



Deep Learning in Neuroimaging for Multiple Sclerosis

Reda Abdellah Kamraoui

► To cite this version:

Reda Abdellah Kamraoui. Deep Learning in Neuroimaging for Multiple Sclerosis. Medical Imaging. Université de Bordeaux, 2023. English. ⟨NNT : 2023BORD0005⟩. ⟨tel-04055255v2⟩

HAL Id: tel-04055255

<https://hal.science/tel-04055255v2>

Submitted on 20 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE
DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Reda Abdellah KAMRAOUI**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

**Apprentissage Profond en Neuroimagerie pour la
Sclérose en Plaques**

Soutenue le 13 Janvier 2023

Devant la commission d'examen composée de :

Dr. Jean-Francois MANGIN	Directeur de Recherche, CEA ...	Président du jury
Dr. Caroline PETITJEAN .	Professeure, Université de Normandie	Rapportrice
Dr. Louis COLLINS	Professeur, Université de McGill ..	Rapporteur
Dr. Aurélie BUGEAU ...	Professeure, Université de Bordeaux	Examinatrice
Dr. Pierrick COUPÉ ...	Directeur de Recherche, CNRS ...	Directeur de thèse

Résumé L'imagerie par résonance magnétique (IRM) est couramment utilisée pour le diagnostic et le pronostic de la sclérose en plaques (SEP). L'analyse et l'extraction des informations de l'IRM pourraient être effectuées manuellement par des radiologues ou des experts, néanmoins, ces tâches sont fastidieuses, chronophages, nécessitent une expertise du domaine et sont sujettes à la variabilité inter-évaluateurs. Ainsi, l'automatisation des tâches d'analyse des IRM a été envisagée pour faire face à ces limitations et traiter la grande quantité de données que nous rencontrons à l'ère des données massives. Dans cette thèse, nous proposons des chaînes de traitement utilisant l'apprentissage profond pour analyser les IRM afin d'en extraire des informations pertinentes pour la SEP. La suite d'outils comprend la segmentation des lésions de SEP, la segmentation/détection de nouvelles lésions et l'estimation du statut d'invalidité à partir de données IRM et clinico-démographiques. Lors de la conception de chaque chaîne de traitement, nous avons proposé des contributions méthodologiques qui ont résolu différents défis, tels que le biais de domaine, la rareté des données et le déséquilibre des données. Nos chaînes de traitement sont hébergées sur la plate-forme volBrain pour les rendre librement et facilement utilisables par la communauté sans avoir besoin de logiciels ou de matériels. Ce faisant, nos utilisateurs bénéficient de performances de pointe en quelques clics sur leur navigateur Web et obtiennent ainsi un rapport compact et facile à lire résumant les résultats.

Mots-clés Apprentissage Profond, Sclérose en Plaques, Données Massives

Laboratoire d'accueil Laboratoire Bordelais de Recherche en Informatique (LaBRI), 351 cours de la Libération, 33405 Talence

Abstract Magnetic Resonance Imaging (MRI) is routinely used for the diagnosis and prognosis of Multiple Sclerosis (MS). Analyzing and extracting information from MRI could be performed manually by radiologists or experts in the field, nonetheless, these tasks are tedious, time-consuming, require domain expertise, and are prone to inter-rater variability. Thus, automatizing MRI analysis tasks would enable to overcome these limitations and process a large amount of data available in our BigData era. In this thesis, we proposed end-to-end pipelines using deep learning to analyze MRI and extract MS-relevant information. The proposed suite of tools includes MS lesion segmentation, new lesion detection, and the estimation of disability status from MRI and clinico-demographic data. During the design of each pipeline, we proposed methodological contributions that solved the different challenges, such as domain bias, data rarity, and data imbalance. Our pipelines are hosted on volBrain online platform, to make them freely and easily available to the MS community without the need of software or hardware requirements. Doing so, our users benefit from state-of-the-art performance in a few clicks on their web browser and obtain compact easy-to-read reports.

Keywords Deep Learning, Multiple Sclerosis, BigDATA

Hosting Laboratory Laboratoire Bordelais de Recherche en Informatique (LaBRI), 351 cours de la Libération, 33405 Talence

Contents

Résumé étendu en français	1
Acknowledgment	5
Publications	7
1 Introduction	9
1.1 Multiple Sclerosis	9
1.1.1 The Neuroanatomical Context of MS	11
1.1.2 The Radiological Context of MS	15
1.1.3 The Clinical Context of MS	19
1.2 Image Processing Tasks for MS	21
1.2.1 Preprocessing	22
1.2.2 Automated Tasks	24
1.3 Thesis overview	29
2 MS Lesions Segmentation	31
2.1 Introduction	32
2.1.1 Problem Description	32
2.1.2 Related Works	33
2.1.3 Proposals	35
2.2 Method and Material	36
2.2.1 Method Overview	36
2.2.2 Implementation Details	39
2.2.3 Method Description	39
2.2.4 Datasets	41
2.2.5 Validation Framework	44
2.3 Results	46
2.3.1 Ablation Study	46
2.3.2 Cross-dataset Testing	48
2.3.3 Same Domain Validation	51

2.3.4	Cross-dataset Segmentation Consistency	53
2.4	Discussion and Conclusion	55
2.4.1	Discussion	55
2.4.2	Conclusion	58
3	New MS Lesions Segmentation and Detection	61
3.1	Introduction	62
3.2	Method and Material	63
3.2.1	Method Overview	63
3.2.2	Transfer-learning from single time-point MS lesion segmen- tation task	65
3.2.3	Time-points Synthesis	66
3.2.4	Data Augmentation	69
3.2.5	Data	70
3.2.6	Implementation Details	72
3.2.7	Validation Framework	72
3.3	Results	73
3.3.1	Internal Validation	74
3.3.2	Challenge Evaluation	75
3.4	Discussion	76
3.5	Conclusion	79
4	Automatic EDSS Estimation based on MRI and Clinico-demographic Data using Deep Learning	81
4.1	Introduction	82
4.2	Method and Material	85
4.2.1	Dataset	85
4.2.2	Multi-phase Training and targeted augmentations	88
4.2.3	Dual path DL model	90
4.2.4	Implementation Details	92
4.2.5	Evaluation Metrics	92
4.2.6	Reference Methods	93
4.3	Results	93
4.3.1	Ablation Study	93
4.3.2	Method Comparison	95
4.3.3	Model Performance based on EDSS ranges	96
4.3.4	Model interpretability by EDSS range	97
4.4	Discussion	98
4.5	Conclusion	99

5	MRI Analysis Pipelines for MS	101
5.1	Introduction	101
5.2	Pipeline Accessibility	101
5.2.1	Containerization	101
5.2.2	Hosting Platform: volBrain	102
5.3	Pipeline Description	102
5.4	Pipeline Report	105
5.4.1	DeepLesionBrain for MS lesion Segmentation	105
5.4.2	Longitudinal DLB for the Detection of New MS lesions . . .	107
5.4.3	DeepLesionBrain update for EDSS estimation	109
5.5	Conclusion	110
6	General Conclusion and Perspectives	111
6.1	Conclusion	111
6.2	Perspectives	113
6.2.1	Extensions related to this Ph.D.	113
6.2.2	General perspectives	114
A	POPCORN: Progressive Pseudo-labeling with Consistency Regularization and Neighboring	115
A.1	Abstract	115
A.2	Introduction	116
A.3	Method	117
A.3.1	Method overview	117
A.3.2	Bottleneck consistency regularization	117
A.3.3	Pseudo-labeling data selection	119
A.3.4	Proximity graph	119
A.4	Experiments	120
A.4.1	Dataset	120
A.4.2	Reference Methods	120
A.4.3	Implementation details	121
A.4.4	Statistical Analysis	121
A.5	Results	122
A.5.1	Ablation study	122
A.5.2	Comparison with state-of-the-art approaches	123
A.6	Conclusion	124
	Acronyms	125
	References	127

List of Figures

1.1	The demyelination caused by MS	10
1.2	MR images of MS lesions	10
1.3	The central nervous system	11
1.4	Gray matter and white matter	12
1.5	The cortex lobes and the inner brain	12
1.6	The cerebellum and the brainstem	14
1.7	The brain ventricles	14
1.8	MRI planes	15
1.9	MS lesion appearance on different MRI modalities	15
1.10	Lesion types based on MRI intensity	17
1.11	MS lesion distribution on the brain	18
1.12	Brain atrophy in MS	19
1.13	Disability progression for MS subtypes	21
1.14	Preprocessing steps	23
1.15	The convolution operation	26
1.16	The fully connected layers	27
2.1	The two-steps training process of DeepLesionBrain	36
2.2	Illustration of the considered U-Net architecture	39
2.3	FLAIR examples from the three used datasets	43
2.4	DLB and SOTA segmentation on image samples from In-house dataset	55
2.5	DLB and SOTA segmentation on image samples from MSSEG'16 dataset	56
2.6	DLB and SOTA segmentation on image samples from ISBI dataset	57
3.1	The pipeline of our new MS lesion segmentation method	64
3.2	Diagram representing our training method	65
3.3	Architecture used for our work	67
3.4	Synthetic time points with new MS lesion generation pipeline . . .	68
3.5	Examples of data augmentation applied on FLAIR images	69
3.6	The validation based on train-set size	75

3.7	Longitudinal DLB vs. expert rater segmentations	78
4.1	The image-based features used for the estimation of EDSS	86
4.2	Histograms by EDSS ranges	89
4.3	DL Architecture used for EDSS estimation	91
4.4	The performance of the EDSS estimation model on each EDSS range	96
4.5	Visualizations of brain regions that contribute the most to the estimation of EDSS	100
5.1	volBrain pipeline selection	103
5.2	volBrain web interface to visualize results	104
5.3	DeepLesionBrain report (MS lesions)	105
5.4	DeepLesionBrain report (Tissues and Macro-structures)	106
5.5	DeepLesionBrain report (Structures segmentation)	106
5.6	Longitudinal DLB report (MS lesions changes)	108
5.7	Longitudinal DLB report (Tissues and whole brain structures changes)	108
5.8	DeepLesionBrain update related to EDSS estimation	109
A.1	The training process of POPCORN	118
A.2	Segmentation comparison of POPCORN and other SOTA methods	124

List of Tables

1.1	MS Disabilities based on lesion location	20
2.1	Description of datasets used in this work.	42
2.2	Ablation study results trained and tested with ISBI	47
2.3	Ablation study results trained on ISBI and tested on MSSEG'16 . .	47
2.4	Results of the different approaches trained on the ISBI training dataset	49
2.5	Results of the different approaches trained on the MSSEG'16 dataset	50
2.6	Results of the different approaches trained on In-house dataset . . .	50
2.7	Summary of the cross-dataset experiment	51
2.8	Results of SOTA approaches trained on the ISBI training dataset .	52
2.9	State-of-the-art published results for the ISBI challenge	52
2.10	The consistency of the segmentations for each approach	54
3.1	Summary of the used datasets	70
3.2	The internal validation results for the ablation study	74
3.3	Results of MSSEG2-challenge evaluation	77
4.1	Dataset metadata	88
4.2	Results of our experiment on different combinations of features as input	94
4.3	Results of our EDSS estimation model compared to SOTA methods	95
4.4	Brain structures with high GradCAM density by EDSS ranges . . .	98
A.1	Results of an ablation study of the key components of POPCORN .	122
A.2	The evaluation of POPCORN compared to other SOAT strategies .	123

Résumé étendu en français

La Sclérose En Plaques (SEP) est une maladie neurologique inflammatoire qui affecte le système nerveux central. Cette pathologie auto-immune engendre une altération de la gaine de myéline qui recouvre les axones, elle induit donc des perturbations dans le flux du signal neuronal. Cette maladie provoque des troubles moteurs, sensoriels et cognitifs. Elle débute par des déficits épisodiques réversibles pouvant évoluer vers une détérioration progressive. L'imagerie par résonance magnétique (IRM) est couramment utilisée pour le diagnostic et le pronostic de la SEP. L'analyse et l'extraction des informations de l'IRM pourraient être effectuées manuellement par des radiologues ou des experts, néanmoins, ces tâches sont fastidieuses, chronophages, nécessitent une expertise du domaine et sont sujettes à la variabilité inter-évaluateurs. Ainsi, l'automatisation des tâches d'analyse IRM a été envisagée pour faire face à ces limitations et traiter la grande quantité de données que nous rencontrons à l'ère des données massives. Dans cette thèse, nous proposons une suite d'outils pour l'automatisation des tâches de neuroimagerie appliquées à la SEP.

