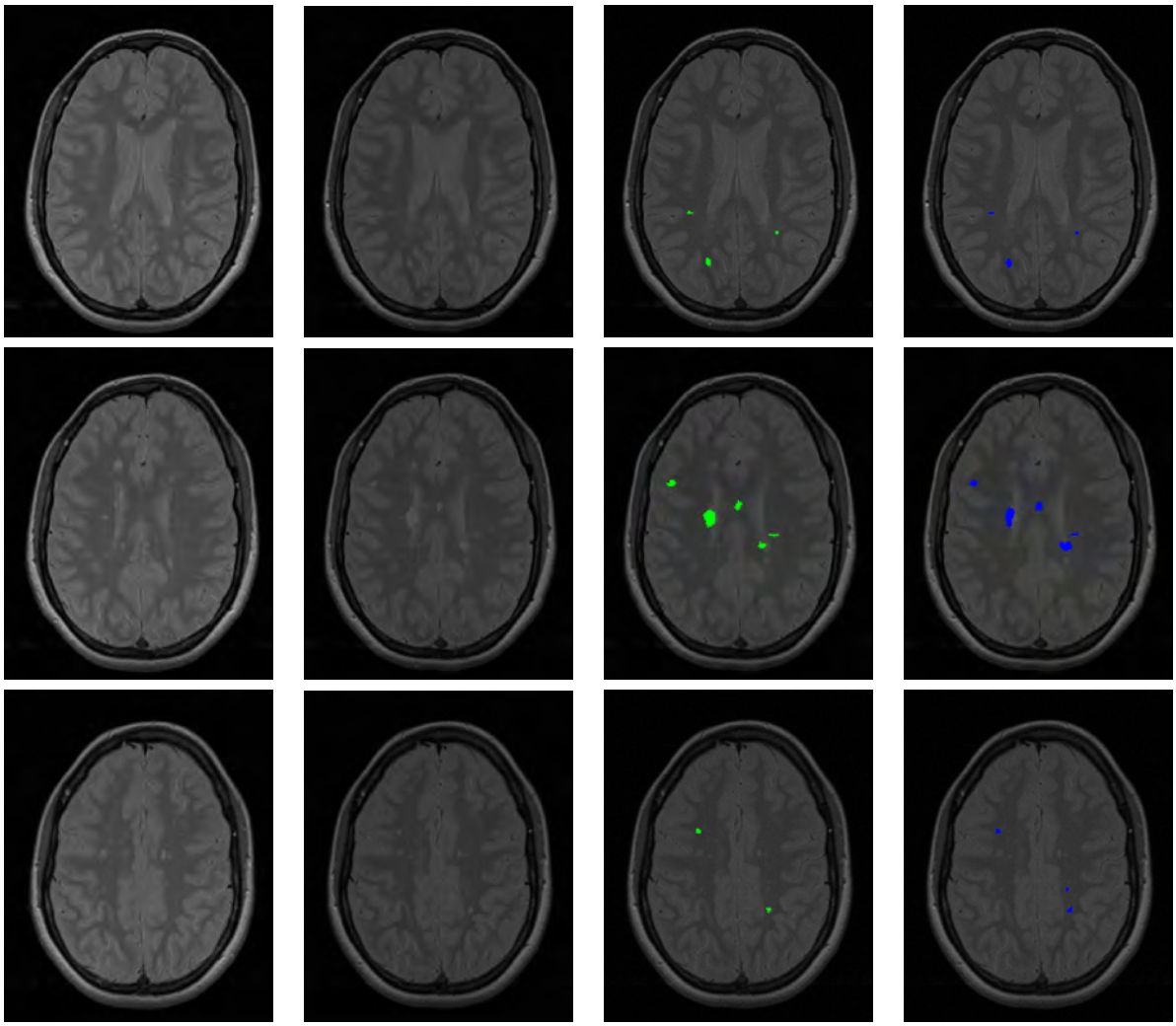


(a) (b) (c) (d)

Figure A.9: Patient9 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.10: Patient10 (12M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	17	7	4	3	1	1	1
TP	16	6	4	3	1	1	1
SENS	0.94	0.86	1.00	1.00	1.00	1.00	1.00
FP	3	1	2	0	0	0	0
FDR	0.16	0.14	0.33	0.00	0.00	0.00	0.00
DSCR	0.89	0.86	0.80	1.00	1.00	1.00	1.00

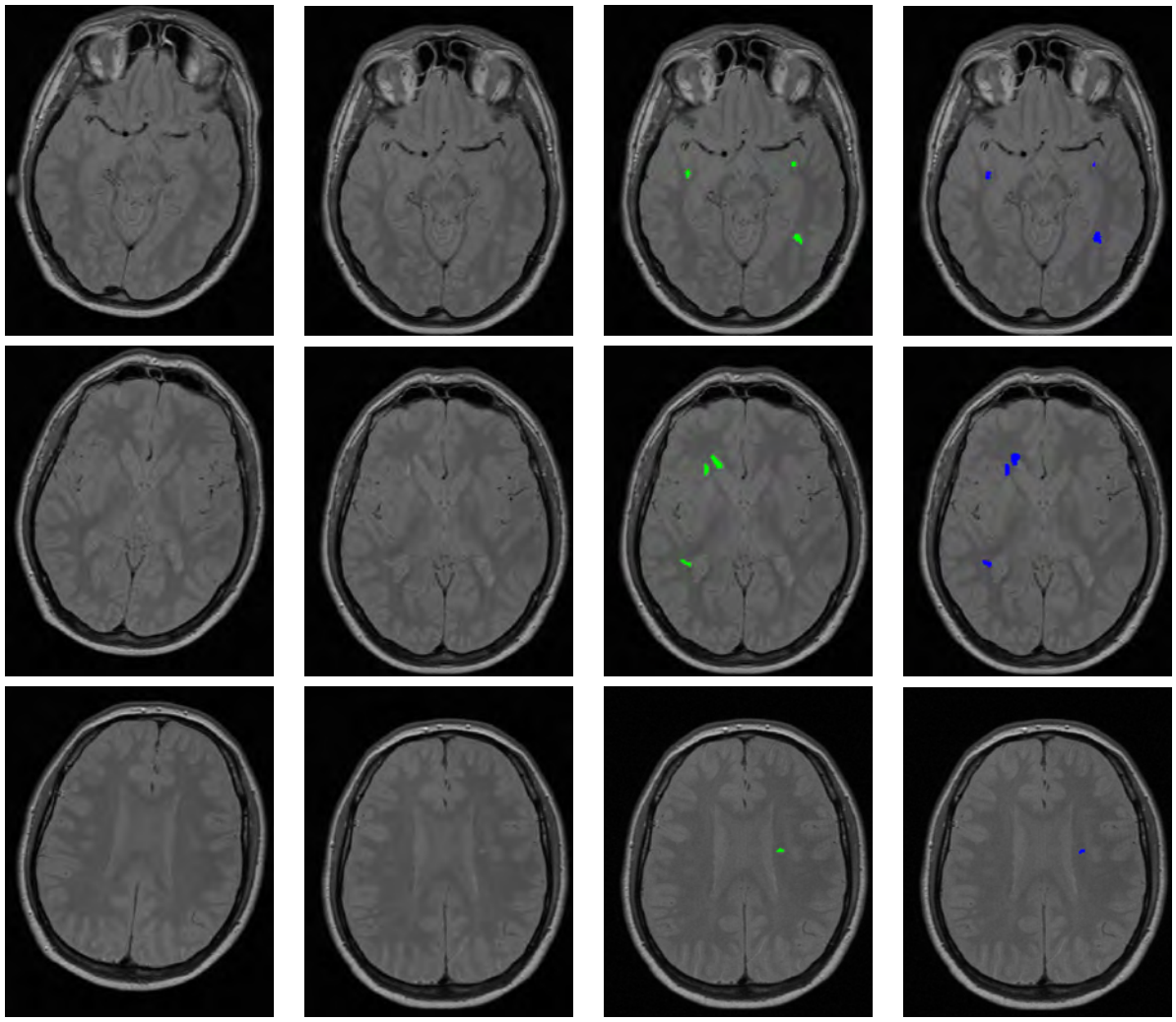


(a) (b) (c) (d)

Figure A.10: Patient10 (12M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.11: Patient11 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	17	4	2	7	2	1	1
TP	13	2	2	6	1	1	1
SENS	0.76	0.50	1.00	0.86	0.50	1.00	1.00
FP	0	0	0	0	0	0	0
FDR	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DSCR	0.87	0.67	1.00	0.92	0.67	1.00	1.00