Dans le Chapitre 1, nous apportons le contexte nécessaire lié à cette thèse. Tout d'abord, nous présentons le contexte neuroanatomique de la SEP qui est le Système Nerveux Central (SNC). Ce dernier est composé du cerveau et de la moelle épinière. Le cerveau contrôle la plupart des fonctions corporelles, y compris la pensée, la perception, la parole, le mouvement, la mémoire, et les sentiments. La moelle épinière transporte des signaux nerveux, permettant la communication entre le cerveau et le système nerveux périphérique innervant l'ensemble du corps. Deuxièmement, nous présentons le contexte radiologique de la SEP à travers l'IRM et sa capacité à visualiser les tissus et les anomalies du SNC. En effet, l'IRM est la technique d'imagerie la plus efficace pour diagnostiquer, évaluer et suivre la progression de la SEP. L'IRM permet d'observer la neuroinflammation (lésions) et la neurodégénérescence (atrophie) provoquées par la SEP. Troisièmement, nous abordons l'aspect clinique de la SEP représenté par ses différents types d'évolution et les poussées (épisode clinique de la SEP). Ces dernières ainsi que la dissémination spatiale et temporelle des lésions font de la SEP une maladie complexe. De plus, le lien entre les manifestations radiologiques et cliniques de la SEP n'est toujours

pas clair. De ce fait, son diagnostic, son suivi ou sa prédiction d'évolution nécessitent le développement de méthodes spécifiques pour aider les cliniciens dans leurs tâches. Enfin, nous expliquons les différentes approches automatiques permettant de résoudre les difficultés liées à l'analyse manuelle des IRM. De nombreuses méthodes entièrement automatiques ont été proposées et peuvent être regroupées en méthodes non supervisées et supervisées.

Dans le Chapitre 2, nous présentons DeepLesionBrain, une nouvelle méthode de segmentation pour les lésions de la SEP qui est robuste au changement de domaine et performante sur des ensembles de données non utilisés lors de l'entraînement. Cette propriété de généralisation résulte de trois apports principaux. Tout d'abord, la méthode utilise un grand nombre de réseaux convolutionnels 3D compacts répartis sur l'ensemble du cerveau avec des sous-volumes d'analyse qui se chevauchent. En associant un réseau distinct à chaque région du cerveau, la stratégie des réseaux spatialement distribués simplifie la segmentation des lésions de la SEP, d'une seule tâche complexe sur l'ensemble du cerveau à plusieurs sous-tâches plus simples sur chaque région. De plus, les régions qui se chevauchent garantissent un consensus cohérent et stable. Deuxièmement, pour extraire des caractéristiques plus pertinentes qui peuvent conduire à une meilleure généralisation, la méthode est entraînée avec un apprentissage hiérarchique. La stratégie d'entraînement en deux étapes consiste à pré-entraîner un seul réseau sur toutes les régions du cerveau, et ensuite de l'utiliser dans l'étape suivante pour initialiser les poids de chaque réseau spatialement distribué. Troisièmement, DeepLesionBrain est entraîné avec une nouvelle méthode d'augmentation des données, qui imite la diversité des données du monde réel en ajoutant des modifications réalistes aux images d'entraînement. Ces augmentations spécifiques contraignent l'apprentissage à être indépendant de la résolution d'acquisition, du contraste ou de la qualité des données. Par conséquent, la stratégie d'augmentation proposée permet une meilleure robustesse au changement de domaine. La généralisation de la méthode a été validée dans des expériences de jeux de données croisés. Au cours de ces expériences, DeepLesionBrain a montré une plus grande précision et une meilleure cohérence de segmentation et de meilleures performances de généralisation par rapport aux méthodes de l'état de l'art. Ce chapitre a fait l'objet d'une publication dans le journal Medical Image Analysis [2].

Dans le Chapitre 3, nous décrivons une chaîne de traitement basée sur l'apprentissage profond abordant la tâche difficile de détecter et de segmenter les nouvelles lésions de SEP apparaissant entre deux examens. En effet, le manque de données longitudinales (images de suivi du patient au cours du temps) annotées pour cette tâche et la rareté des cas d'apparition de nouvelles lésions sont des facteurs limitants pour l'apprentissage de modèles robustes et généralisables. Ces problèmes de rareté de données et de déséquilibre des classes sont abordés à travers trois

contributions principales. Tout d'abord, l'apprentissage par transfert est proposé pour exploiter des jeux de données plus vastes et plus diversifiés disponibles pour la tâche de segmentation des lésions de SEP. En effet, les jeux de données utilisés pour entraîner la segmentation des lésions de la SEP sont plus facilement disponibles et la classe positive est plus fréquente (moins impactée par le déséquilibre des classes) par rapport à la tâche de détection de nouvelles lésions de la SEP. Par conséquent, l'exploitation des connaissances d'une tâche plus facile et similaire avec un jeu d'entraînement plus riche a considérablement amélioré la détection de nouvelles lésions. Deuxièmement, la chaîne de traitement comprend une nouvelle stratégie de synthèse de données pour générer des données longitudinales réalistes avec de nouvelles lésions à l'aide de données transversales (un seul examen sans suivi du patient). La stratégie combine l'utilisation d'un générateur de lésions et d'un effaceur de lésions (tous deux entraînés séparément et avant la chaîne de traitement décrite) pour générer "à la volée" des IRM synthétiques du même patient avec une évolution des lésions. De cette manière, le modèle est entraîné sur de grands jeux de données annotées synthétiques. Troisièmement, une version améliorée de la méthode d'augmentation de données présentée au Chapitre 2 est proposée pour simuler une plus grande diversité de données en IRM. La version améliorée introduit plus de modifications à la fois dans l'espace spatial et fréquentiel. Ainsi, la méthode d'augmentation de données permet de mieux suréchantillonner les échantillons rares avec de nouvelles lésions et rend le modèle entraîné plus robuste à la diversité des images. Notre validation a montré que chaque contribution conduit à une amélioration de la précision de la segmentation. En utilisant la chaîne de traitement proposée, nous avons obtenu le meilleur score pour la segmentation et la détection de nouvelles lésions de SEP lors du challenge international MICCAI MSSEG2. Ce chapitre a fait l'objet d'une publication dans le journal *Frontiers in Neuroscience* [1].

Au Chapitre 4, nous proposons une nouvelle méthode pour l'estimation de l'Échelle de Statut d'Invalidité Étendue (EDSS en anglais) à partir d'informations IRM et clinico-démographiques à l'aide d'apprentissage profond. Pour y parvenir, nous proposons dans un premier temps d'extraire des biomarqueurs neurodégénératifs et neuroinflammatoires pertinents. Ensuite, nous proposons une architecture de réseaux convolutionnels qui combine efficacement des données images et des tableaux. De plus, notre modèle est entraîné avec un nouvel apprentissage multiphase et une augmentation des données pour atténuer l'effet de déséquilibre des données. Lors de la validation, notre méthode par apprentissage profond a surpassé les méthodes conventionnelles. Pour mieux comprendre quelles régions du cerveau contribuent le plus à la prédiction de notre modèle d'estimation EDSS, nous avons proposé d'utiliser des méthodes d'interprétabilité de modèle utilisant l'apprentissage profond. Les visualisations obtenues ont montré une cohérence avec

les travaux associant les troubles fonctionnels de la SEP durant différents stades de la maladie, aux structures cérébrales responsables de ces fonctions. Ce chapitre a fait l'objet d'un article soumis au journal Artificial Intelligence in Medicine [4].

Dans le Chapitre 5, nous proposons des chaînes de traitements complètes pour le traitement des données de neuroimagerie multimodale et l'extraction des informations pertinentes pour la SEP sur la base des travaux proposés dans les chapitres 2 à 4. Tout d'abord, nous proposons d'utiliser la conteneurisation et d'héberger nos chaînes de traitements sur une plate-forme en ligne pour permettre à la communauté travaillant sur la SEP de les utiliser facilement sans avoir à gérer l'installation et les exigences matérielles. Ensuite, nous proposons des chaînes de traitements faciles à utiliser qui ne nécessitent pas d'expertise technique pour exécuter et récupérer leurs résultats. Enfin, nous proposons de générer des rapports offrant des informations compactes, organisées et faciles à lire. Le développement logiciel décrit dans ce chapitre a fait l'objet d'un dépôt à l'agence pour la protection des programmes [7].

Enfin, nous finissons ce manuscrit par une conclusion générale et nos perspectives futures.

Acknowledgment

First and foremost, I am extremely grateful to my supervisor, Dr. Pierrick Coupé for his assistance, his guidance, and dedicated involvement during my Ph.D. study. For this, I will be forever in his dept.

I would like to thank Boris Mansencal, Huy Dung Nguyen, Nicolas Papadakis, José V Manjón, Thomas Tourdias, and Ismail Koubiyr for their help and instructive comments.

I would like to express my sincere gratitude to all the members of my thesis committee. I would like to thank Dr. Aurélie Bugeau from Bordeaux University, Dr. Louis Collins from McGill University, Dr. Jean-Francois Mangin from Paris-Saclay University, and Dr. Caroline Petitjean from Rouen University for spending their valuable time reviewing my thesis.

I wish to show my appreciation to Tarek Bouzar and Noureddine Mouhoub for sticking up with me in one of the most challenging times I underwent. Special thanks to Mehdi Rahou, Alaa Belarbi, Bakhti Yassine, Abdelwahid Benzerrouk, and Nasreddine Barigou. I also want to thank Sarra Zenagui for the beautiful moments we spent together and for taking part in my journey.

Last but not the least, I must express my very profound gratitude and deepest love to my parents, my brother, and my sister for all the support and kindness they gave me during my whole life and especially these last three years.

Publications

Articles Accepted in Peer-reviewed Journals

- [1] Reda Abdellah Kamraoui, Boris Mansencal, José V Manjon, and Pierrick Coupé. Longitudinal detection of new MS lesions using Deep Learning. *Frontiers in Neuroscience*, 2022.
- [2] Reda Abdellah Kamraoui, Vinh-Thong Ta, Thomas Tourdias, Boris Mansencal, José V Manjon, and Pierrick Coupé. DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis*, 76:102312, 2022.
- [3] T Yamamoto, C Lacheret, H Fukutomi, Reda Abdellah Kamraoui, L Denat, B Zhang, V Prevost, L Zhang, A Ruet, B Triaire, et al. Validation of a denoising method using deep learning-based reconstruction to quantify multiple sclerosis lesion load on fast FLAIR imaging. *American Journal of Neuroradiology*, 43(8):1099–1106, 2022.

Articles in Preparations or Submitted

- [4] Reda Abdellah Kamraoui, Boris Mansencal, Ismail Koubiyr, Bruno Brochet, Aurélie Ruet, Thomas Tourdias, Pierrick Coupé, and on behalf of OFSEP invistigators. Automatic EDSS estimation based on MRI and clinico-demographic data using deep learning. *Artificial Intelligence in Medicine*, 2023.
- [5] Ismail Koubiyr, Reda Abdellah Kamraoui, Pierrick Coupé, Thomas Tourdias, and et al. Thalamic nuclei at CSF interface are the timer of neurodegeneration progression in multiple sclerosis. 2023.

Articles Accepted in Peer-reviewed Conferences

- [6] Reda Abdellah Kamraoui, Vinh-Thong Ta, Nicolas Papadakis, Fanny Compaire, José V Manjon, and Pierrick Coupé. POPCORN: Progressive pseudo-labeling with consistency regularization and neighboring. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 373–382. Springer, 2021.

Protected Software

- [7] Pierrick Coupé, José V Manjon, Reda Abdellah Kamraoui, and Boris Mansencal. Deeplesionbrain: Brain lesion segmentation. *IDDN.FR.001.190024.000.S.C.2022.000.31230*, 2022.

Posters and Short-papers

- [8] Reda Abdellah Kamraoui, Vinh-Thong Ta, José V Manjon, and Pierrick Coupé. Draw and erase to learn better. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 33, 2021.
- [9] Reda Abdellah Kamraoui, Vinh-Thong Ta, José V Manjon, and Pierrick Coupé. Image quality data augmentation for new MS lesion segmentation. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 37, 2021.
- [10] Reda Abdellah Kamraoui, Vinh-Thong Ta, José V Manjon, and Pierrick Coupé. New MS lesion segmentation with lesion-wise metrics learning. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 29, 2021.

Chapter 1

Introduction

1.1 Multiple Sclerosis

Multiple Sclerosis (MS) is an inflammatory and neurodegenerative disease that affects Central Nervous System (CNS). It is one of the most common neurological disorders among young adults, with a risk of developing the disease 2.3 times higher for women compared to men. MS affects more than 2.8 million people worldwide¹. Although the cause of MS is unknown to this day, a combination of genetics and environmental factors appears to be responsible.

MS is an autoimmune condition in which the myelin sheath, an insulating layer that covers the axons and helps the propagation of action potentials, is damaged in the brain and spinal cord (see Figure 1.1). Thus, the loss of myelin (demyelination) causes a disruption in the ability of the nerves to conduct electrical impulses to and from the CNS and can evolve into permanent axonal damage, causing handicaps in the later stages of the disease. The attacks cause the myelin layer to become inflamed in small accumulations, commonly called MS lesions. Using *in vivo* Magnetic Resonance Imaging (MRI), MS lesions can be visualized as high-intensity spot areas (see Figure 1.2).

This disease impacts the CNS leading to motor, sensorial, and cognitive impairments. Depending on the severity and the form of MS, patients experience symptoms ranging from small discomfort (such as numbness and tingling), minimal motor disability (difficulty walking), severe disability (need to use a wheelchair) and, in the most extreme cases, the requirement of permanent medical assistance and MS-related death.

In the following, we will detail the neuroanatomical, radiological, and clinical contexts of MS.