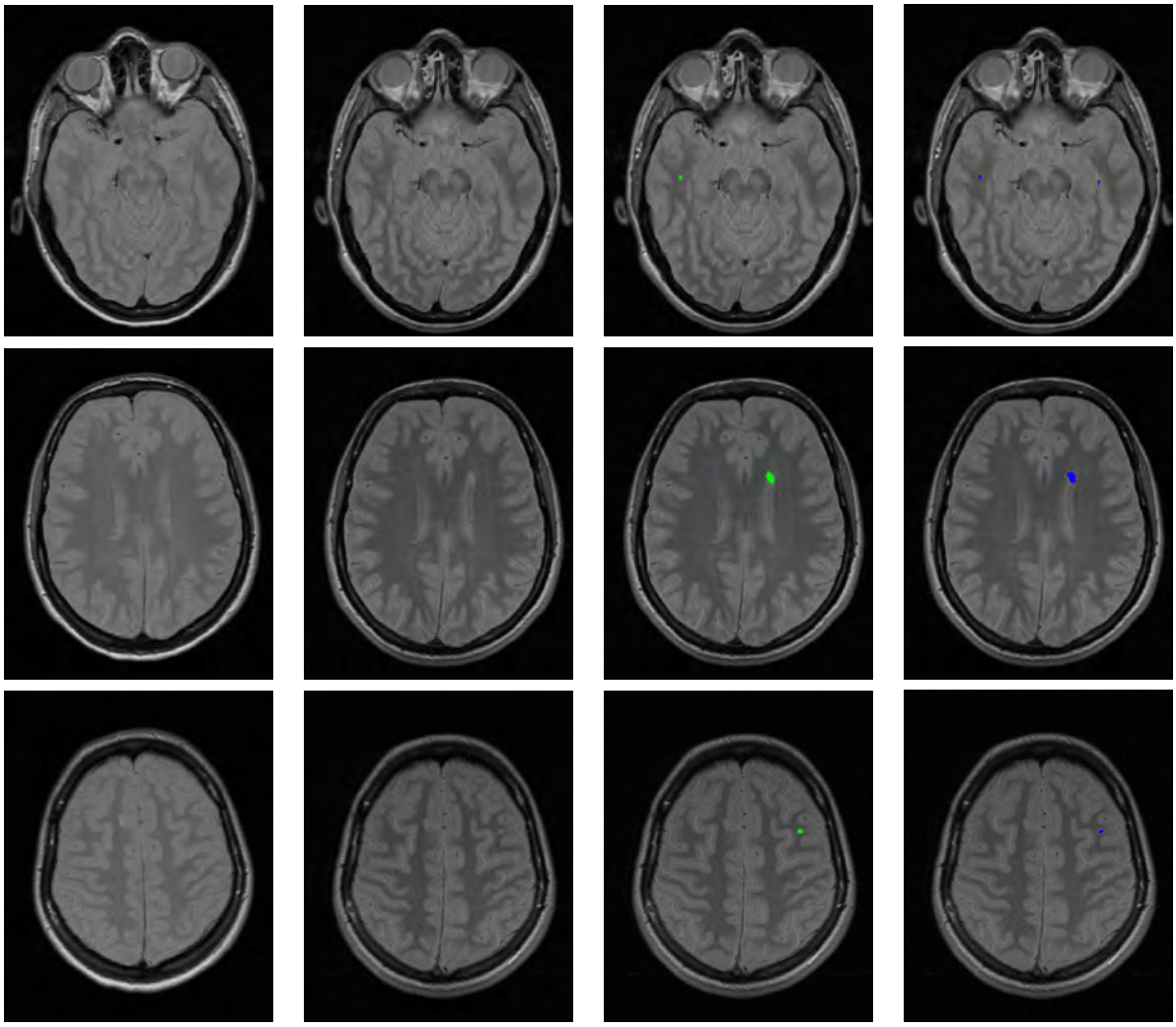


(a) (b) (c) (d)

Figure A.11: Patient11 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.12: Patient12 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	6	2	2	0	1	1	0
TP	3	1	1	-	1	0	-
SENS	0.50	0.50	0.50	-	1.00	0.00	-
FP	6	5	0	1	0	0	0
FDR	0.67	0.83	0.00	-	0.00	-	-
DSCR	0.40	0.25	0.67	-	1.00	0.00	-

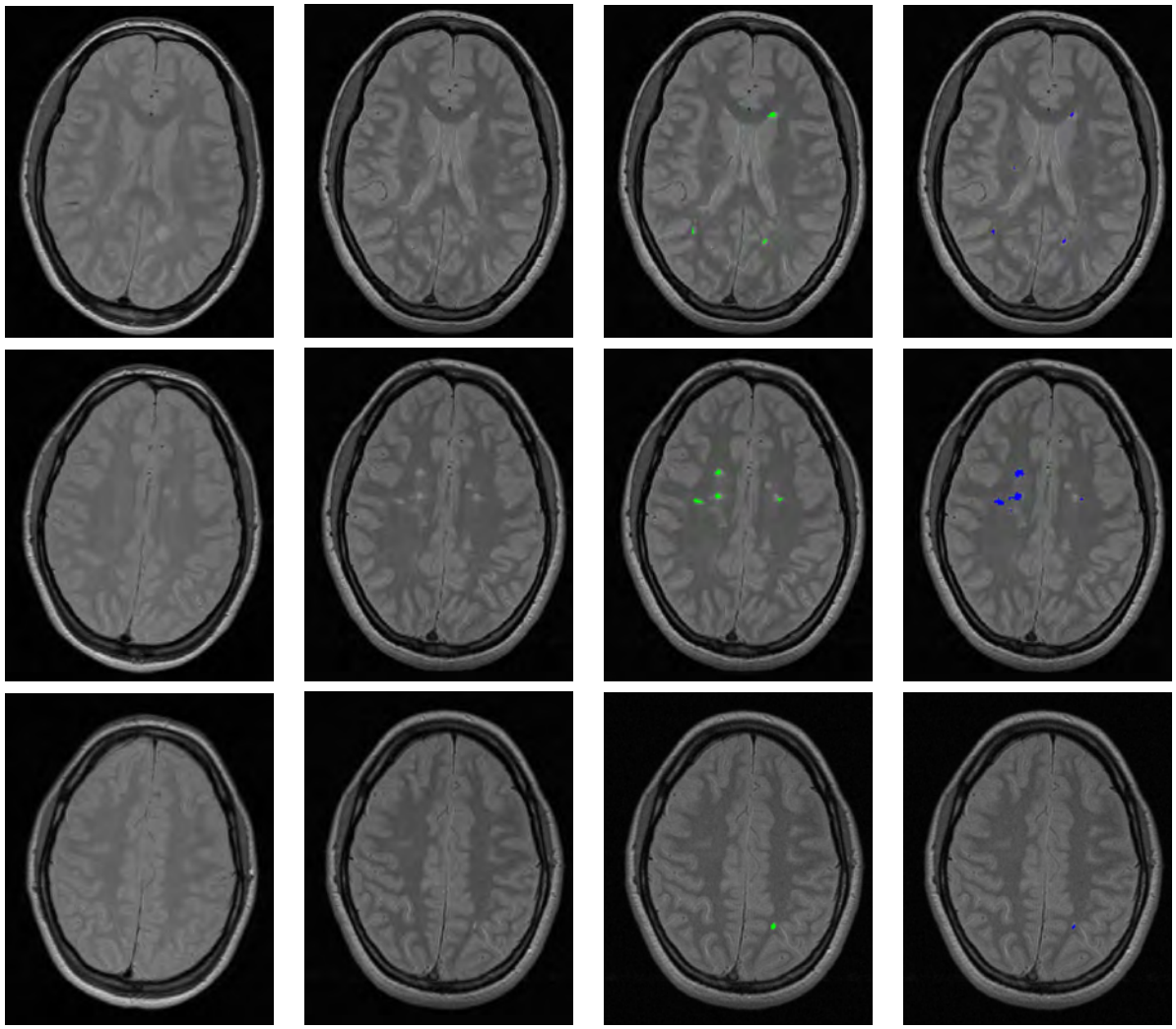


(a) (b) (c) (d)

Figure A.12: Patient12 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.13: Patient13 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	16	1	3	6	6	0	0
TP	16	1	3	6	6	-	-
SENS	1.00	1.00	1.00	1.00	1.00	-	-
FP	2	0	1	0	0	1	0
FDR	0.11	0.00	0.25	0.00	0.00	-	-
DSCR	0.94	1.00	0.86	1.00	1.00	-	-