¹www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms#about

²www.alamyimages.fr

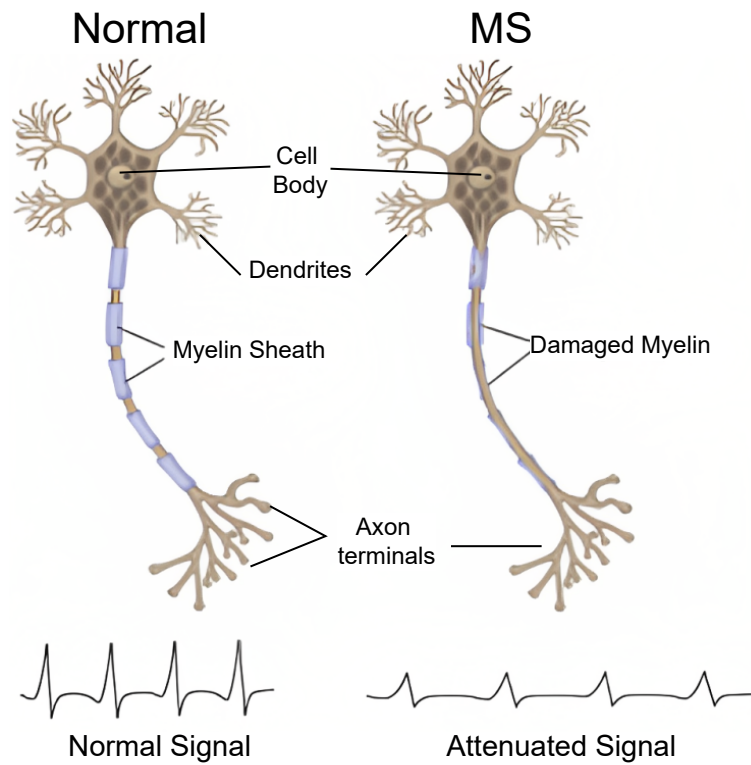


Figure 1.1: The demyelination caused by Multiple Sclerosis causing signal attenuation. Modified from: ².

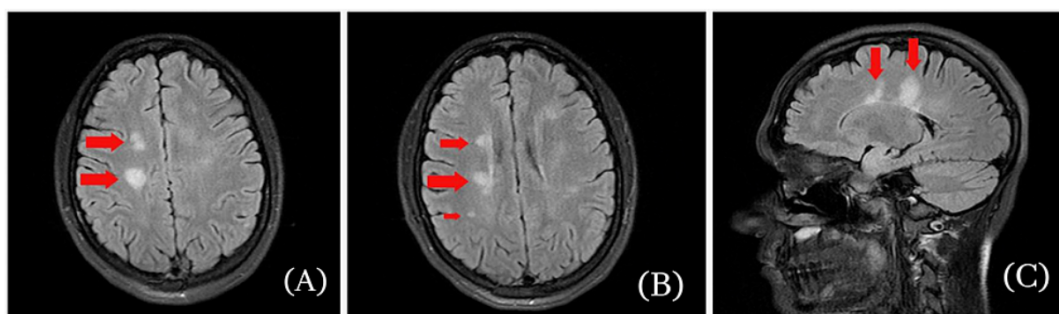


Figure 1.2: MR images of MS lesions with the T2 sequence. A, B: Axial views and C: Sagittal view. Source: [Al-Midfai et al., 2022].

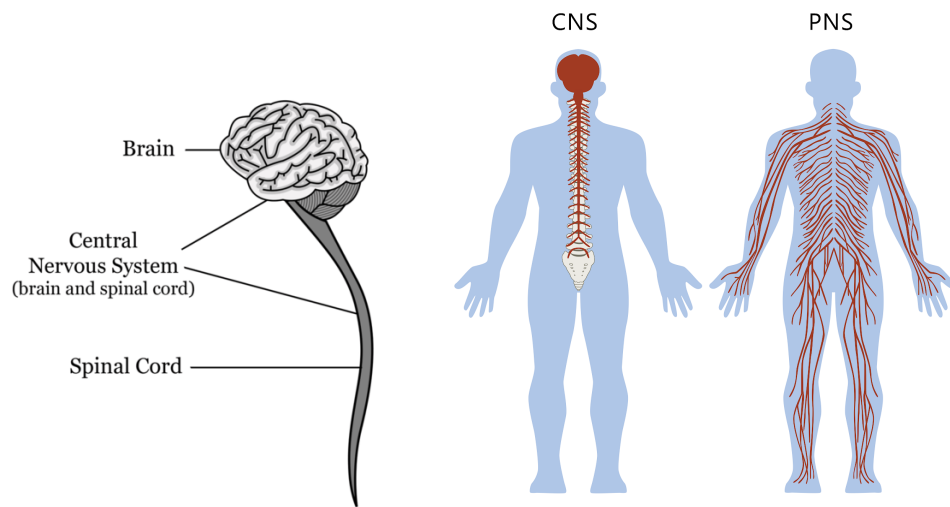


Figure 1.3: On the left, (a) the CNS is composed of the brain and spinal cord. On the right, (b) the CNS and the Peripheral Nervous System (PNS).
Modified from: ³.

1.1.1 The Neuroanatomical Context of MS

The CNS includes the brain and the spinal cord (see Figure 1.3 (a)). The brain controls most bodily functions, including perception, movement, feeling, thought, speech, and memory. The spinal cord attaches to the brain by the brainstem and is protected by the vertebrae, which form the vertebral column. Nerves emerge from the spinal cord to innervate both sides of the body. The spinal cord carries nerve signals, allowing communication between the brain and the peripheral nervous system, composed of the nerves throughout the rest of the body (see Figure 1.3 (b)).

In the following of this section, we will focus on the brain since it is the most complex part of the CNS and extensively studied in MS.

Brain tissues can be categorized into Gray Matter (GM) and White Matter (WM), as shown in Figure 1.4. The GM of the brain contains the cell bodies of nerve cells (neurons) while the WM contains the nerve fibers (axons of nerve cells) surrounded by a protective myelin sheath. Myelin, which gives the color white, acts as an insulator that facilitates the transmission of signals transmitted by nerve fibers. This is the neuron part that is attacked and damaged by MS disease (see Figure 1.1).

³www.wikipedia.org/wiki/Syst%C3%A8me_nerveux_central

⁴www.operativeneurosurgery.com/doku.php?id=white_matter

⁵www.wikipedia.org/wiki/Basal_ganglia

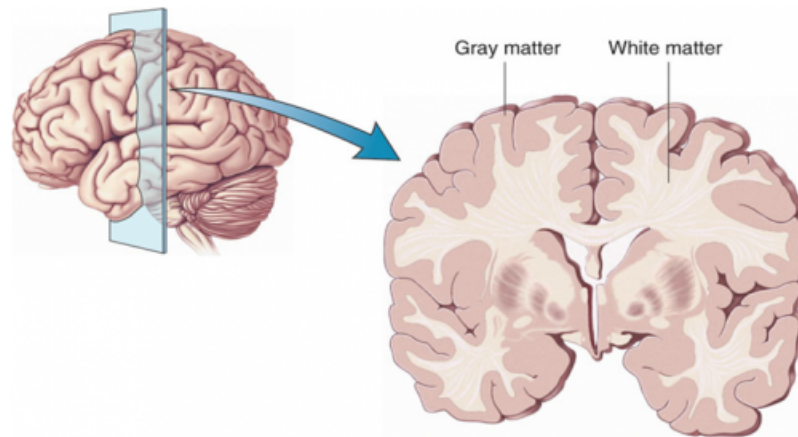


Figure 1.4: Gray Matter and White Matter. Source: ⁴.

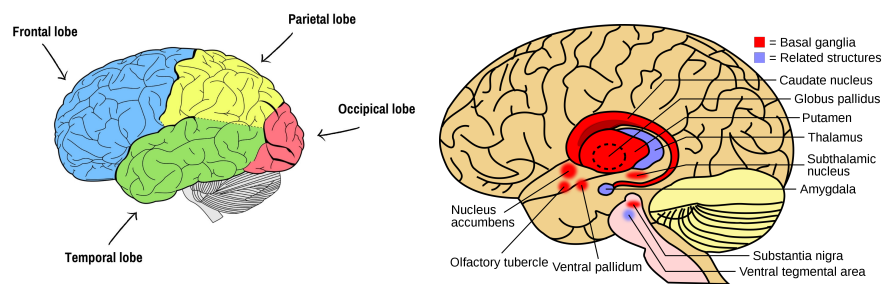


Figure 1.5: On the left, (a) the cortex lobes of the cerebrum. On the right, (b) the inner brain. Source: ⁵.

1.1.1.1 The Cerebrum

The cerebrum is the largest structure of the brain. It is part of the forebrain (additionally to the inner brain 1.1.1.2). Its prominent outer part, the cerebral cortex, not only processes sensory and motor information, but is also the origin of consciousness. Cortical tissue is composed primarily of neuronal cell bodies (GM), and its folds and fissures (known as gyri and sulci) give the forebrain its characteristic “crumpled” surface. The cerebral cortex has a right and left hemisphere. Each hemisphere can be divided into four lobes: the frontal lobe, the temporal lobe, the parietal lobe, and the occipital lobe, as shown in Figure 1.5 (a). These lobes correspond to functional segments specializing in various areas such as thinking, memory, planning, decision-making, speech, and sensory perception.

1.1.1.2 The Inner Brain

The inner brain is a set of structures buried deep under the cortex composed of the thalamus, the hypothalamus, the hippocampus, and the basal ganglia (see Figure 1.5 (b)). The thalamus determines which signals require conscious awareness, and which need to be available for learning or memory. The hypothalamus helps to take over the sensory impulses of smell, taste, and sight. The hypothalamus is also the center of visceral control which regulates the endocrine system and internal functions. The hippocampus is a memory indexer used for long-term storage and retrieval of memories. The basal ganglia are clusters of nerve cells surrounding the thalamus composed of the caudate nucleus, putamen, nucleus accumbens, globus pallidus, ventral pallidum, substantia nigra, and subthalamic nucleus. They are responsible for initiating and integrating movements.

1.1.1.3 The Cerebellum

The cerebellum is the second largest part of the brain (see Figure 1.6). It lies below the posterior (occipital) lobes of the cerebrum. Like the forebrain, the cerebellum has a right hemisphere and a left hemisphere. An intermediate region, the vermis, connects them. The primary function of the cerebellum is to maintain posture and balance, but it also participates in cognition.

1.1.1.4 The Brainstem

The brainstem connects the spinal cord to the higher centers of thought, located in the brain, as shown in Figure 1.6. It is made up of three structures: the medulla oblongata (rachidian bulb), the pons, and the midbrain. In addition to

⁶www.acikders.ankara.edu.tr/pluginfile.php/104007/mod_resource/content/1/week%203.pdf

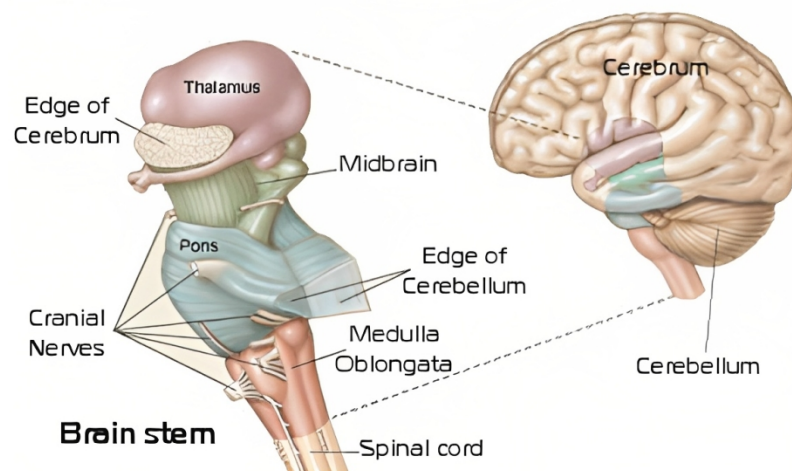


Figure 1.6: The cerebellum and the brainstem. Source: ⁶.

relaying motor and sensory signals, brainstem structures also direct involuntary functions. The medulla oblongata is in charge of breathing, digestion, circulation, and reflexes. The pons helps to control the breathing rhythm. The midbrain contributes to motor control, vision, and hearing, as well as reflexes related to sight and hearing.

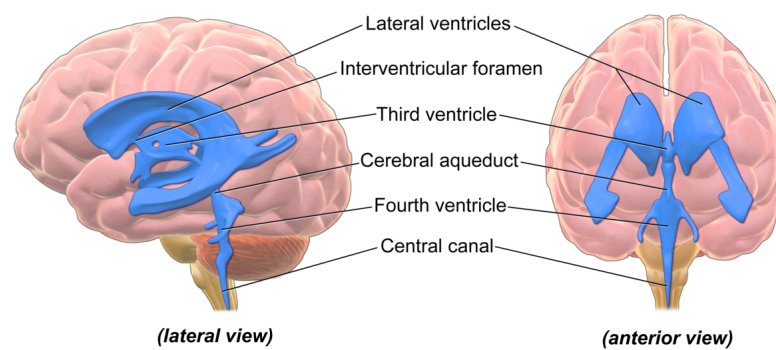


Figure 1.7: The brain ventricles. Source: [Medical, 2014]

1.1.1.5 The Brain Ventricles

The ventricles of the brain are cavities that produce and store cerebrospinal fluid (CSF) (see Figure 1.7). This fluid surrounds the brain and spinal cord, cushioning them and protecting them from trauma. It is also responsible for removing waste

and delivering nutrients to the brain. Of the four cavities comprising the ventricular system, the first and second are lateral ventricles. These C-shaped cavities are arranged symmetrically in each hemisphere. The third ventricle is a narrow funnel-shaped structure located between the right and left thalamus and above the brainstem. The fourth ventricle is a diamond-shaped structure that runs along the brainstem.

1.1.2 The Radiological Context of MS

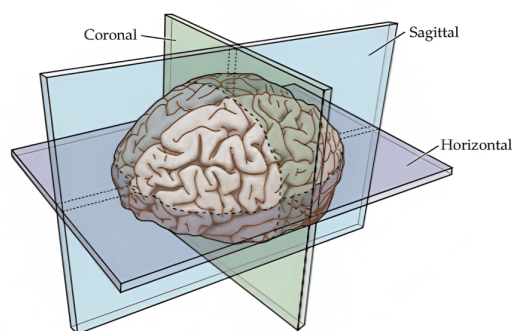


Figure 1.8: The coronal, sagittal, and axial planes of the MRI image of the brain. Source: ⁷

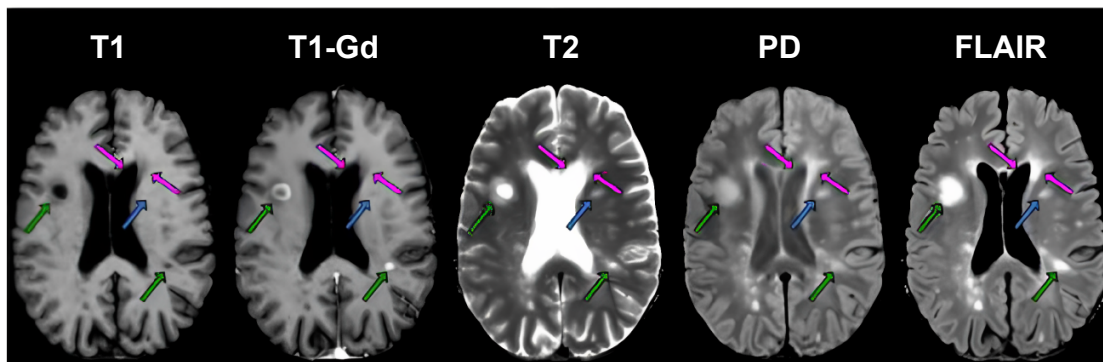


Figure 1.9: MS lesion appearance on different MRI modalities. Arrows highlight the advantages and disadvantages of each MRI modality in showing MS lesions. Modified from: [Gedamu, 2011].

⁷www.chegg.com/flashcards/biole201-3-53cd3f69-6300-4a12-9f88-acacc00e7a90/deck

Magnetic Resonance (MR) Imaging (MRI) is a non-invasive medical imaging technique that creates three-dimensional *in-vivo* images. The resulting image can be visualized alongside three planes: the coronal, sagittal, and axial (see Figure 1.8). MRI is based on applying a strong and stable magnetic field (B_0) to align protons' spin moment. Then, by applying weaker oscillating magnetic fields, the protons alignments are slightly modified and stabilized again which produces a measurable electromagnetic signal. Since the molecules composing different organic matters (*e.g.*, water, bone, muscle, fat, etc.) have a different response to the magnetic fields, MRI provides images with a high contrast of soft tissues and structure fidelity.

MRI sequences (modalities) provide different image contrast, optimized by tuning MRI parameters (echo time and repetition time), to better visualize specific tissues or abnormalities (*e.g.*, MS lesions).

MRI of the CNS is the most effective imaging technique to diagnose, assess, and follow up the progression of MS. It allows us to observe the neuroinflammation (*i.e.*, lesions) and neurodegeneration (*i.e.*, atrophy) caused by MS. As shown in Figure 1.9, several modalities can be used: T1-weighted (T1), T2-weighted (T2), Proton Density-weighted (PD), Fluid-Attenuated Inversion Recovery (FLAIR), and Gadolinium-enhanced T1 (T1-Gd). Other advanced modalities can also be helpful in MS monitoring such as Double Inversion Recovery (DIR), Diffusion Tensor Imaging (DTI), or Magnetization Transfer Ratio (MTR).

1.1.2.1 MS Lesions

MS lesions are one of the most important pathological hallmarks of MS. An MS lesion is defined on MRI as an area of focal hyperintensity (high-intensity spot) on the T2 and its variants (FLAIR or similar), or PD-weighted sequence. MS lesions have ovoid shapes and their size ranges from a few millimeters to more than 2 centimeters in diameter. To fulfill the diagnostic criteria and exclude artifacts, a lesion should be at least 3 millimeters in its long axis, and it should be visible on at least two consecutive slices [Filippi et al., 2019b].

1.1.2.2 Lesion Types by MRI Sequence

MS lesions can be divided into several categories depending on their appearance in different MRI modalities, associated with their inflammatory activity [Kaur et al., 2020]. Using different MRI sequences helps to better estimate the time when a lesion appears. The age of lesions and their state of inflammation are helpful for MS diagnosis and prognostic.

First, T2 lesions appear as hyperintense compared to WM in T2, PD, and FLAIR. This type of lesions is also known as White Matter Hyperintensities

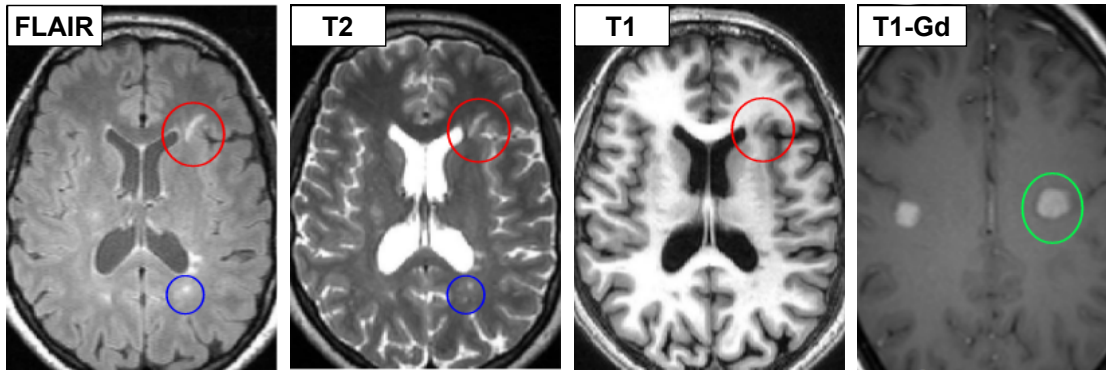


Figure 1.10: black holes, T2 lesions, T1-Gd lesions are surrounded by red, blue, and green circles respectively. Modified from: [García-Lorenzo, 2010].

(WMH). WMH can indicate other pathologies than MS such as vascular dementia, ischemia, and micro-hemorrhages, and they can also appear with aging (*i.e.*, frequent in the elderly population). Second, Gadolinium-enhanced lesions manifest with an increase in the intensity of T1-Gd after gadolinium injection compared with T1 pre-injection, they are also associated with hyperintense T2, PD, and FLAIR. Gadolinium-enhanced lesions indicate active inflammations or the age of MS lesions. If a lesion lights up, it means that active inflammation has occurred usually within the last two to three months. Third, black holes are long-time presence lesions associated with hypointensities in T1 that do not enhance with contrast product. These lesions represent more severe tissue damage and axonal loss. Figure 1.10 shows black hole, T2, and T1-Gd enhanced lesions.

The work of [Narayana et al., 2020] suggests that the most important modality for MS lesion (*i.e.*, neuroinflammation activity) visualization and delineation is FLAIR. As for contrast agents, the best practice is to limit their usage to necessary cases, as recommended by the French Observatory of MS [Brisset et al., 2020]. Indeed, excessive usage of gadolinium (Gd) can lead to renal failure and Gd deposit accumulation in the brain. For other cases and follow-ups, it is recommended to use 3D FLAIR, considering it to be the most relevant sequence. Additionally, T1 is important for MS since it is helpful to visualize black hole lesions but also to monitor the neurodegeneration activity caused by MS (see 1.1.2.4).

1.1.2.3 Classification of Lesion by Localization

The location of MS lesions is also an important aspect of both the diagnosis and the follow-up of the disease. Several MRI lesions are defined according to their location in the CNS anatomy.

Juxtacortical lesions are T2-hyperintense cerebral WM lesions juxtaposed (ad-

jacent) to the cortex, and not separated from it by WM. Infratentorial lesions are T2-hyperintense lesions in the brainstem or cerebellum. Periventricular lesions are WM hyperintense cerebral lesions located in the surrounding region of the lateral ventricles without WM in between, including lesions in the corpus callosum (bundle of axons interconnecting the two cerebral hemispheres) but excluding lesions in deep GM structures. Cortical MRI lesions are within the cerebral cortex. Typically, special MRI techniques such DIR, or high-resolution MRI (7T MRI) are required to visualize these lesions. Care is needed to distinguish potential cortical lesions from neuroimaging artifacts. Deep GM lesions are located in the inner brain (deep GM nuclei), but they are most common in the thalamus, where lesions can be located along the surface of the ventricle or around blood vessels. These lesions tend to have less of an inflammatory pattern than WM lesions but more so than cortical lesions. Spinal cord MRI lesions are located in the cervical, thoracic, or lumbar spinal cord, usually visible on PD sequence.

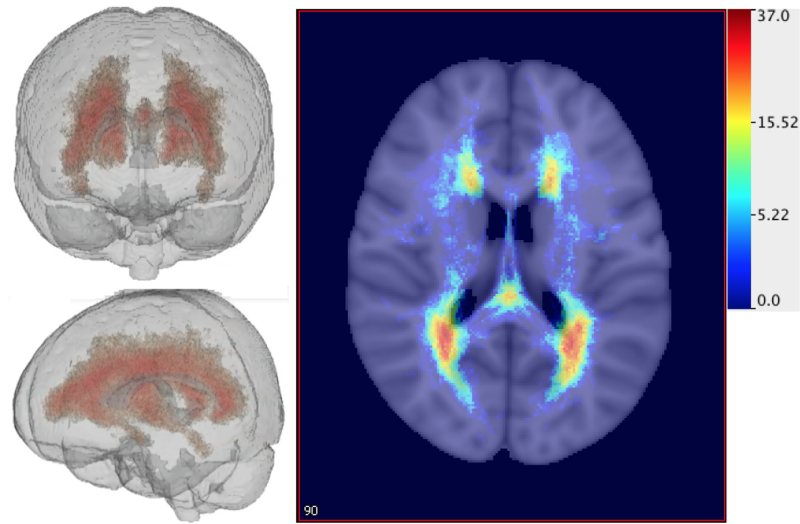


Figure 1.11: The histogram of brain lesions for 98 patients based on a rigid registration of the images to a template brain indicating the number of patients out of 98 having lesions at each voxel. Source: [Eloyan et al., 2014].

While lesions may occur in any CNS region, they tend to affect specific WM regions, periventricular and juxtacortical WM, corpus callosum, infratentorial areas, and the spinal cord. Figure 1.11 shows the statistical distribution of MS lesions in a cohort of 98 patients [Eloyan et al., 2014].

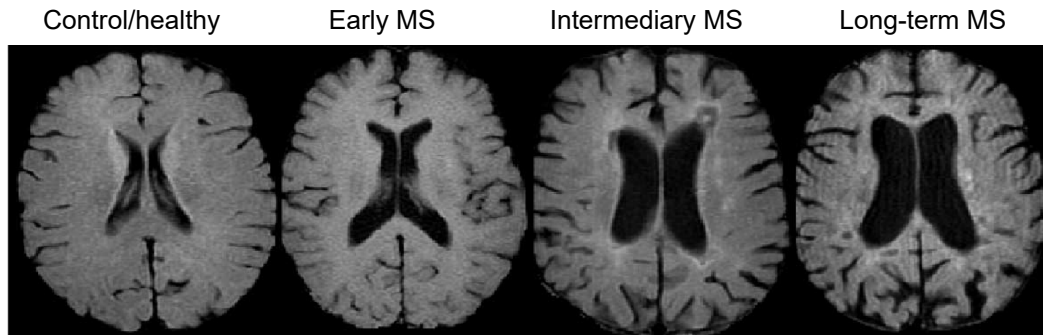


Figure 1.12: Brain atrophy in different stages of MS. Atrophy can be seen by the enlargement of lateral ventricles (the inner side of the brain) and by the reduction of the cortical thickness (the outer side of the brain). Modified from: [Bermel and Bakshi, 2006]

1.1.2.4 Brain Atrophy

Brain atrophy is the gradual loss of brain volume and a clinically relevant component of disease progression in MS. The brain volume of patients suffering from MS decreases far off the limits of normal aging, with approximately 0.5–1.35% per year [Andravizou et al., 2019]. The atrophy of brain tissue is widespread, affecting the entire brain and all its parts, including the lobes, WM, brainstem, and cerebellum [Bakshi et al., 2001] (see Figure 1.12).

Brain atrophy begins early in the disease course and can accelerate with disease progression [Andravizou et al., 2019]. Although severe atrophy is the long-term effect of inflammation, many studies showed that GM atrophy is observable from the earliest stages of MS, and is only moderately related to lesion accumulation [Dalton et al., 2004]. Early atrophy can be located in specific cortical regions but also in the inner brain, affecting structures such as the thalamus or the caudate [Zivadinov et al., 2008].