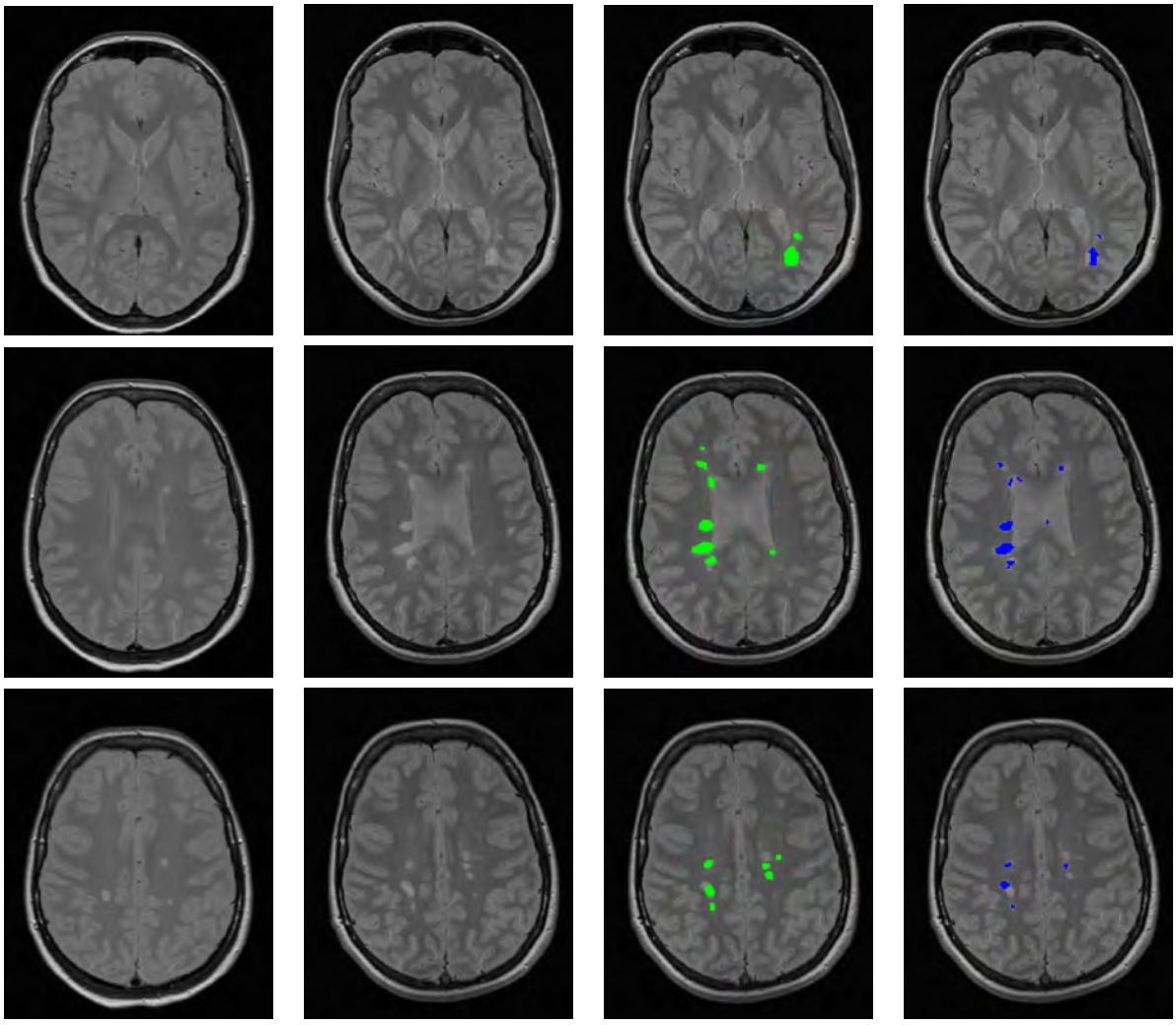


(a) (b) (c) (d)

Figure A.13: Patient13 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.14: Patient14 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	38	2	5	5	13	6	7
TP	22	1	1	2	6	5	7
SENS	0.58	0.50	0.20	0.40	0.46	0.83	1.00
FP	2	2	0	0	0	0	0
FDR	0.08	0.67	0.00	0.00	0.00	0.00	0.00
DSCR	0.71	0.40	0.33	0.57	0.63	0.91	1.00

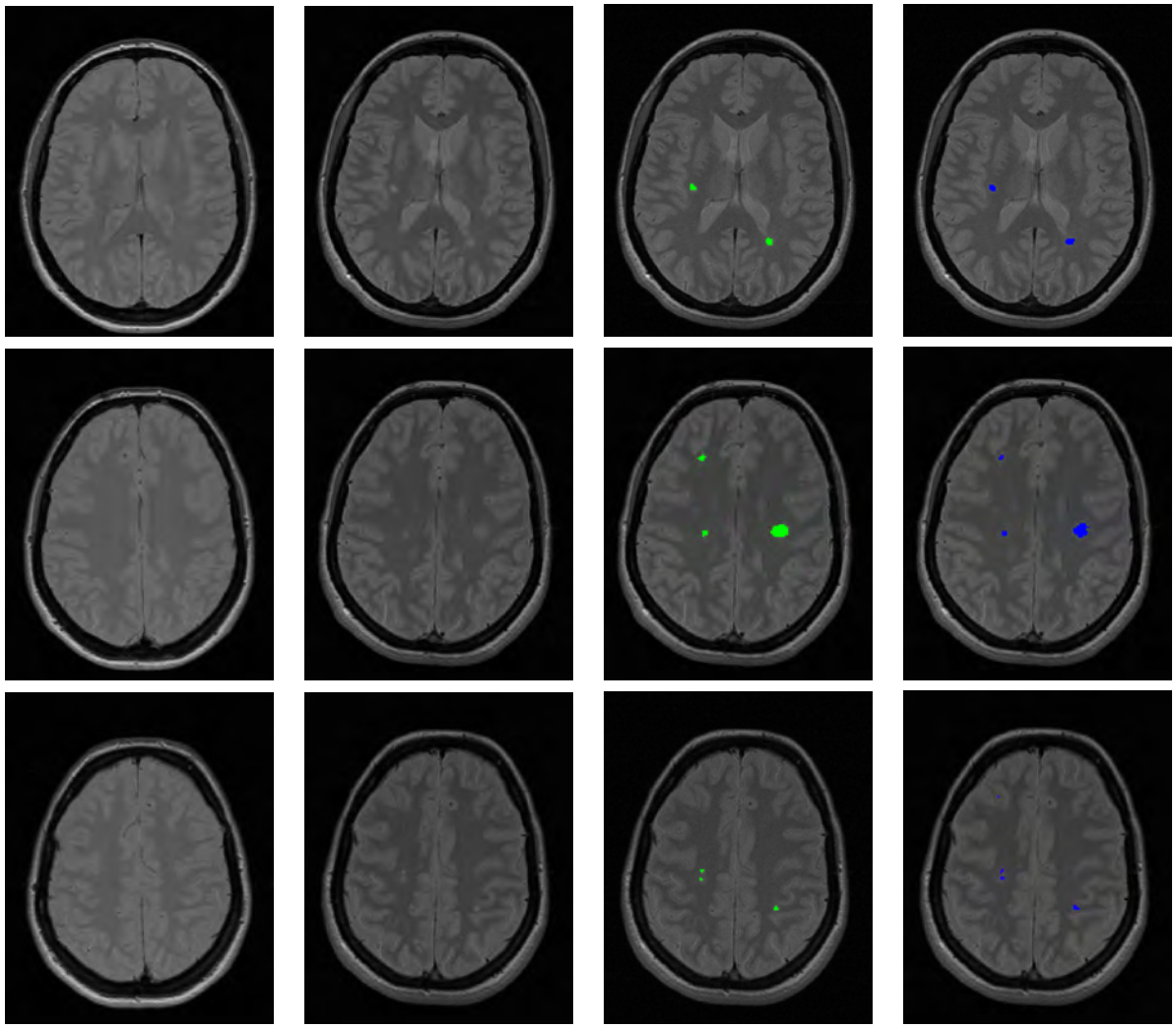


(a) (b) (c) (d)

Figure A.14: Patient14 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.15: Patient15 (48M): Performance of the unsupervised pipeline by lesion size.

	Overall	**Size**					
		3-6	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	13	4	0	2	5	1	1
TP	10	1	-	2	5	1	1
SENS	0.77	0.25	-	1.00	1.00	1.00	1.00
FP	1	1	0	0	0	0	0
FDR	0.09	0.50	-	0.00	0.00	0.00	0.00
DSCR	0.83	0.33	-	1.00	1.00	1.00	1.00

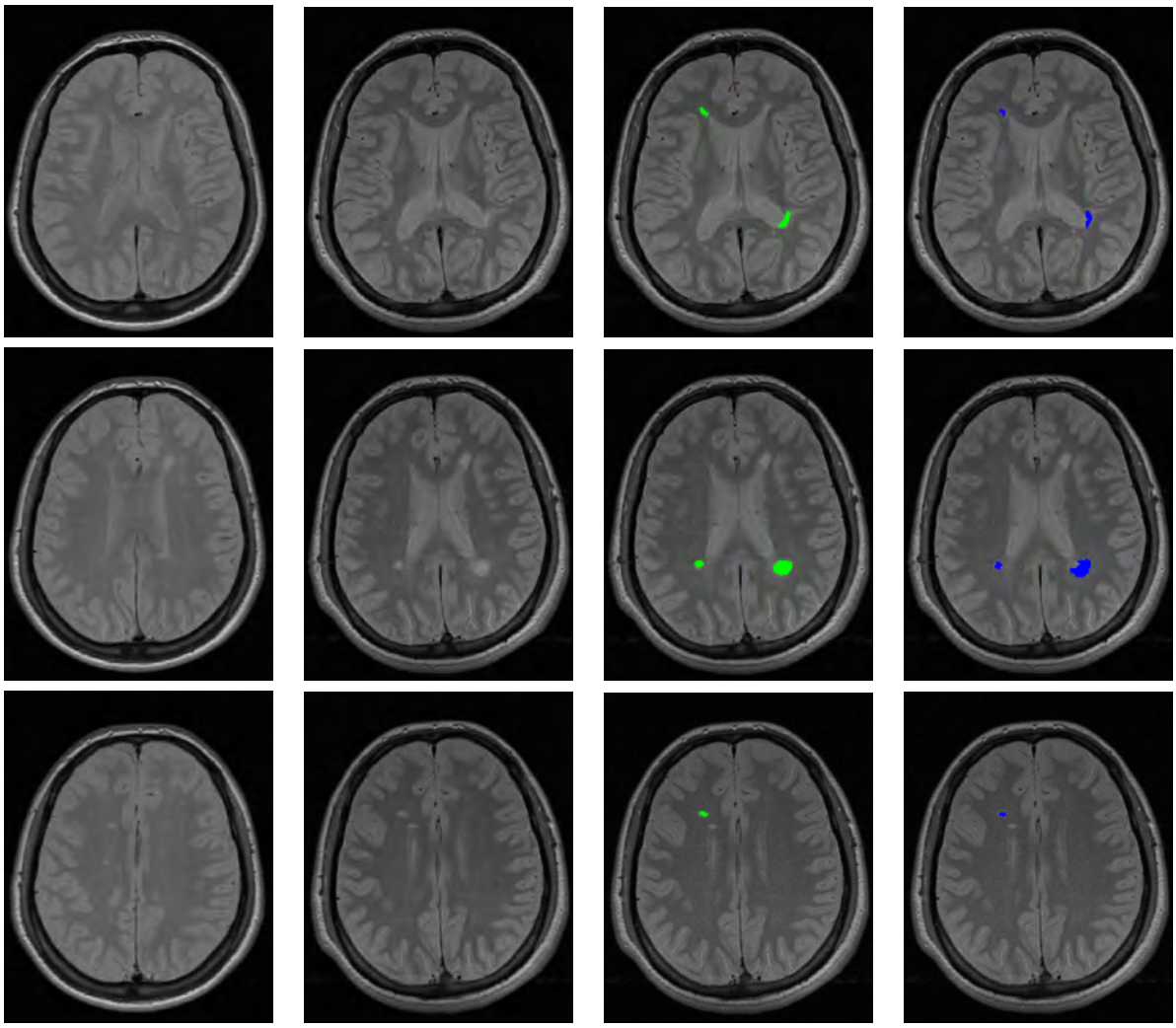


(a) (b) (c) (d)

Figure A.15: Patient15 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.16: Patient16 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	9	1	1	3	3	0	1
TP	9	1	1	3	3	0	1
SENS	1.00	1.00	1.00	1.00	1.00	-	1.00
FP	1	1	0	0	0	0	0
FDR	0.10	0.50	0.00	0.00	0.00	-	0.00
DSCR	0.95	0.67	1.00	1.00	1.00	-	1.00

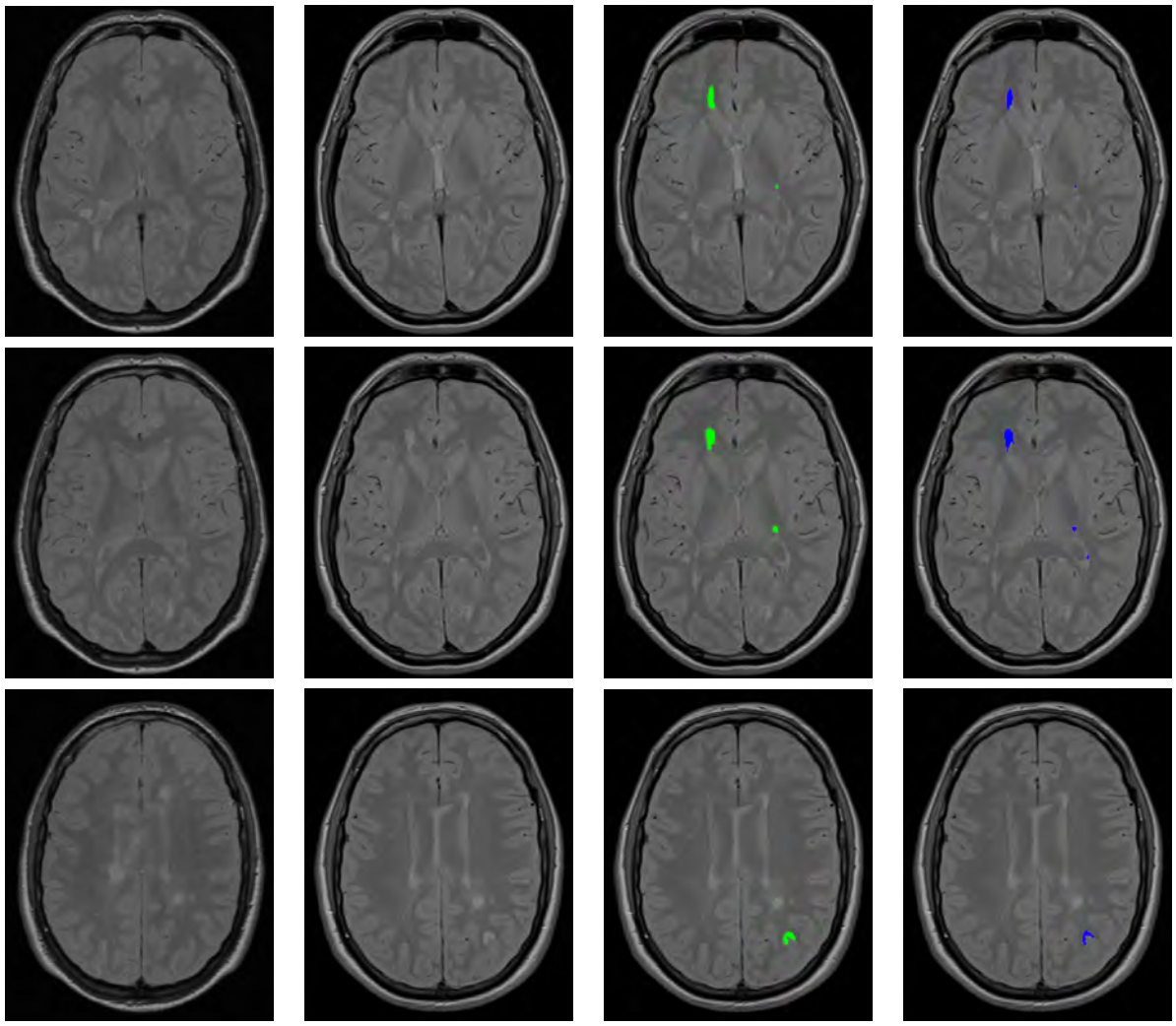


(a) (b) (c) (d)

Figure A.16: Patient16 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.17: Patient17 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	15	0	0	4	6	2	3
TP	12	-	-	3	5	1	3
SENS	0.80	-	-	0.75	0.83	0.50	1.00
FP	2	1	1	0	0	0	0
FDR	0.14	-	-	0.00	0.00	0.00	0.00
DSCR	0.83	-	-	0.86	0.91	0.67	1.00

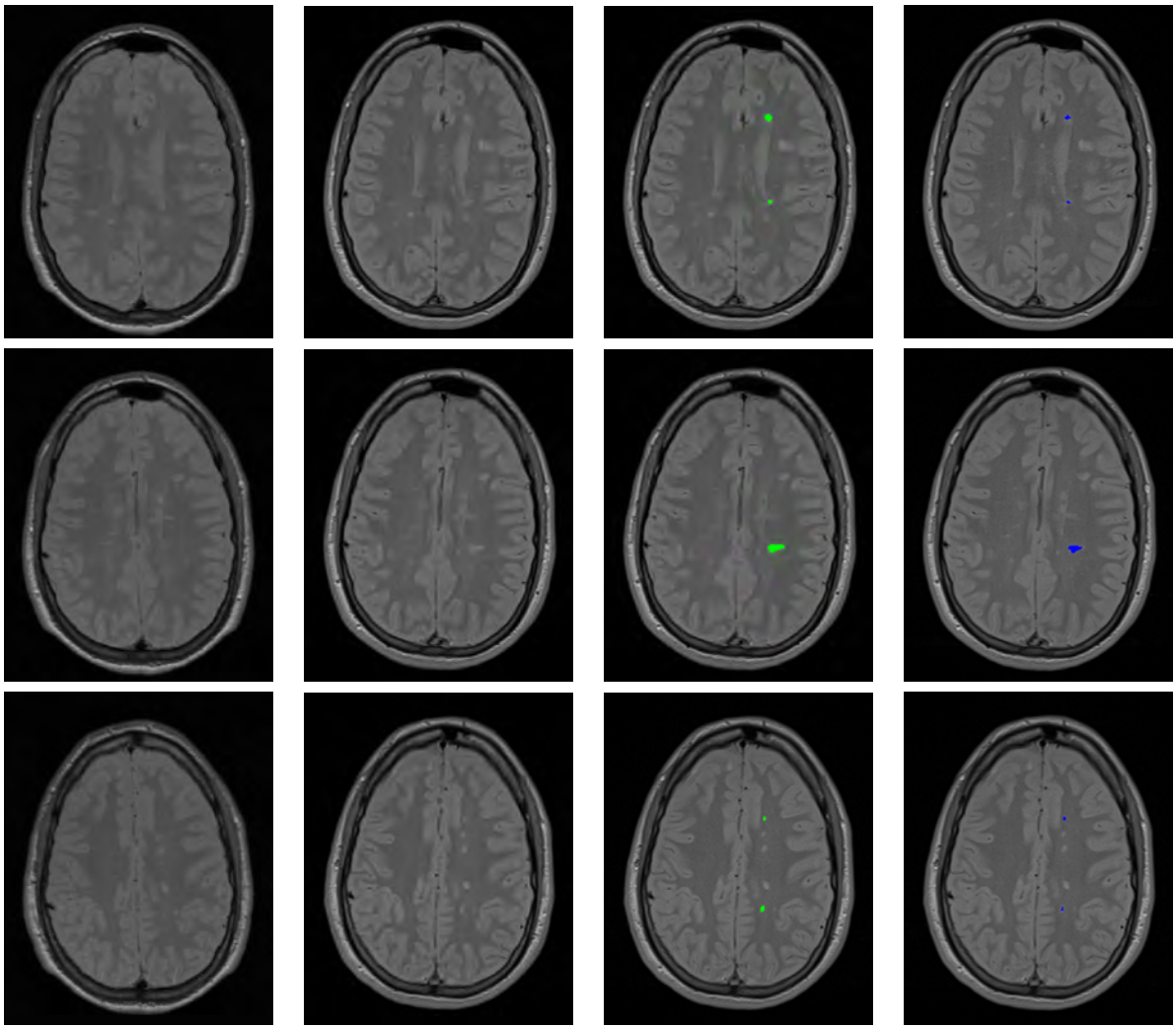


(a) (b) (c) (d)