The increasing amount of data linking brain atrophy to clinical impairments suggests that irreversible tissue destruction is an important determinant of disease progression to a greater extent than can be explained by conventional lesion assessments [Bermel and Bakshi, 2006]. Besides, some studies suggested that GM atrophy is more sensitive for identifying progressive neurodegeneration compared to WM or whole-brain atrophy [Zivadinov et al., 2008].

1.1.3 The Clinical Context of MS

As a consequence of MS inflammation activity (lesions), relapse is a period during which people with MS experience new symptoms. Relapse is defined by one or

Table 1.1: Symptoms and Disabilities associated with MS based on lesion location⁸. These associations are far from absolute, and it is possible for an individual to have MS lesions without any obvious accompanying symptoms, or to have symptoms even though a lesion cannot be clearly visualized on MRI scans.

Lesion location	Possible associated symptoms
Outermost section of the brain (cortical or juxtacortical regions, including the cortex and cerebrum)	Cognitive and memory impairment Depression Fatigue Weakness or numbness
Central section of the brain (periventricular region)	Impaired cognition and executive function Numbness or other abnormal sensations Problems related to movement Fatigue
Lower back of the brain (infratentorial region, including the cerebellum and brainstem)	Double vision Swallowing difficulty Weakness or unusual sensations in the face Impaired balance and coordination
Spinal cord	Muscle weakness or stiffness Trouble with coordination and balance Pain, tingling Sexual dysfunction Bladder and bowel problems
Optic nerves	Vision problems (<i>i.e.</i> , blurry vision) Painful eye movements

more neurological signs occurring for more than 24 hours, in the absence of fever or infection. These signs can be ocular, sensory, motor, and other disorders. The first experienced relapse by a patient is called Clinically Isolated Syndrome (CIS). With no further evidence, CIS is not yet considered (diagnosed) as MS.

Depending on their localization, lesions can have different effects on patients. Table 1.1 summarizes some of the most important associations between lesion locations and possible symptoms.

1.1.3.1 MS Disease Course

MS patients may be grouped into three categories, also called MS phenotypes, depending on the course and progression of the disability with time [Kaur et al., 2020, Goldenberg, 2012]. MS disease courses are Relapsing-Remitting MS (RRMS), Secondary Progressive MS (SPMS), and Primary Progressive MS (PPMS) (see Figure

⁸www.multiplesclerosisnewstoday.com/ms-lesions/

1.13)

RRMS is the most common form of MS and affects around 85% of patients, it is characterized by relapses of the symptoms followed by an improvement or cease. SPMS is the evolution of RRMS. This type is characterized by the development of the disease with or without periods of remission and with or without inflammatory activity. PPMS manifests as a slow progression of the symptoms without picks nor remission periods and resists treatment.

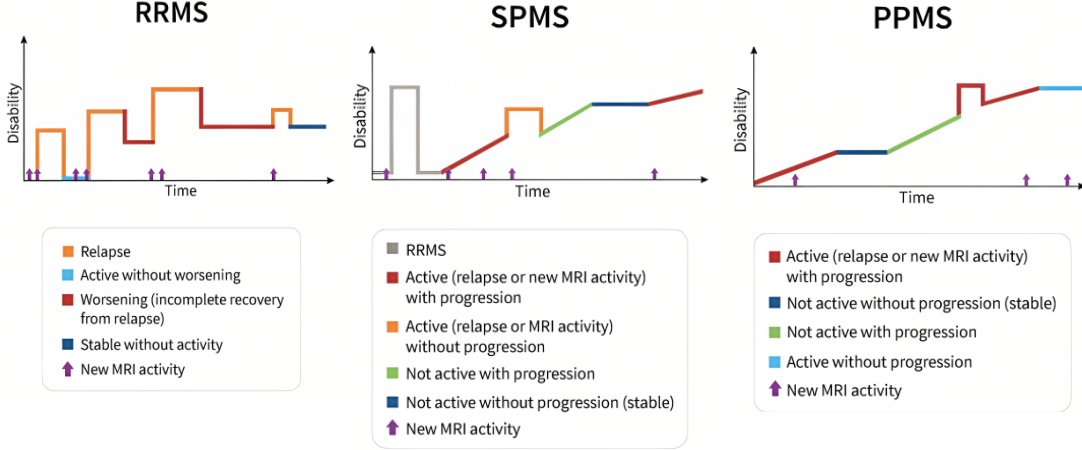


Figure 1.13: Disability progression for MS subtypes [Lublin et al., 2014]

1.1.3.2 The Diagnosis of Multiple Sclerosis

The diagnosis of MS is made when the patient fulfills specific and precise criteria. The latest iteration and refinement is the 2017 McDonald criteria for MS (refer to [Thompson et al., 2018] for further details). After the first clinically isolated symptoms are observed/detected, the clinician needs radiologic evidence to establish the diagnosis. This radiologic inspection is based on two major characteristics of MS – Dissemination In Space (DIS) and Dissemination In Time (DIT). DIS is the development of lesions in distinct anatomical locations within the CNS whereas DIT is the development or appearance of new lesions over time.

1.2 Image Processing Tasks for MS

The diagnosis and the follow-up of MS are heavily dependent on extracting information from MRI. First, MS clinicians need to closely analyze the multi-sequence MRI to spot WMH, delineate the lesions based on their expert knowledge, and discard any false positive (*i.e.*, image artifact such as noise). As mentioned in

section 1.1, MRI lesions are an essential biomarker for MS. Second, the detection of new-appearing lesions in longitudinal MRI is required for neurologists to understand the evolution of the disease and to assess and adapt treatment. Third, the recognition of brain structures and measurement of their volumes is essential to understand how MS patients are impacted by their disease. As highlighted in sections 1.1.2.3 and 1.1.2.4, both the location of lesions with regard to brain structures and the brain atrophy are linked to MS symptoms and disabilities. Finally, all the previously mentioned information extracted from multi-modal MRI needs to be jointly analyzed in addition to demographic data, clinical tests, and comorbidities (*i.e.*, other conditions that may interact with MS). This overall assessment is needed for a personalized efficient treatment and a better prediction of the disease progression.

The extraction of information from MRI could be performed manually by radiologists or experts in the field using visualization and annotation software (such as ITKsnap [Yushkevich et al., 2016]). Although manual extraction can be accurate when performed carefully, it suffers from multiple limitations. These tasks are tedious, time-consuming, expensive, require domain expertise, prone to inter-rater and intra-rater variability. Thus, automatizing these tasks was eventually considered.

1.2.1 Preprocessing

MRI task automation can be very complex due to the diversity and inhomogeneity of the obtained images. To reduce automatic task complexity and make MRI analysis easier in general, image preprocessing steps are applied to improve the image quality and normalize the overall representation. Indeed, MR images are usually affected by different types of artifacts that degrade image quality during the acquisition process due to hardware, calibration, subject motion, and other magnetic characteristics.

The most important preprocessing steps in neuroimaging are denoising, inhomogeneity correction, registration, and intensity standardization (see Figure 1.14). In the following, these preprocessing steps will be briefly addressed. For further details, refer to [Manjón, 2017].

1.2.1.1 Denoising

MR images are inherently corrupted by random noise from the image acquisition process which introduces uncertainties in the measurement of quantitative biomarkers. Although denoising could be achieved with a classic high-frequency filter, the critical aspect of medical imaging makes it risky to remove at the same time high-frequency components necessary for the analysis. Current stable state-

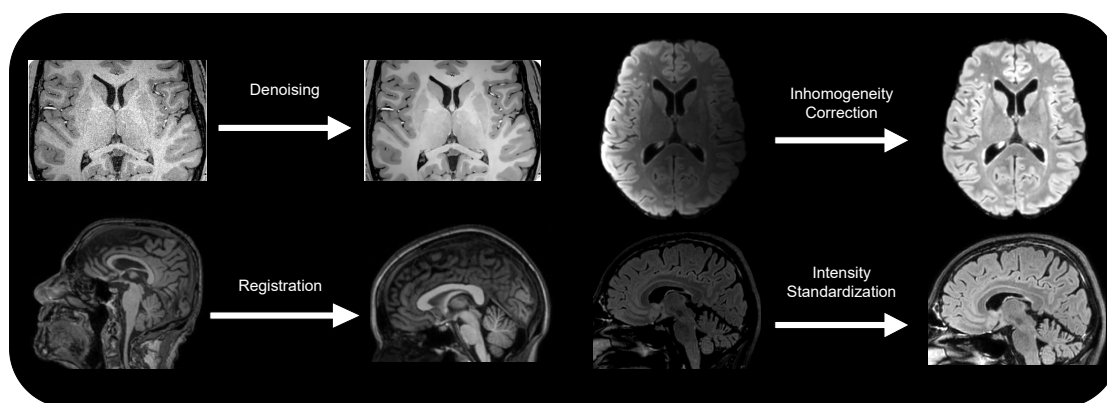


Figure 1.14: Most important preprocessing steps in neuroimaging

of-the-art denoising methods are based on patch-wise image processing approaches exploiting sparseness and self-similarity properties of the medical images such as [Manjón et al., 2010, Coupé et al., 2008] or based on deep learning [Manjón and Coupé, 2018, Ran et al., 2019].

1.2.1.2 Inhomogeneity Correction

MR images signal intensity inhomogeneity is mainly produced by imperfections in the radio-frequency coils and object-dependent interactions. Such an artifact is perceived as a low-frequency variation of the signal intensity across the image. Indeed, a given tissue should be represented by similar voxel intensities throughout the data but intensity inhomogeneity, as the name suggests, makes similar tissues have different intensities. The most used methods are N3 [Sled et al., 1998], N4 [Tustison et al., 2010], and SPM8/SPM12 [Ashburner, 2002] (SPM is limited to brain imaging).

1.2.1.3 Registration

Image registration is the process of finding the optimal geometric transformation to spatially align two different images and represent them in a similar space. Registration is necessary to compare different subjects having similar properties (same anatomical region) or analyze different images of the same subject (multimodal or longitudinal images). Registration is also used to represent images in a common and standard space, such as the Montreal Neurological Institute (MNI) space. The registration process first estimates the transformation parameters needed to map the different images, then, applies them to the moving image/s. Depending on the complexity of the transformation, we can have linear or nonlinear registrations.

There are several available registration methods such as SyN [Avants et al., 2008], ITK [Avants et al., 2014], and SPM's DARTEL Toolbox.

1.2.1.4 Intensity Standardization

Due to scanner configuration, MR images acquired with a similar protocol (for instance T1 or FLAIR) are not always encoded with the same intensity levels. Moreover, even within the same scanner and setting, there could be variability in the intensity patterns of the acquired images when comparing different sessions. Several methods deal with intensity standardization such as Kernel Density Estimation (KDE), Z-score using the brain mask, or histogram matching [Nyúl et al., 2000].

1.2.2 Automated Tasks

To address the difficulties that come with the manual processing and analysis of MRI related to MS, many fully automatic methods have been proposed. These methods tackled the automation of MS-related tasks such as the detection and segmentation of MS lesions [Danelakis et al., 2018, García-Lorenzo et al., 2013], structure and tissue segmentation [González-Villà et al., 2016, Akkus et al., 2017], and MS course classification [Ion-Mărgineanu et al., 2017a, Kocavar et al., 2016]. Since we cover several tasks related to MS analysis automation in this section, we will focus on the common underlying mechanisms shared by these methods. In chapters 2, 3, and 4 we will detail the specificities of works related to our methods. The algorithms used for MS task automation can be grouped into the broad categories of supervised and unsupervised methods.

1.2.2.1 Supervised Methods

Supervised models search for a function that maps an input to an output based on a labeled dataset containing input-output pairs. The inferred function obtained by analyzing the training data can be used for mapping new input samples to their respective output. Due to the exponential increase in computational power and the explosion of BigData, most supervised methods are based on Machine Learning (ML). The principle of ML is to give computers the ability to "learn" from data by improving their performance to solve tasks without being explicitly programmed.

Classical Machine Learning

ML-based methods generally consist of a feature extraction step, followed by a step to correlate the extracted feature to the objective output (*e.g.*, classification

or regression). In the first step, the aim is to simplify the learning by using input features that are more relevant to the task, instead of raw MRI voxel intensities. These set of features (a.k.a hand-crafted features) are manually chosen by experts in the domain.

In the case of features derived from MRI or medical imaging that capture tissue, lesion characteristics, and other complex patterns, they are referred to as radiomics. Radiomics include several categories of features such as **first-order features** (*i.e.*, gray-level intensity mean, maximum, minimum, standard deviation, percentiles, and skewness), **texture features** (*i.e.*, same statistical descriptors but for the absolute gradient [Benoit-Cattin, 2006]). Other features derived from spatial relation and spatial distribution about pairs, groups, or neighboring pixels/voxels such as GLRLM [Galloway, 1975], GLCM [Haralick et al., 1973], GLSZM/GLDZM [Thibault et al., 2013], and NGTDM [Amadasun and King, 1989]), **shape-based feature** (*i.e.*, geometric descriptor [Zhang et al., 2019a, Al-Zubi et al., 2002] including diameters, surface, compactness, sphericity), and **transform-based features** (*i.e.*, Fourier [Yonar et al., 2018], Gabor [Bodis-Wollner and Brannan, 1997], and Haar wavelet transforms [Laine and Fan, 1993]).