Figure A.17: Patient17 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.18: Patient18 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	8	3	1	1	1	1	1
TP	7	3	1	0	1	1	1
SENS	0.88	1.00	1.00	0.00	1.00	1.00	1.00
FP	6	3	1	1	0	1	0
FDR	0.46	0.50	0.50	1.00	0.00	0.50	0.00
DSCR	0.67	0.67	0.67	0.00	1.00	0.67	1.00

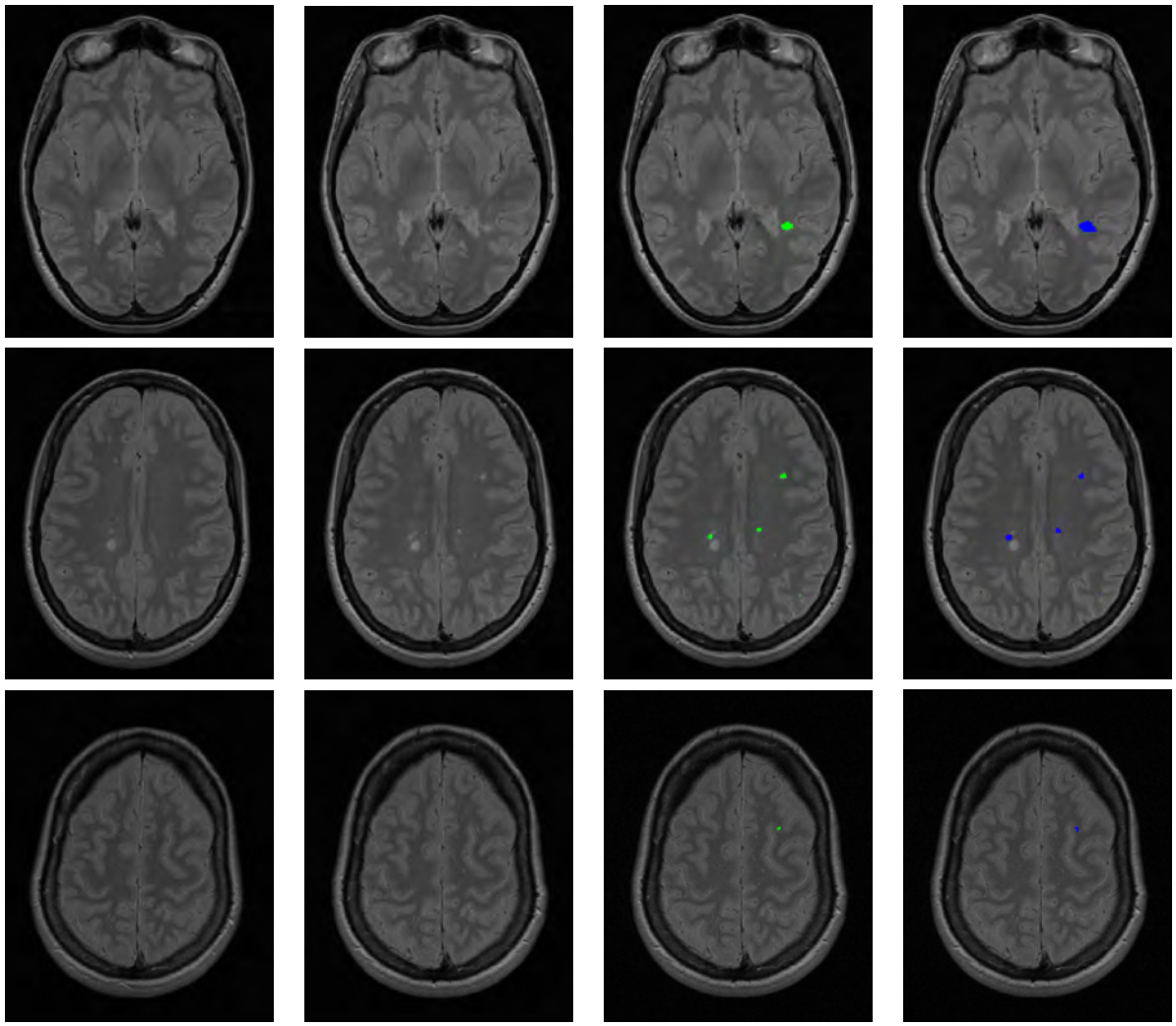


(a) (b) (c) (d)

Figure A.18: Patient18 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.19: Patient19 (48M): Performance of the unsupervised pipeline by lesion size.

	Size						
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	13	4	1	2	4	1	1
TP	13	4	1	2	4	1	1
SENS	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FP	0	0	0	0	0	0	0
FDR	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DSCR	1.00	1.00	1.00	1.00	1.00	1.00	1.00

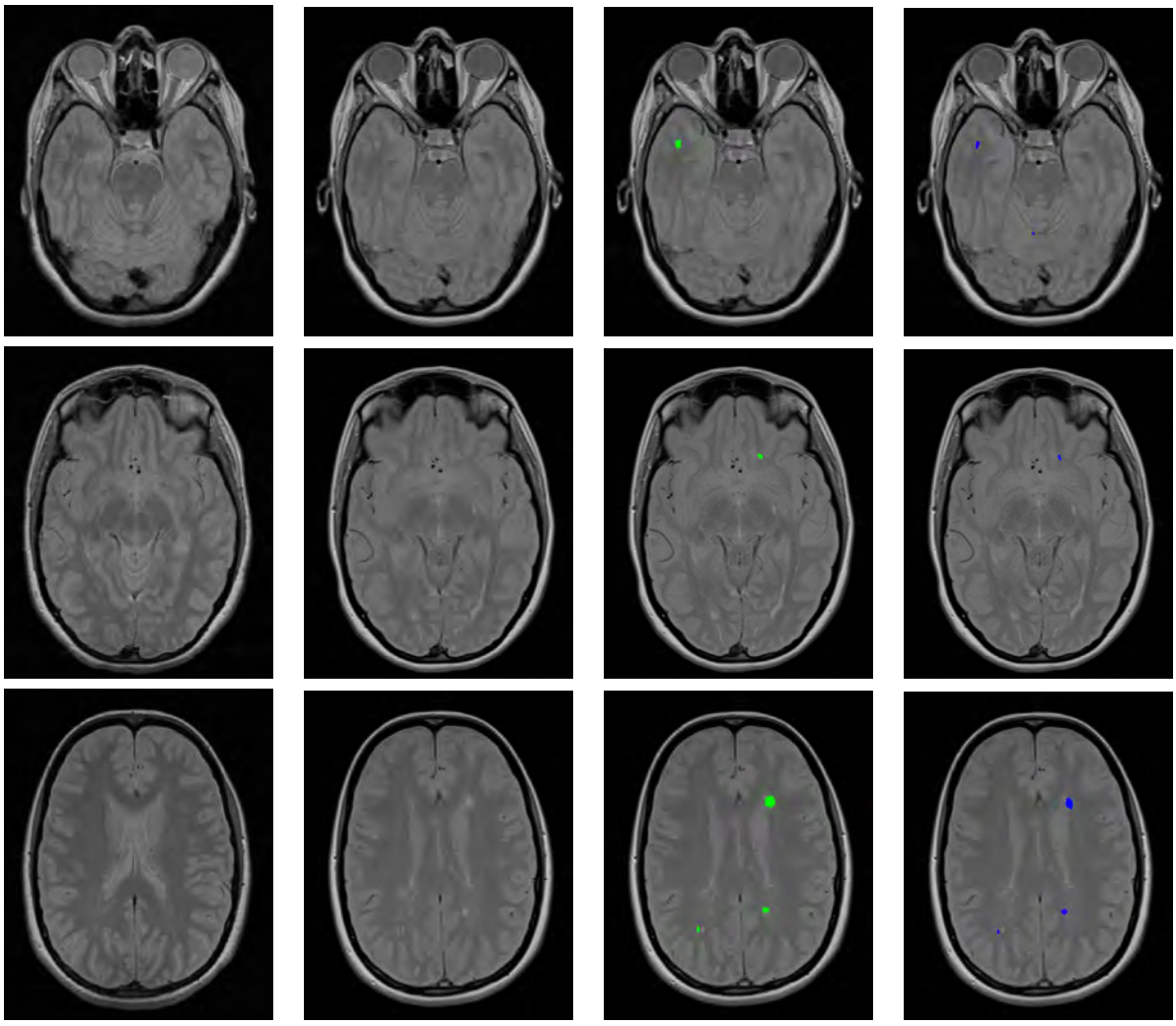


(a) (b) (c) (d)

Figure A.19: Patient19 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Table A.20: Patient20 (48M): Performance of the unsupervised pipeline by lesion size.

		Size					
	Overall	**3-6**	**7-10**	**11-20**	**21-50**	**51-100**	**101+**
N	17	4	0	7	3	3	0
TP	12	2	-	5	2	3	-
SENS	0.71	0.50	-	0.71	0.67	1.00	-
FP	5	1	1	1	2	0	0
FDR	0.29	0.33	-	0.17	0.50	0.00	-
DSCR	0.71	0.57	-	0.77	0.57	1.00	-



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

Figure A.20: Patient20 (48M): Visual examples of automated detections of new lesions. (a) baseline image, (b) follow-up image, (c) manual segmentation, (d) automated detection.

Bibliography

[1] MS lesion segmantation challange 2008. `http://www.ia.unc.edu/MSseg`. Accessed: 15/08/2014.

[2] The McConnell Brain Imaging Center (2010) BrainWeb: Simulated Brain Database. `http://mouldy.bic.mni.mcgill.ca/brainweb/`. Accessed: 29/12/2010.

[3] P. Anbeek, K. Vincken, and M. Viergever. Automated MS-lesion segmentation by K-nearest neighbor classification. 2008.

[4] P. Anbeek, K.L. Vincken, and M.J. Van Osch. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Medical Image Analysis*, 8:205–215, 2004.

[5] S.B. Antel, D.L. Collins, N. Bernasconi, F. Andermann, R. Singhal, R.E. Kearney, D. Arnold, and A. Bernasconi. Automated detection of focal cortical dysplasia lesions using computational models of their MRI characteristics and texture analysis. *IEEE Trans Med Imaging*, 19(4):1748–1759, 2003.

[6] H. Arimura, T. Magome, Y. Yamashita, and D. Yamamoto. Computer-aided diagnosis systems for brain diseases in magnetic resonance images. *Algorithms*, 2(3):925–952, 2009.

[7] E.A. Ashton, C. Takahashi, M.J. Berg, A. Goodman, S. Totterman, and S. Ekholm. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. *Journal Of Magnetic Resonance Imaging*, 17:300–308, 2003.

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

[9] R. Bakshi, A.J. Thompson, M.A. Rocca, D. Pelletier, V. Dousset, F. Barkhof, M. Inglese, C.R. Guttmann, M.A. Horsfield, and M. Filippi. MRI in multiple sclerosis: current status and future prospects. *The Lancet Neurology*, 7(7):615–625, 2008.

[10] F. Barkhof, M. Filippi, D.H. Miller, P. Scheltens, A. Campi, C.H. Polman, G. Comi, H.J. Adèr, N. Losseff, and J. Valk. Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain*, 120(11):2059–2069, 1997.

[11] M. Battaglini, F. Rossi, R.A. Grove, M.L. Stromillo, B. Whitcher, P.M. Matthews, and N. De Stefano. Automated identification of brain new lesions in multiple sclerosis using subtraction images. *Journal of Magnetic Resonance Imaging*, 2013. Article in Press.

[12] A. Ben-Hur and J. Weston. A user's guide to support vector machines. In Oliviero Carugo and Frank Eisenhaber, editors, *Data Mining Techniques for the Life Sciences*, volume 609, pages 223–239. Humana Press, 2010.

[13] M.A. Bernstein, K.E. King, and X.J.F. Zhou. *Handbook of MRI Pulse Sequences*. Academic Press Elsevier, 2004.

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

[15] Evert J. Blink. *MRI : Basic Physics*. Ebook, 2004.

[16] K. Boesen, K. Rehm, K. Schaper, S. Stoltzner, R. Woods, E. Lüders, and D. Rottenberg. Quantitative comparison of four brain extraction algorithms. *NeuroImage*, 22(3):1255–1261, 2004.

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

[18] M. Brant-Zawadzki, G.D. Gillan, and Nitz W.R. MP RAGE: A three-dimensional, T1-weighted, gradient-echo sequence initial experience in the brain. *Radiology*, 182(3):769–775, 1992.

[19] W. Brück, A. Bitsch, H. Kolenda, Y. Brück, M. Stiefel, and Lassmann H. Inflammatory central nervous system demyelination: correlation of magnetic resonance imaging findings with lesion pathology. *Ann. Neurol.*, 42(5):783–793, 1997.

[20] M. Cabezas, M. BachCuadra, A. Oliver, X. Lladó, J. Freixenet, J.C. Vilanova, L. Valls, Ll. Ramió-Torrentà, E. Huerga, D. Pareto, and A. Rovira. A pipeline approach with spatial information for segmenting multiple sclerosis lesions on brain magnetic resonance imaging. In *Proc. Europ. Comm. Treatm. Res. Mult. Scl.*, page 381, 2011.

[21] 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.

[22] G. Calcagno, A. Staiano, G. Fortunato, V. Brescia-Morra, E. Salvatore, R. Liguori, S. Capone, A. Filla, G. Longo, and L. Sacchetti. A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients. *Information Sciences*, 180(21):4153–4163, 2010.

[23] I. Catalaa, J.C. Fulton, X. Zhang, J.K. Udupa, D. Kolson, M. Grossman, L. Wei, J.C. McGowan, M. Polansky, and R.I. Grossman. MR imaging quantitation of gray matter involvement in multiple sclerosis and its correlation with disability measures and neurocognitive testing. *American Journal of Neuroradiology*, 20(9):1613–1618, 1999.

[24] 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.

[25] A. Cerasa, E. Bilotta, A. Augimeri, A. Cherubini, P. Pantano, G. Zito, P. Lanza, P. Valentino, M.C. Gioia, and A. Quattrone. A cellular neural network methodology for the automated segmentation of multiple sclerosis lesions. *Journal of neuroscience methods*, 203(1):193–199, 2012.

[26] D.L. Collins, A.P. Zijdenbos, V. Kollokian, and J.G. Sled. Design and construction of a realistic digital brain phantom. *IEEE Trans. Med. Imag.*, 17(3):463–468, 1998. `http://www.bic.mni.mcgill.ca/brainweb`.

[27] A. Compston and A. Coles. Multiple sclerosis. *Lancet*, 359(9313):1221–1231, 2002.

[28] A. Compston and A. Coles. Multiple sclerosis. *Lancet*, 372:1502–1517, 2008.

[29] C. Confavreux, S. Vukusic, T. Moreau, and P. Adeleine. Relapses and progression of disability in multiple sclerosis. *New Engl. J. Med.*, 343(20):1430–1438, 2000.

[30] W.L. Curati, E.J. Williams, A. Oatridge, J.V. Hajnal, N. Saeed, and G.M. Bydder. Use of subvoxel registration and subtraction to improve demonstration of contrast enhancement in MRI of the brain. *Neuroradiology*, 38:717–723, 1996.

[31] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[32] Y. Diez, A. Oliver, M. Cabezas, S. Valverde, R. Martí, J.C. Vilanova, L. Ramió-Torrentà, A. Rovira, and X. Lladó. Intensity based methods for brain MRI longitudinal registration. a study on multiple sclerosis patients. *Neuroinformatics*, pages 1–15, 2013. Article in Press.

[33] Y. Diez, A. Oliver, X. Lladó, J. Freixenet, R. Martí, J. Martí, and J. C. Vilanova. Revisiting intensity-based image registration applied to mammography. *IEEE Trans. Inform. Technol. Biomed.*, 15(5):716–725, 2011.

[34] K. Doi. Diagnostic imaging over the last 50 years: Research and development in medical imaging science and technology. *Physics in Medicine and Biology*, 51(13):R5–R27, 2006.

[35] K. Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[36] Y. Duan, P.G. Hildenbrand, and M.P. Sampat. Segmentation of subtraction images for the measurement of lesion change in multiple sclerosis. *AJNR Am J Neuroradiol*, 29:340–346, 2008.

[37] C. Elliott, D.L. Arnold, D.L. Collins, and T. Arbel. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Trans. Med. Imag.*, 32(8):1490–1503, 2013.

[38] G.J. Ettinger, W.E.L. Grimson, T. Lozano-Perez, W.M. Wells III, S.J. White, and R. Kikinis. Automatic registration for multiple sclerosis change detection. pages 297–306, 1994.

[39] F. Fazekas, F. Barkhof, M. Filippi, R.I. Grossman, D.K.B. Li, W.I. McDonald, H.F. McFarland, D.W. Paty, J.H. Simon, J.S. Wolinsky, and D.H. Miller. Criteria for

an increased specificity of MRI interpretation in elderly subjects with suspected multiple sclerosis. *Neurology*, 38:1822–1825, 1988.

[40] Bernd Fischer and Jan Modersitzki. FLIRT: A Flexible Image Registration Toolbox. In James C. Gee, J.B. Antoine Maintz, and Michael W. Vannier, editors, *Biomedical Image Registration*, volume 2717 of *Lecture Notes in Computer Science*, pages 261–270. Springer Berlin Heidelberg, 2003.

[41] Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L. Arnold, and D. Louis Collins. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med. Image Anal.*, 17(1):1 – 18, 2013.

[42] Y. Ge. Multiple sclerosis: The role of MR imaging. *Am J Neuroradiol*, 27(6):1165–1176, 2006.

[43] E. Geremia, O. Clatz, B.H. 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.

[44] G. Gerig, J. Martin, R. Kikinis, O. Kübler, M. Shenton, and F.A. Jolesz. Unsupervised segmentetion of 3-D dual-echo MR head data. *Image and Vision Computing*, 10(6):349–360, 1992.

[45] G. Gerig, D. Welti, C.R.G. Guttmann, A.C.F. Colchester, and G. Székely. Exploring the discrimination power of the time domain for segmentation and characterization of active lesions in serial MR data. *Medical image analysis*, 4(1):31–42, 2000.

[46] C.R.G. Guttman, R. Kikinis, M.C. Anderson, M. Jakab, S.K. Warfield, R.J. Killiany, H.L. Weiner, and F.A. Jolesz. Quantitative follow-up of patients with multiple sclerosis using MRI: reproducibility. *J. Magn. Reson. Imaging*, 9:509–518, 1999.

[47] S.W. Hartley, A. I. Scher, E.S.C. Korf, L.R. White, and L.J. Launer. Analysis and validation of automated skull stripping tools: A validation study based on 296 MR images from the Honolulu Asia aging study. *NeuroImage*, 30(4):1179–1186, 2006.

[48] R.H. Hashemi, W.G. Bradley Jr, and C.J. Lisanti. *MRI: The Basics*. Lippincott Williams and Wilkins, 3 edition, 2010.

[49] D.L. Hill, P.G Batchelor, M. Holden, and D.J. Hawkes. Medical image registration. *Phys Med Biol*, 46(3):R1–R45, 2001.