Besides radiomics which do not contain apriori nor expert-infused knowledge, several high-level features can be extracted from the different imaging modalities. High-level features are task-specific and can be extracted either manually or using other automatic methods and tools such as FreeSurfer [Fischl, 2012]. High-level features related to MS may include MS lesion features (*e.g.*, lesion load [Zijdenbos et al., 2002], lesion count, and location), brain structure and tissues statistics (*i.e.*, volume and asymmetry), brain curvature, cortical thickness, and fractional anisotropy [Pfefferbaum et al., 2003, Ciccarelli et al., 2001].

Both radiomics and high-level features can be used alone or combined with complementary information such as clinical (*e.g.*, treatment, comorbidities, disease duration, and other disease-related statistics) and demographic (*e.g.*, age, sex, and ethnicity) information, for tasks automation and analysis [Mayerhoefer et al., 2020, Yip and Aerts, 2016, Ion-Mărgineanu et al., 2017b].

In the second step, ML models are trained to learn a correlation between the aforementioned features and the output objective. Many ML algorithms have been used for MS automatizing tasks, such as the bayesian approach [Johnston et al., 1996], logistic regression [Barkhof et al., 1997], artificial neural networks [Zijdenbos et al., 1994], K-NN [Vinitski et al., 1999], and Random forests [Boucekine et al., 2013].

These algorithms are either used for **segmentation** (*i.e.*, MS lesion segmentation methods such as BIANCA [Griffanti et al., 2016], OASIS [Sweeney et al., 2013], MIMoSA [Valcarcel et al., 2018] and LST toolbox for SPM [Schmidt and Wink, 2017]), **classification** (*i.e.*, MS disease course classification [Zhao et al., 2017,

Pinto et al., 2020]), or **regression** (*i.e.*, the estimation and prediction of MS disability score [Roca et al., 2020, Pontillo et al., 2021]).

Deep Learning

Deep Learning (DL) is a specific type of ML based on artificial neural networks (a.k.a deep neural networks) in which multiple layers of processing are used to extract progressively higher-level features from input data. Contrarily to classical ML, DL requires reduced to no feature engineering, since both feature extraction and correlation are learned simultaneously during training. DL methods have gradually overwhelmed classical ML approaches during the last years due to the advancement of graphical computing (*i.e.*, Graphical Processor Unit (GPU)) and the increasing amount of available training data. Deep neural networks feedforward (the transformation of the input to the output) is guided by networks weights or parameters. These parameters are learned by backpropagation which computes the gradient of the loss function (*e.g.*, mean squared error, binary cross entropy) with respect to the network weights based on a labeled dataset. This efficiency makes it feasible to use gradient methods for training multilayer networks such as gradient descent and its variants (*e.g.*, stochastic gradient descent [Bottou, 2010], ADAM [Kingma and Ba, 2014]).

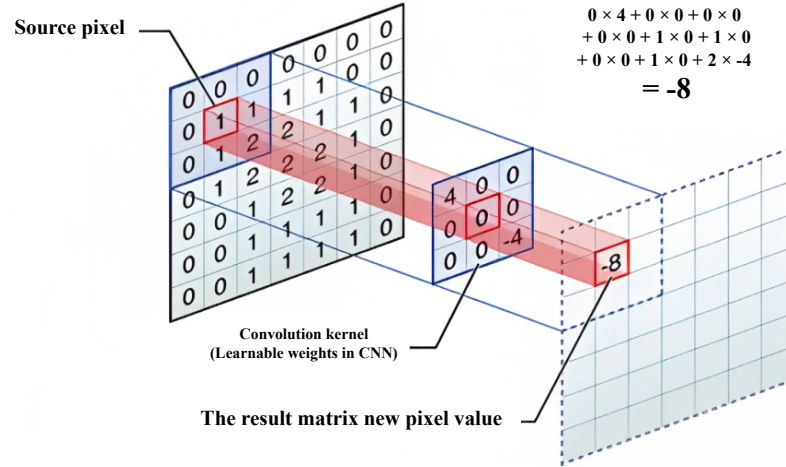


Figure 1.15: The 2D convolution operation used in CNN learns to extract priorities of the pixel neighborhood. Modified from: ⁹

DL methods have gained popularity for MS-related tasks, especially with the Convolutional Neural Network (CNN). CNN takes its name from convolution, a

⁹www.medium.com/@bdhuma/6-basic-things-to-know-about-convolution-daef5e1bc411

¹⁰www.stanford.edu/~shervine/l/fr/teaching/cs-230/

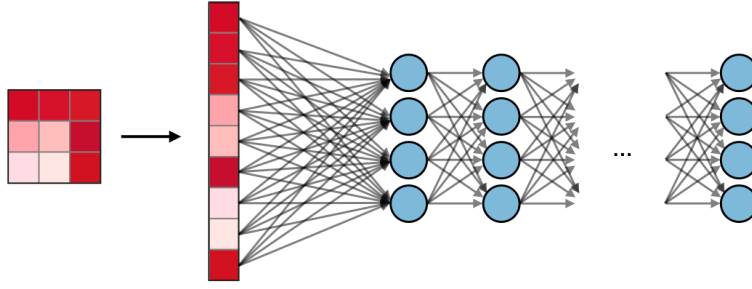


Figure 1.16: The fully connected layers learn the mapping of every element in a layer to each element of the next layer. Modified from: ¹⁰

mathematical linear operation between matrices that is able to represent pixel/voxel neighborhood proprieties (see Figure 1.15). CNNs have multiple layers including convolutional layers, activation layers (*i.e.*, mathematical operations that introduce non-linearity such as rectified linear units, hyperbolic tangent, sigmoid, softmax), pooling layers (*i.e.*, to reduce the complexity and aggregate features), and fully-connected layers. While both the convolutional and fully-connected layers have learnable parameters, the strength of CNN is in exploiting the spatial aspect of 2D or 3D images by learning the kernel mask coefficients of the convolution operation, as opposed to fully-connected layers which learn the linear coefficients for every element of a layer to output a single element of the subsequent layer (see Figure 1.16). The kernel size defines the limited range of convolution which also represents the limitation of CNN since they capture local features. Recent works propose solutions to learn global features such as non-local neural networks [Wang et al., 2018], and visual transformers [Dosovitskiy et al., 2020].

For MS lesion segmentation several patch-wise CNN have been proposed such as nicMSLesions [Valverde et al., 2017], DeepMedic [Kamnitsas et al., 2015, Kamnitsas et al., 2016], FLEXCONN [Roy et al., 2018], and MSWS [Ghafoorian et al., 2017].

Fully Convolutional Networks (FCN) are a special case of CNNs that do not contain fully-connected layers. These networks take a 2D/3D input and produce correspondingly-sized output with very efficient inference and learning [Long et al., 2015]. Indeed, FCN permits the design of pixel/voxel-based models able to produce the segmentation in a single pass which makes them significantly faster compared to their patch-wise counterparts. FCN gave rise to several pixel/voxel-based models for MS-related segmentation tasks such as 2D and 3D Unet [Ronneberger et al., 2015, Çiçek et al., 2016], nnUnet [Isensee et al., 2019], [Brosch et al., 2016], and 2D Tiramisu [Zhang et al., 2019b].

1.2.2.2 Unsupervised Methods

Unsupervised Learning methods analyze and find distinctive patterns in data without the need for human-labeled output. In this section, we will briefly discuss how unsupervised algorithms can be used in MS analysis and task automation.

First, early segmentation methods were implemented with explicitly programmed algorithms such as thresholding, intensity analysis, and morphology/topology analysis [Iheme and Unay, 2005, Boesen et al., 2004, Goldberg-Zimring et al., 1998]. Contrary to learning-based techniques, these algorithms specifically defined the task of segmentation based on image priorities (*e.g.*, intensity and object shape). Although these simple methods have limited performance, especially for complex tasks, they can be very helpful as a baseline or when building a labeled dataset. Indeed, on one hand, the use of complex learning-based methods is often motivated by a significant increase in performance compared to simpler baseline methods. On the other hand, when labeling data for a new task with no available learning-based automatic method, these simple algorithms can make labeling easier by providing the first prediction that a human rater can correct and refine manually.

Second, atlas-based methods are data-centric techniques that use healthy brain anatomy represented by the atlas to discriminate deviant tissues and anomalies such as MS lesions. The atlas of the brain is a map representing either statistical or topological information. The statistical atlas provides the prior probability of each voxel belonging to a particular tissue class whereas the topological ones encode a specific topology for each structure and group of structures. Both can be used for MS-related tasks since MS lesions have a statistical probability of distribution and are considered topological outliers, such as in TOPOLOGY-preserving Anatomical Segmentation (TOADS) for MS lesions [Shiee et al., 2010]. Atlas-based methods require a non-linear registration to fit the input data to the atlas which makes their processing time longer compared to their learning-based counterparts. However, their simple mechanism can make them more efficient for the detection of very rare anomalies.

Third, unsupervised methods can also be used to cluster and separate data. This can be advantageous when performing an automatic task (*i.e.*, the discrimination of MS lesions from healthy tissues [Pham and Prince, 1999, Desolneux et al., 2003]). Besides that, recent works [Eshaghi et al., 2021] used unsupervised clustering methods, to uncover data-driven MS disease subtypes with distinct temporal progression patterns and different trajectories of abnormalities stages (*e.g.*, atrophy, tissue damage, lesions). This suggests that the analysis of MRI-based information is able to better predict worsening and response to treatment compared to clinical phenotypes (*i.e.*, RRMS, SPMS, PPMS).

1.3 Thesis overview

In this thesis, we aim to propose a suite of tools for the automation of MS neuroimaging tasks. As shown previously in this chapter, MS is a complex disease having disseminations in space and time. Moreover, the link between the radiological and clinical manifestations of MS is still unclear. Therefore, its diagnosis, monitoring, or progression prediction requires development of specific methods to help clinicians in their tasks. In the following chapters, we will present the methods developed during this Ph.D. to address *i)* lesion segmentation (Chapter 2), *ii)* detection of new lesions (Chapter 3) and *iii)* disability estimation (Chapter 4). All these chapters correspond to published or submitted articles during the preparation of the thesis.

In Chapter 2, we present DeepLesionBrain, a novel MS lesion segmentation method that is robust to domain shift and well-performing on unseen datasets. This generalization property results from three main contributions. First, the method uses a large group of compact 3D CNNs spatially distributed over the brain with overlapping receptive fields between regions. By associating a distinct network with each region of the brain, the spatially distributed networks strategy simplifies the MS lesion segmentation from a single complex task on the whole brain to multiple simpler sub-tasks on each region. Moreover, the overlapping regions ensure consistent and stable consensus. Second, to extract more relevant features for MS segmentation that may lead to better generalization, the method is trained with the proposed Hierarchical Specialization Learning. The two-step training strategy consists of a single network pre-trained on all brain regions that are used in the subsequent step to initialize the weights of each spatially distributed network. Third, DeepLesionBrain is trained with a novel image quality data augmentation method, which mimics real-world data diversity by adding realistic alterations to the training images. These specific augmentations constrain task learning to be independent of source data acquisition resolution, data contrast, or data quality. Consequently, the proposed augmentation strategy enables domain shift robustness. The method generalization was validated in cross-dataset experiments on a couple of reference datasets, and our in-house datasets. During experiments, DeepLesionBrain showed higher segmentation accuracy, better segmentation consistency, and greater generalization performance compared to state-of-the-art methods.

In Chapter 3, we describe a DL-based pipeline addressing the challenging task of detecting and segmenting new MS lesions. Indeed, the lack of annotated longitudinal data for this task and the rarity of cases with new-appearing lesions are limiting factors for the training of robust and generalizing models. These data scarcity and class imbalance problems are tackled in three main contributions. First, transfer learning is proposed to exploit the larger and more diverse datasets

available for the task of single-point MS lesion segmentation. The datasets used to train MS lesion segmentation are more easily available and the positive class is more frequent (less impacted by class imbalance) compared to the task of new MS lesion segmentation. Therefore, exploiting knowledge from an easier and similar task with a richer training set improved considerably the task of new MS lesion segmentation. Second, the pipeline includes a novel data synthesis strategy to generate realistic longitudinal time-points with new lesions using single time-point scans. The strategy combine the use of a lesion generator and lesion eraser models (both trained separately and before the described pipeline) to generate “on the fly” synthetic 3D patches that represent longitudinal scans of the same patient with evolution in their lesion mask. In this way, the model is trained on large synthetic annotated datasets. Third, an improved version of the Image Quality Data Augmentation (presented in Chapter 2) is proposed to simulate data diversity in MRI. The improved version introduces more alterations both in the spatial and frequency space (k-space). Thus, the augmentation method helps to better oversample the scarce samples with new lesions and makes the trained model more robust to image diversity. The ablation study showed that each contribution lead to an enhancement of the segmentation accuracy. Using the proposed pipeline, we obtained the best score for the segmentation and the detection of new MS lesions in the MSSEG2 MICCAI challenge.

In Chapter 4, we propose a novel method for the estimation of EDSS from MRI-based and clinicodemographic information with DL. To achieve this, we first propose to extract relevant neurodegenerative and neuroinflammatory biomarkers. Next, we propose CNN architecture that effectively combines image-based and tabular information. Moreover, our model is trained with novel multi-phase learning and data augmentation to mitigate the data imbalance effect. During validation, our DL method surpassed conventional state-of-the-art methods. To better understand which brain regions contribute the most to the prediction of our EDSS estimation model, we propose to use model interpretability visualizations adapted for our task. The obtained visualizations show consistency with works associating MS functional impairments at different stages of the disease with brain structures responsible for these functions.