[50] F.G. Hillary and B.B. Biswal. Automated detection and quantification of brain lesions in acute traumatic brain injury using MRI. *Brain Imaging and Behavior*, 3:111–112, 2009.

[51] M. Holden, D.L.G. Hill, E.R.E. Denton, J.M. Jarosz, T.C.S. Cox, T. Rohlfing, J. Goodey, and D.J. Hawkes. Voxel similarity measures for 3-D serial MR brain image registration. *IEEE Transactions on Medical Imaging*, 19(2):94–102, 2000.

[52] Z. Hou. A review on MR image intensity inhomogeneity correction. *International Journal of Biomedical Imaging*, 2006:1–11, 2006.

[53] M.A. Jacobs, P. Mitsias, H. Soltanian-Zadeh, S. Santhakumar, A. Ghanei, R. Hammond, D.J. Peck, M. Chopp, and S. Patel. Multiparametric MRI tissue characterization in clinical stroke with correlation to clinical outcome: part 2. *Stroke*, 32(4):950–957, 2001.

[54] M.A. Jacobs, Z.G. Zhang, R.A. Knight, H. Soltanian-Zadeh, A.V. Goussev, D.J. Peck, and M. Chopp. A model for multiparametric MRI tissue characterization in experimental cerebral ischemia with histological validation in rat: part 1. *Stroke*, 32(4):943–949, 2001.

[55] M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[56] M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

[57] B. Johnston, M.S. Atkins, B. Mackiewich, and M. Anderson. Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. *IEEE Transactions on Medical Imaging*, 15(2):154–169, 1996.

[58] LH. Juang and MN. Wu. MRI brain lesion image detection based on color-converted k-means clustering segmentation. *Measurement*, 43:941–949, 2010.

[59] H. Jun Yoon, B. Zheng, B. Sahiner, and D.P. Chakraborty. Evaluating computer-aided detection algorithms. *Medical physics*, 34(6):2024–2038, 2007.

[60] A. Kassner and R.E. Thornhill. Texture analysis: A review of neurologic MR imaging applications. *American Journal of Neuroradiology*, 31(5):809–816, 2010.

[61] R. Kikinis, C.R.G. Guttmann, and D. Metcalf. Quantitative follow-up of patients with multiple sclerosis using MRI: technical aspects. *J. Magn. Reson. Imaging*, 9:519–530, 1999.

[62] A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M. . Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, J.H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R.P. Woods, J.J. Mann, and R.V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, 2009.

[63] S. Klein, M. Staring, K. Murphy, M.A. Viergever, and J.P.W. Pluim. Elastix: a toolbox for intensity-based medical image registration. *Medical Imaging, IEEE Transactions on*, 29(1):196–205, 2010.

[64] Z. Lao, D. Shen, D. Liu, A.F. Jawad, E.R. Melhem, L.J. Launer, R.N. Bryan, and C. Davatzikos. Computer-Assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic Radiology*, 15(3):300–313, 2008.

[65] M.A. Lee, S. Smith, J. Palace, and P.M. Matthews. Defining multiple sclerosis disease activity using MRI T2-weighted difference imaging. *Brain*, 121:2095–2102, 1998.

[66] T.M. Lehmann, C. Gönner, and K. Spitzer. Survey: Interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 18(11):1049–1075, 1999.

[67] T.M. Lehmann, C. Gönner, and K. Spitzer. Addendum: B-spline interpolation in medical image processing. *IEEE Transactions on Medical Imaging*, 20(7):660–665, 2001.

[68] L. Lemieux, U. Wieshmann, N. Moran, D. Fish, and S. Shorvon. The detection and significance of subtle changes in mixed-signal brain lesions by serial MRI scan matching and spatial normalization. *Medical Image Analysis*, 2(3):227–242, 1998.

[69] H. Lester and S.R. Arridge. A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129–149, 1999.

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

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

[72] F.D. Lublin, S.C. Reingold, and National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Defining the clinical course of multiple sclerosis: Results of an international survey. *Neurology*, 46(4):907–911, 1996.

[73] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[74] J.B.A. Maintz and M.A. Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998.

[75] J. Martola, J. Bergström, S. Fredrikson, L. Stawiarz, J. Hillert, Y. Zhang, O. Flodmark, A. Lilja, A. Ekbom, P. Aspelin, and M.K. Wiberg. A longitudinal observational study of brain atrophy rate reflecting four decades of multiple sclerosis: a comparison of serial 1D, 2D, and volumetric measurements from MRI images. *Neuroradiology*, 52(2):109–117, 2010.

[76] D. Mattes, D.R. Haynor, H. Vesselle, T. Lewellen, and W. Eubank. Nonrigid multimodality image registration. In *Proc. of SPIE - The International Society for Optical Engineering*, volume 4322, pages 1609–1620, 2001.

[77] W.I. McDonald, A. Compston, G. Edan, D. Goodkin, H.-P. Hartung, F. D. Lublin, H.F. McFarland, D.W. Paty, C. H. Polman, S.C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. Van Den Noort, B.Y. Weinshenker, and J.S. Wolinsky. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Ann. Neurol.*, 50(4):121–127, 2001.

[78] D.S. Meier and C.R.G. Guttmann. Time-series analysis of MRI intensity patterns in multiple sclerosis. *NeuroImage*, 20:1193–1209, 2003.

[79] D. Metcalf, R. Kikinis, C. Guttmann, L. Vaina, and F. Jolesz. 4D connected component labelling applied to quantitative analysis of MS lesion temporal development. In *IEEE Eng. Med. Biol. Society*, pages 945–946, 1988.

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

[81] M. Modat, G.R. Ridgway, Z.A. Taylor, M. Lehmann, J. Barnes, D.J. Hawkes, N.C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[82] P.D. Molyneux. Precision and reliability for measurement of change in MRI lesion volume in multiple sclerosis: A comparison of two computer assisted techniques. *Journal of Neurology Neurosurgery and Psychiatry*, 65(1):42–47, 1998.

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

[84] X. Montalban, M. Tintoré, J. Swanton, F. Barkhof, F. Fazekas, M. Filippi, J. Frederiksen, L. Kappos, J. Palace, C. Polman, M. Rovaris, N. De Stefano, A. Thompson, T. Yousry, A. Rovira, and D.H. Miller. MRI criteria for MS in patients with clinically isolated syndromes. *Neurology*, 74(5):427–434, 2010.

[85] B. Moraal, D.S. Meier, and P.A. Poppe. Subtraction MR images in a multiple sclerosis multicenter clinical trial setting. *Radiology*, 250:506–514, 2009.

[86] B. Moraal, M.P. Wattjes, and J.J.G. Geurts. Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. *Radiology*, 255(1), 2010.

[87] D. Mortazavi, A.Z. Kouzani, and H. Soltanian-Zadeh. Segmentation of multiple sclerosis lesions in MR images: A review. *Neuroradiology*, 54(4):299–320, 2012.

[88] Otsu N. A threshold selection method from gray-level histogram. *IEEE Trans. Syst. Man Cybern*, 9:62–66, 1979.

[89] L.G. Nyul, J.K. Udupa, and Xuan Zhang. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imag.*, 19(2):143–150, 2000.

[90] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4):622–639, 2011.

[91] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache. Reconstructing a 3D structure from serial histological sections. *Image and Vision Computing*, 19(1-2):25–31, 2001.

[92] Sébastien Ourselin, Radu Stefanescu, and Xavier Pennec. Robust registration of multi-modal images: Towards real-time clinical applications. In Takeyoshi Dohi and Ron Kikinis, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002*, volume 2489 of *Lecture Notes in Computer Science*, pages 140–147. Springer Berlin Heidelberg, 2002.

[93] J. Patriarche and B. Erickson. A review of the automated detection of change in serial imaging studies of the brain. *J. Digital Imaging*, 17(3):158–174, 2004.

[94] P. Pieperhoff, M. Sudmeyer, L. Homke, K. Zilles, A. Schnitzler, and K. Amunts. Detection of structural changes of the human brain in longitudinally acquired MR images by deformation field morphometry: methodological analysis,validation and application. *NeuroImage*, 43(2):269–287, 2008.

[95] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images: A survey. *IEEE Trans. Med. Imag.*, 22(8):986–1004, 2003.

[96] C.H. Polman, S.C. Reingold, B. Banwell, M. Clanet, J.A. Cohen, M. Filippi, K. Fujihara, M. Hutchinson E. Havrdova and, L. Kappos, F.D. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A.J. Thompson, E. Waubant, B. Weinshenker, and J.S. Wolinsky. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Ann. Neurol.*, 69(2):292–302, 2011.

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

[98] C.M. Poser, D.W. Paty, and L. Scheinberg. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann. Neurol.*, 13:227–231, 2000.

[99] 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.

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

[101] A. Roche, G. Malandain, X. Pennec, and N. Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 1115–1124, 1998.

[102] K.N. Rode and R.T. Patil. Texture analysis of MRI using SVM & ANN for multiple sclerosis patients. *International Journal of Engineering Research and Applications (IJERA)*, 2(4):1925–1928, 2012.

[103] A. Rovira and A. León. MR in the diagnosis and monitoring of multiple sclerosis: An overview. *Eur. J. Radiol.*, 67(3):409–414, 2008.

[104] A. Rovira, J. Swanton, M. Tintoré, E. Huerga, F. Barkhof, M. Filippi, J. L. Frederiksen, A. Langkilde, K. Miszkiel, C. Polman, M. Rovaris, J. Sastre-Garriga, D. Miller, and X. Montalban. A single, early magnetic resonance imaging study in the diagnosis of multiple sclerosis. *Arch. Neurol.*, 66(5):587–592, 2009.

[105] M.A. Sahraian and A. Eshaghi. Role of MRI in diagnosis and treatment of multiple sclerosis. *Clin Neurol Neurosurg*, 112(7):609–615, 2010.

[106] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D.L. Arnold, D.L. Collins, and T. Arbel. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical image analysis*, 15(2):267–282, 2011.

[107] S. Shen, A. Szameitat, and A. Sterr. Detection of infarct lesions from brain MRI images using inconsistency between voxel intensity and spatial location. A 3D automatic approach. *IEEE Trans Inf Technol Biomed*, 12(4):532–540, 2008.

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

[109] J.G. Sled, A.P. Zijdenbos, and A.C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imag.*, 17(1):87–97, 1998.

[110] S.M. Smith. Fast robust automated brain extraction. *Hum. Brain Mapp.*, 17(3):143–155, 2002.

[111] S.M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P.M. Matthews, A. Federico, and N. De Stefano. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479–489, 2002.

[112] J. Solomon and A. Sood. 4-D lesion detection using expectation-maximization and hidden markov model. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Macro to Nano*, volume 1, pages 125–128, 2004.

[113] J.C. Souplet, C. Lebrun, N. Ayache, and G. Malandain. An automatic segmentation of T2-FLAIR multiple sclerosis lesions. In *Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge*, pages 1–11, 2008.

[114] S. Srivastava, F. Maes, D. Vandermeulen, W.V. Paesschen, P. Dupont, and P. Suetens. Automatic detection of focal cortical dysplastic lesions. *NeuroImage*, 27:253–266, 2005.

[115] C. Studholme, V. Cardenas, E. Song, F. Ezekiel, A. Maudsley, and M. Weiner. Accurate template-based correction of brain MRI intensity distortion with application to dementia and aging. *IEEE Transactions on Medical Imaging*, 23(1):99–110, 2004.

[116] C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[117] J.K. Swanton, A. Rovira, M. Tintore, D.R. Altmann, F. Barkhof, M. Filippi, E. Huerga, K.A. Miszkiel, G.T. Plant, C. Polman, M. Rovaris, A.J. Thompson, X. Montalban, and D.H. Miller. MRI criteria for multiple sclerosis in patients presenting with clinically isolated syndromes: A multicentre retrospective study. *Lancet Neurology*, 6(8):677–686, 2007.

[118] E. M. Sweeney, R.T. Shinohara, C.D. Shea, D.S. Reich, and C.M. Crainiceanu. Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *Am. J. Neuroradiol.*, 34(1):68–73, 2013.

[119] I.L. Tan, R.A.V. Schijndel, F. Fazekas, M. Filippi, P. Freitag, D.H. Miller, T.A. Yousry, P.J.W. Pouwels, H.J. Adèr, and F. Barkhof. Image registration and subtraction to detect active T2 lesions in MS: An interobserver study. *J. Neurology*, 249:767–773, 2002.

[120] I.L. Tan, R.A.V. Schijndel, M.A.A. van Walderveen, M. Quist, R. Bos, P.J.W. Pouwels, P. Desmedt, H.J. Adèr, and F. Barkhof. Magnetic resonance image regis-

tration in multiple sclerosis: comparison with repositioning error and observer-based variability. *J. Magn. Reson. Imag.*, 15(5):505–510, 2002.

[121] Jean-Philippe Thirion. Non-rigid matching using demons. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 245–251, 1996.

[122] J.P. Thirion and G. Calmon. Deformation analysis to detect and quantify active lesions in three-dimensional medical image sequences. *IEEE Trans. Med. Imag.*, 18(5):429–441, 1999.

[123] A. J. Thompson and P. Baneke. Atlas of MS 2013. In *Atlas Of MS 2013*. Multiple Sclerosis International Federation, Modern Colour Solutions, 2013.

[124] W. Tian, T. Zhu, J. Zhong, X. Liu, P. Rao, B.M. Segal, and S. Ekholm. Progressive decline in fractional anisotropy on serial DTI examinations of the corpus callosum: A putative marker of disease activity and progression in SPMS. *Neuroradiology*, 54(4):287–297, 2012.

[125] M. Tintoré, A. Rovira, M.J. Martínez, J. Rioa, P.D Villosladaa, L. Brievaa, C. Borrása, E. Grivéa, J. Capelladesa, and X. Montalbana. Isolated demyelinating syndromes: comparison of different MR imaging criteria to predict conversion to clinically definite multiple sclerosis. *Am. J. Neuroradiol.*, 21:702–706, 2000.

[126] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, and J.C. Gee. N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imag.*, 29(6):1310–1320, 2010.

[127] J.K. Udupa and S. Samarasekera. Fuzzy connectedness and object definition: Theory, algorithms, and applications in image segmentation. *Graphical Models and Image Processing*, 58(3):246–261, 1996.

[128] J.K. Udupa, L. Wei, S. Samarasekera, Y. Miki amnd M.A. Van Buchem, and R.I. Grossman. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. *IEEE Trans. Med. Imag.*, 16(5):598–609, 1997.

[129] M. Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

[130] M. Unser. Splines: A perfect fit for medical imaging. In *Proceedings of SPIE - The International Society for Optical Engineering*, volume 4684 I, pages 225–236, 2002.

[131] M. Unser, A. Aldroubi, and M. Eden. B-spline signal processing. part I. theory. *IEEE Transactions on Signal Processing*, 41(2):821–833, 1993.

[132] M. Unser, A. Aldroubi, and M. Eden. B-spline signal processing. part II. efficient design and applications. *IEEE Transactions on Signal Processing*, 41(2):834–848, 1993.

[133] I.J. Van den Elskamp, B. Boden, V. Dattola, D.L. Knol, M. Filippi, L. Kappos, F. Fazekas, K. Wagner, C. Pohl, R. Sandbrink, C.H. Polman, B.M. Uitdehaag, and F. Barkhof. Cerebral atrophy as outcome measure in short-term phase 2 clinical trials in multiple sclerosis. *Neuroradiology*, 52(10):875–881, 2010.

[134] I. M. Vavasour, D.K.B. Li, C. Laule, A. L. Traboulsee, G.R.W. Moore, and A.L. MacKay. Multi-parametric MR assessment of T1 black holes in multiple sclerosis: Evidence that myelin loss is not greater in hypointense versus isointens T1 lesions. *Journal of neurology*, 254(12):1653–1659, 2007.

[135] 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, 2007.

[136] L. Wang, H. . Lai, G.J. Barker, D.H. Miller, and P.S. Tofts. Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. *Magnetic Resonance in Medicine*, 39(2):322–327, 1998.

[137] S.K. Warfield, M. Kaus, and F.A. Jolesz. Adaptive, template moderated, spatially varying statistical classification. *Med. Image Anal.*, 4(1):43–55, 2000.

[138] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[139] X. Wei, S.K. Warfield, K.H. Zou, Y. Wu, X. Li, A. Guimond, J.P. Mugler III, R.R. Benson, L. Wolfson, H.L. Weiner, and C.R.G. Guttmann. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. *J. Magn. Reson. Imaging*, 15:203–209, 2002.

[140] H.L. Weiner, C.R. Guttmann, S.J. Khoury, E.J. Orav, M.J. Hohol, R. Kikinis, and F.A. Jolesz. Serial magnetic resonance imaging in multiple sclerosis: correlation with attacks, disability, and disease stage. *Journal of Neuroimmunology*, 104:164–173, 2000.

[141] W.M. Wells III and W.E.L. Grimson. Adaptive segmentation of MRI data. *IEEE Trans. Med. Imag.*, 15:429–443, 1996.

[142] W.M. Wells III, W.E.L. Grimson, R. Kikinis, and F.A. Jolesz. Statistical intensity correction and segmentation of MRI data. pages 13–24. SPIE Conf. Visualization Biomed. Computing, 1994.

[143] D. Welti, G. Gerig, E.W. Radü, L. Kappos, and G. Székely. Spatio-temporal segmentation of active multiple sclerosis lesions in serial MRI data. In *Int. Conf. Inform. Proc. Medical Imaging*, pages 438–445, 2001.

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

[145] D. Yamamoto, H. Arimura, S. Kakeda, T. Magome, Y. Yamashita, F. Toyofuku, M. Ohki, Y. Higashida, and Y. Korogi. Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. *Computerized Medical Imaging and Graphics*, 34(5):404–413, 2010.

[146] E.I. Zacharaki, S. Kanterakis, R.N. Bryan, and C. Davatzikos. Measuring brain lesion progression with a supervised tissue classification system. *Med Image Comput Assist Interv*, 11:620–627, 2008.

[147] J. Zar. Measures of dispersion and variability. *Biostatistical analysis*, pages 27–39, 1984.

[148] J. Zhang, L. Tong, L. Wang, and N. Li. Texture analysis of multiple sclerosis: A comparative study. *Magnetic resonance imaging*, 26(8):1160–1166, 2008.

[149] J. Zhang, L. Wang, and L. Tong. Feature reduction and texture classification in MRI-texture analysis of multiple sclerosis. In *2007 IEEE/ICME International Conference on Complex Medical Engineering, CME 2007*, pages 752–757, 2007.

[150] Y. Zhang. MRI texture analysis in multiple sclerosis. *International Journal of Biomedical Imaging*, 2012, 2012. Article ID 762804.

[151] 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.

[152] A.P. Zijdenbos, R. Forghani, and A.C. Evans. Automatic Pipeline analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. *IEEE Trans. Med. Imag.*, 21(10):1280–1291, 2002.

[153] B. Zitová and J. Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21(11):977–1000, 2003.