In Chapter 5, we propose complete pipelines for the processing of multimodal neuroimaging data and the extraction of MS-relevant information based on the works proposed in chapters 2 to 4. First, we propose to use containerization and host our pipelines on a web-based platform to allow the MS community to use them easily without the need of dealing with installation and hardware requirements. Second, we propose easy-to-use pipelines that do not require technical expertise to execute and retrieve their results. Finally, we propose to generate reports offering compact, organized, and easy-to-read information.

Chapter 2

MS Lesions Segmentation

This chapter corresponds to the following publication:

Reda Abdellah KAMRAOUI, Vinh-Thong TA, Thomas TOURDIAS, Boris Mansencal, José V. Manjon, and Pierrick Coupé. DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis*, 2022, vol. 76, p. 102312 [2].

Recently, segmentation methods based on Convolutional Neural Networks (CNN) showed promising performance in automatic Multiple Sclerosis (MS) lesions segmentation. These techniques have even outperformed human experts in controlled evaluation conditions such as Longitudinal MS Lesion Segmentation Challenge (ISBI Challenge). However, state-of-the-art approaches trained to perform well on highly-controlled datasets fail to generalize on clinical data from unseen datasets. Instead of proposing another improvement of the segmentation accuracy, we propose a novel method robust to domain shift and performing well on unseen datasets, called DeepLesionBrain (DLB). This generalization property results from three main contributions. First, DLB is based on a large group of compact 3D CNNs. This spatially distributed strategy aims to produce a robust prediction despite the risk of generalization failure of some individual networks. Second, we propose a hierarchical specialization learning (HSL) by pre-training a generic network over the whole brain, before using its weights as initialization to locally specialized networks. By this end, DLB learns both generic features extracted at global image level and specific features extracted at local image level. Finally, DLB includes a new image quality data augmentation to reduce dependency to training data specificity (*e.g.*, acquisition protocol). DLB generalization was validated in cross-dataset experiments on MSSEG’16, ISBI challenge, and in-house datasets. During experiments, DLB showed higher segmentation accuracy, better segmentation consistency and greater generalization performance compared to state-of-the-art methods. Therefore, DLB offers a robust framework well-suited for clinical practice.

2.1 Introduction

2.1.1 Problem Description

In recent years, the medical imaging community has witnessed a rapid increase in image processing methods based on DL. These novel techniques came with remarkable performance in many tasks including MS lesion segmentation. Some automated algorithms have even reached human-level performance in controlled evaluation conditions (see [Carass et al., 2017]). Unlike over-controlled conditions where most DL approaches have been validated, real-world data exhibit high diversity. Consequently, clinical use of MS lesion segmentation based on DL is still limited mainly because of their poor generalization on new data coming from medical sites that have not been covered during training (unseen domains). This lack of generalization of DL methods can result from several factors such as the selected solution during optimization, the diversity of the training dataset or the genericity of the extracted features.

DL is based on the assumption that training and test data are independent but come from the same distribution. This assumption is usually not respected in medical imaging data especially for MRI where acquisition protocols, MRI scanner, patient populations, and software processing may vary depending on the clinical center or the cohort. As a result of these differences of data distribution (covariate shift), a decrease in performance is observed between the training data (source domain) and the test data derived from different distributions (target domain). This is known as the domain shift.

An intuitive method to reduce this problem is to train on a wider and more heterogeneous dataset (as shown by [Mårtensson et al., 2020]). However, this requires a large dataset annotated by experts which are rarely available and tedious to produce. Some deal with this phenomenon by applying extensive data augmentation (such as [Zhang et al., 2020b]). Others use few available labeled images from the target domain to reduce the covariate shift, such as few-shot and one-shot learning strategies (see [Snell et al., 2017, Valverde et al., 2019]).

Besides, DL requires the tuning of a large number of parameters relative to the number of training data samples. Thus, it usually ends up converging to one of the many possible local minima as opposed to the theoretical best parameter configuration which leads to the global minimum. Consequently, the generalization ability of the model depends on the selected solution. The selection of the best generalizing local minimum is still an open question. On one hand, some works have proposed to select it using the local characteristics (*e.g.*, flatness) of the loss function (see [Keskar et al., 2016, Wu et al., 2017]). On the other hand, an alternative strategy consists in combining several local minima to improve the generalization capability of the method. This can be done by averaging several local minima of

one model (*e.g.*, snapshot ensemble [Huang et al., 2017]) or by combining outputs of different models trained independently (*e.g.*, classical ensemble [Zhang et al., 2019b] and spatially distributed networks [Coupé et al., 2020, Huo et al., 2019]).

Unlike classical methods that use hand-crafted features, Convolutional Neural Networks (CNNs) automatically extract the most suitable set of features for a particular task. Although this strategy is very efficient to extract relevant features for a particular source domain, this set of features may not generalize well for the target domain. Some works proposed to learn invariant features that coexist across different source domains [Motiian et al., 2017, Muandet et al., 2013, Yang and Gao, 2013]. They tried to apply a regularization to learn an abstract representation of the specific computer vision task (*i.e.*, just like humans understand high-level concepts). Indeed, the extraction of generalizing features lies between the freedom of the optimization process to find the optimal combination from data and the constraints used for minimizing domain bias.

The successful deployment of DL based methods for MS lesion segmentation requires generalization capabilities that can guarantee high performance for unseen domains. First, such methods should ensure the convergence of the DL model to generalizing minima. Second, the training process should anticipate the reduction of the covariate shift. Moreover, the method should be enforced to learn MS lesion generalizing features from the source domain, to effectively delineate lesions despite the target domain distribution. Finally, this solution should not require additional annotation in case of processing unseen domains.

2.1.2 Related Works

Recently, many works have been proposed for MS lesion segmentation using CNNs. First, [Brosch et al., 2016] proposed a deep 3D encoder-decoder network, with joint training of the encoder and the decoder. The authors used shortcut connections between the two interconnected pathways for integrating high and low-level features. This pioneering work demonstrated the high potential of deep learning for MS lesion segmentation.

[Valverde et al., 2017] proposed a cascade of two patch-wise 3D CNNs, composed of a first sensitive network to reveal possible lesion candidates followed by a second network to reduce misclassified voxels. This cascade allows refined segmentation but it uses a small receptive field that prevents capturing the global context. Later, the authors [Valverde et al., 2019] improved their method by proposing a one-shot domain adaptation model which uses transfer learning and partial fine-tuning. However, this domain adaptation needs a labeled example from the new domain. Moreover, such strategies lead to different versions of the method after each adaptation, this results in discrepancies in the segmentation.

[Hashemi et al., 2018] considered the problem of data imbalance (*i.e.*, the under-

sampling of the lesion class) by using an asymmetric similarity loss function based on Tversky index to train a 3D CNN that performed better than Dice or cross-entropy measures. This result suggests that further work should be done on choosing an adequate loss function. Although the proposed loss can be tuned for the optimal trade-off between precision and recall in a particular domain, the generalization to unseen domains has to be proven.

[Zhang et al., 2019b] used a fully convolutional densely connected network for MS lesion segmentation. They stacked adjacent 2D slices of different modalities with a channel-wise concatenation, before forwarding this stack through a 2D CNN. The final segmentation is based on a majority vote along different orientations. While this method showed competitive performance on a well-controlled challenge, the stacking using only the two directly adjacent slices gives a weak insight into the 3D nature of the data. Moreover, 2D features may not be considered as generalizing features when processing 3D volumes and can result in the limited generalization of the method.

[Aslani et al., 2019] proposed an end-to-end encoder-decoder 2D network with multiple downsampling branches, one for each input modality, and a decoder part where features from the different modalities are put together at multiple scales. This separation in encoder branches enables the model to encode information efficiently from each modality, before combining them in a later stage. However, this 2D approach does not combine features based on axial, coronal, and sagittal orientations that may greatly reduce its generalization on 3D images.

[Feng et al., 2019] considered MRI modality unavailability during segmentation by introducing sequence dropout. This is an important point since the availability of all the modalities is not always ensured between datasets that can greatly reduce the generalization capacity of a method. This framework randomly drops specific MRI sequences during training, with the intent to learn the intrinsic information of each sequence. This technique showed it can produce acceptable segmentation even in the absence of one or two modalities. Nonetheless, it is less efficient than other state-of-the-art methods when all modalities are available (will be detailed in section 2.3.3).

[Aslani et al., 2020] tackled the problem of generalization to new domains by integrating a regularization network to the traditional encoder-decoder network. The regularizer penalizes the network when the latter learns features that allow the prediction of MRI scanning sites. However, [Li et al., 2018] have argued that such strategies suffer from overfitting, the obtained representation could well generalize for all the source domains but poorly for the unknown target domains.

All the cited MS methods [Brosch et al., 2016, Valverde et al., 2017, Feng et al., 2019, Hashemi et al., 2018, Zhang et al., 2019b, Aslani et al., 2019, Feng et al., 2019] focused on obtaining accurate segmentation within a same domain

evaluation. However, the use of out-of-domain datasets is essential to ensure a good evaluation of the generalization capabilities of a method. This question is right now a hot topic (see [Mårtensson et al., 2020], and [Bron et al., 2021]) and an important recommendation from the clinical world (see [Omoumi et al., 2021]). Experiments using training and testing images derived from the same domain are known to be biased [Omoumi et al., 2021] and do not ensure generalization. Therefore, a model used in clinical conditions should produce accurate segmentation for new domain images without the need of retraining with expert segmentation on the new domain.

2.1.3 Proposals

In this chapter, we propose DeepLesionBrain (DLB), a novel method for MS lesion segmentation robust to domain shift, validated on out-of-domain testing (cross-dataset testing).

First, we use a large group of compact 3D CNNs spatially distributed over the brain with overlapping receptive fields between regions.

By associating a distinct network with each region of the brain, the spatially distributed networks (see [Coupé et al., 2020, Huo et al., 2019]) strategy simplifies the MS lesion segmentation from a single complex task on the whole brain to multiple simpler sub-tasks on each region. Moreover, the overlapping regions ensure consistent and stable consensus.

Applied to brain segmentation, this strategy demonstrated good generalization on unseen domains (*e.g.*, child’s brain or patients with Alzheimer’s disease) when trained on healthy adult brains [Coupé et al., 2020].

Second, to extract more relevant features for MS that may lead to better generalization, we focus on feature learning strategy. We consider that a generalizing model should learn two types of features: first global and generic features, and second local and specific features. Therefore, we propose Hierarchical Specialization Learning (HSL) to efficiently extract those features in two-steps. In the first step, a single network (the generic network) is trained on all brain regions. In the second step, each network of the spatially distributed networks is initialized with the generic network weights and specialized for a specific region of the brain.

Third, DLB is trained with a novel Image Quality Data Augmentation (IQDA) method, which mimics real-world data diversity by adding realistic alterations to the training images. As shown in the works of [Zhang et al., 2020a], such a type of regularization technique aims to reduce covariate shift. IQDA proposes specific augmentations that constrain task learning to be independent from source data acquisition resolution, data contrast, or data quality. Consequently, the proposed augmentation strategy enables domain shift robustness.

Finally, we propose a method using only two modalities (T1 and FLAIR) to ensure its compatibility with a large number of datasets. Most of the methods [Feng et al., 2019, Brosch et al., 2016, Valverde et al., 2017, Zhang et al., 2019b] optimize their segmentation using T1, FLAIR, PD, and T2 modalities. However, in clinical conditions, not all these sequences are always available. Therefore we focused our work on developing a robust approach using only two modalities.

2.2 Method and Material

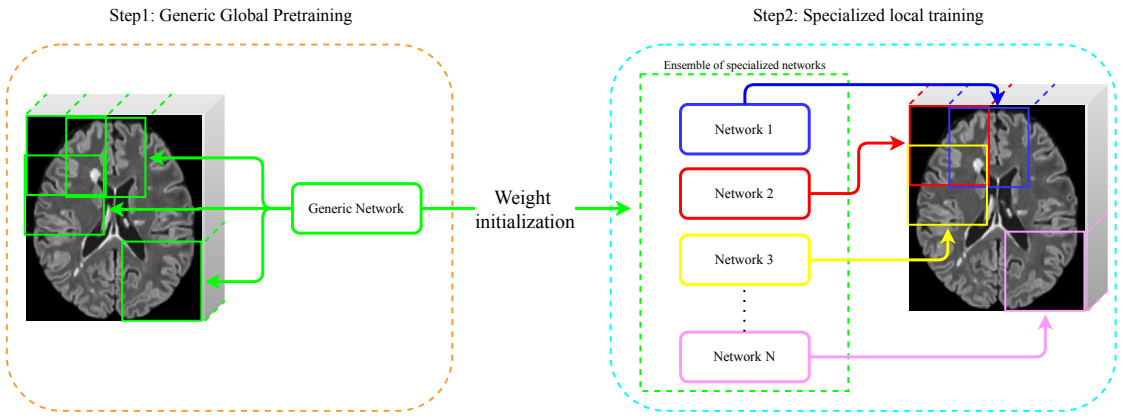


Figure 2.1: The two-steps training process of DeepLesionBrain (see Sect. 2.2.1.1 for more details).

2.2.1 Method Overview

2.2.1.1 Spatially Distributed Networks with Hierarchical Specialization

Spatially Distributed Networks Strategy

DLB is based on a spatially distributed networks strategy, proposed by [Coupé et al., 2020] and [Huo et al., 2019]. Such a strategy uses a group of compact networks, where each network is specialized in a particular region of the brain, and processes a sub-volume of the global volume. The receptive fields of the neighboring networks overlap with one another, and the final segmentation of the whole volume is obtained with a majority vote of the local predictions. Employing our spatially distributed compact networks is equivalent to a big network with more filters and a higher receptive field (see Fig. 2.1). This particular configuration with

overlapping receptive fields aims to produce a more robust segmentation compared to an individual network. As shown in the work of [Huo et al., 2019] that compared between 8 networks configuration with no-overlapping and 27 networks with overlapping, having an average prediction over overlapping regions led to significantly better performance on the evaluation data-sets including out-of-domain (on infants) data (*i.e.*, better generalization). Similarly, [Coupé et al., 2020] confirmed that larger overlapping led to better performance until the limit of 125 networks where performance peaked and stayed stable for 125, 216, and 343 networks. Additionally, we assume that averaging the prediction of a group of networks that have been trained separately on different sub-volumes is less likely to collapse than having a single network trained on the same set of data.

Hierarchical Specialization Learning (HSL)

To improve generalization for the task of MS lesion segmentation, we take inspiration from MS lesion features and propose a better learning strategy. MS lesions features can be grouped into two categories:

First, some lesion characteristics are considered generic and shared among lesion types. Such features are independent of lesion localization. They have a common and inherent significance at the global scale of the brain volume, we will refer to them in this chapter as “generic global features”.

Second, other relevant features for MS lesions depend on brain structure and some distinct regions (see [Filippi et al., 2019a]). In this work, we refer to these features as “specialized local features”.

On the one hand, training each specialized network on a specific sub-region of the brain (see Fig. 2.1 right) would prevent the efficient learning of “generic global features”, since each specialized member of our group would be trained on a particular region of the whole brain. On the other hand, using a single 3D CNN to learn “specialized local features” over the whole brain volume would require a large model which may not fit into memory and a large dataset to train it.

To overcome this limitation, we propose a novel Hierarchical Specialization Learning (HSL). Fig. 2.1 shows our two-step learning process. First, the “generic network” is trained with data samples from all over brain regions to learn general knowledge about lesions by extracting “generic global features”. Second, each network in the spatially distributed strategy is specialized over a specific sub-volume of the brain.

The generic network is used as an initialization for each network of our spatially distributed networks, by transferring the generic network weights to each individual specialized network. The knowledge gained from this transfer learning transmits the ability to extract “generic global features”, while the specialized network training will specialize them in extracting local “specialized local features”.

In our ablation study, we will show that this hierarchical specialization learning of the specialized networks performs better than training a single network over the whole brain, or training the specialized networks without HSL.

2.2.1.2 Image Quality Data Augmentation (IQDA)

The quality of the MRI greatly varies between datasets. In fact, the quality of the images depends on several factors such as signal-to-noise ratio, contrast-to-noise ratio, resolution, or slice thickness. To address this issue, we propose a data augmentation strategy that considers image quality disparity. During training, we simulate “on the fly” altered versions of 3D patches. We randomly introduce at each iteration either blur, edge enhancement, or axial subsampling distortion (2D FLAIR are usually acquired along the axial direction). For the blur, a gaussian kernel is used with a randomly selected standard deviation ranging between $[0.5, 1.75]$. For edge enhancement, we use unsharp masking with the inverse of the blur filter. For axial subsampling distortion, we simulate acquisition artifacts that can result from the varying slice thickness. We use a uniform filter (a.k.a mean filter) on the axial direction with a size of $[1 \times 1 \times sz]$ where $sz \in 2, 3, 4$. Ground truth is kept the same as the original version. This process reduces the domain bias when learning to extract relevant features caused by data variability.

2.2.1.3 Selection of the Required Modalities

To use a trained model for MS lesion segmentation with optimal performance, it usually requires the use of the same set of modalities that have been chosen during training. DLB proposes a method that needs only T1 and FLAIR sequences to be compatible with all benchmark MS datasets and most already available MS patients data.

Our method is built with the purpose to generalize on unseen datasets, thus it uses the minimum necessary modalities. Indeed, increasing the number of sequences requires a longer scan acquisition time. Besides, it needs more complex processing which may be prone to error, such as multimodal image registration. Furthermore, the use of more sequences during the training on a dataset may reduce the generalization to other image domains.

In addition to the wide use of T1 and FLAIR for MS diagnosis, the choice of these modalities has also been motivated by the fact that FLAIR is the most relevant sequence for revealing most MS lesions (see [Narayana et al., 2020]), while T1 can provide complementary structural information needed for accurate segmentation.

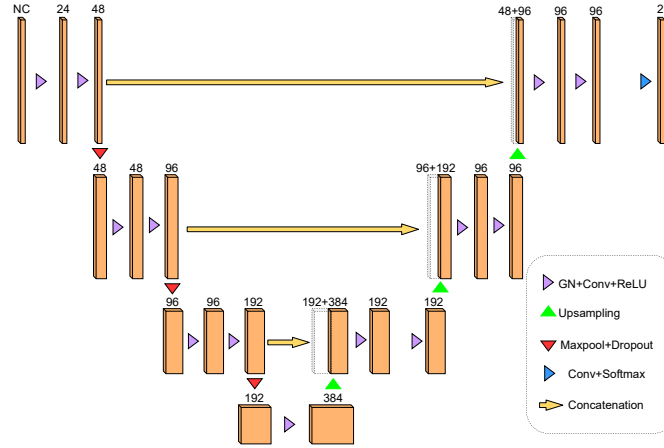


Figure 2.2: Illustration of the considered U-Net architecture. The number of input channels (NC) depends on the modality number (*i.e.*, NC= 2, for using T1 and FLAIR). Each block is composed of group normalization (GN), Convolution (Conv) and Rectified Linear Unit (ReLU) activation.

2.2.2 Implementation Details

The network architecture used in our spatially distributed strategy is based on 3D U-Net composed of a downsampling part and an upsampling one, linked with one another by skip connections at the multiple scales. This 3D CNN architecture, shown in Figure 2.2, has been used for all the networks in our approach. Dropout with 0.5 rate is used after max-pooling layers to prevent the overfitting of our model to the training data. Due to GPU memory constraints, we trained with a batch size of 1, and so we used Group Normalization (GN) [Wu and He, 2018] with 8 groups before each convolution. We have chosen Rectified Linear Units (ReLU) to introduce non-linearity after convolution layers. DLB is optimized with Adam [Kingma and Ba, 2014] using a learning rate of 0.0001 and a momentum of 0.9. The experiments have been performed with Keras 2.2.4 [Chollet et al., 2015] and Tensorflow 1.12.0 [Abadi et al., 2016] on NVIDIA Titan Xp 12 GB GPU.

2.2.3 Method Description

To obtain sub-volumes for each image, we first divide our whole MRI registered into the Montreal Neurological Institute template space (MNI space) into multiple overlapping 3D patches. We perform a cropping operation over the whole image using a sliding window of the size (Px, Py, Pz) , and a stride of (Sx, Sy, Sz) . We take $Sx < Px$, $Sy < Py$, $Sz < Pz$ to ensure the overlapping.

Sub-volumes from different images, that represent the same receptive field into the MNI space (the same sub-volume region of the whole volume), are grouped

together. They are used for training a network specialized for that particular region.

In this work, we explored many combinations of sub-volume sizes and numbers. We chose a configuration with 125 sub-volumes by taking experimentally $Px = Py = Pz = 96$, $Sx = Sy = 76$, and $Sz = 67$ as a good trade-off between the overall performance and computation resources.

2.2.3.1 Symmetrical Training

To limit redundant training and to use the most possible data for a particular brain region, we choose to train specialized networks only on one hemisphere. By flipping (mirroring) the sub-volumes of the second hemisphere, it is possible to train a single specialized network for both sides. Thus, we can use twice the amount of data for each region while reducing the number of networks to train to nearly a half (due to sub-volume overlapping, the plane of symmetry cuts through the median networks which cover equivalent symmetrical regions from both hemispheres). Consequently, unlike previous works with spatially distributed networks (*i.e.*, [Coupé et al., 2020]), instead of using 125 networks we use only 75 specialized networks. We experimentally verified that using 125 networks or only 75 trained with twice the number of patches, produced similar segmentation accuracy.

2.2.3.2 Loss Function

MS lesion segmentation task suffers from class imbalance since lesion volume is considerably low compared to healthy volume. Thus, we use a smooth version of the Generalized Jaccard Loss (GJL), which considers this issue [Manjón et al., 2022].

$$GJL = 1 - \frac{\sigma + \sum_{c=1}^{Nc} w_c \sum_{i=1}^N p_{ci} t_{ci}}{\sigma + \sum_{c=1}^{Nc} w_c \left(\sum_{i=1}^N (p_{ci} + t_{ci}) - \sum_{i=1}^N p_{ci} t_{ci} \right)} \quad (2.1)$$

Where $w_c = 1/(1 + \sum_{i=1}^N t_{ci})$, σ is the smoothness factor, N is the number of voxels, Nc is the number of classes, p_{ci} and t_{ci} are respectively the predicted probability and the ground truth probability of the voxel i for the class c .

During inference, we combine the overlapping predictions in a straightforward majority vote technique. The class of each voxel (either lesion or healthy tissue) is chosen based on the most predicted class among the networks which cover that voxel.

2.2.4 Datasets

To assess the robustness of a model, it is crucial to test its ability to generalize on unseen domains. Therefore, DLB has been trained and validated using different datasets to assess its domain generalization ability (see 2.3.2). These datasets exhibit high heterogeneity in terms of resolution, data processing, acquisition sites, delineation protocols, and they also cover a large variety of clinical scenarios.

2.2.4.1 ISBI Longitudinal Multiple Sclerosis Lesion

The ISBI dataset [Carass et al., 2017] consists of five subjects for training, fourteen subjects for testing, with a mean of 4.4 time-points per subject (21 images for training and 61 images for testing). Two human expert raters delineated MS lesions, from the four available modalities acquired on 3.0 Tesla MRI scanner: 3D MPRAGE T_1 -weighted (T_1) of $0.82 \times 0.82 \times 1.17 \text{ mm}^3$ voxel size, 2D T_2 -weighted (T_2), 2D T_2 -weighted fluid attenuated inversion recovery (FLAIR), and 2D Proton Density weighted (PD), of $0.82 \times 0.82 \times 2.2 \text{ mm}^3$ voxel size each.

For the training, we used the ISBI training dataset with available annotations from the two experts. For test and evaluation, we segmented the test data with no available expert annotation, and submitted our results to the ISBI challenge website¹. The ISBI pipeline already included preprocessing. Each first time-point T_1 was inhomogeneity-corrected using N4 [Tustison et al., 2010], skull-stripped [Carass et al., 2007], dura stripped [Shiee et al., 2014], followed by a second N4 inhomogeneity correction, and rigid registration to a 1 mm^3 isotropic MNI template. Then, this image was used as a target for the remaining T_1 time-points and all modalities for the same subject. These images were N4 corrected and then rigidly registered to the T_1 in the MNI space. The skull and dura-stripped mask from the target T_1 was applied, which were then N4 corrected again. We added an intensity normalization step using kernel density estimation for all images.

2.2.4.2 MICCAI2016 MS Challenge Dataset

The MSSEG’16 training dataset [Commowick et al., 2016] contains 15 patients from 3 different clinical sites. Five modalities are available for each patient: 3D FLAIR, 3D T_1 , 3D T_1 -Gd, 2D PD, and 2D T_2 . The images were acquired on 1.5T and 3T MRI scanners with multiple resolutions: 3D FLAIR modalities ranging from $1 \times 0.5 \times 0.5$ to $1.25 \times 1.04 \times 1.04 \text{ mm}^3$, and 3D T_1 sequences between $0.85 \times 0.74 \times 0.74$ and $1.08 \times 1.08 \times 0.9 \text{ mm}^3$.

Seven human experts have manually segmented the multiple sclerosis lesions. Each patient modalities have been preprocessed with the same pipeline. First, each

¹<https://smart-stats-tools.org/lesion-challenge-upload-results>

sequence was denoised using the non local means algorithm [Coupé et al., 2008]. Second, a rigid registration of each modality on the FLAIR was performed [Commonick et al., 2012]. Then, skull stripping of T1 was performed using the volBrain platform [Manjón and Coupé, 2016], the same mask is applied to other modalities. Finally, bias field correction was applied using the N4 algorithm [Tustison et al., 2010]. In addition to these steps that have been performed on the available images, each modality was registered to the MNI space for our experiments. Similarly, to the ISBI images, we used kernel density estimation for the normalization step.

2.2.4.3 In-house Dataset

For further evaluation of our approach, we used an In-house 3D MRI dataset, with 3D T1 and 3D FLAIR modalities [Coupé et al., 2018]. This dataset contains 43 subjects diagnosed with MS. The images were acquired with different scanners and multiple resolutions ($0.6 \times 0.6 \times 0.65 \text{ mm}^3$, $0.5 \times 0.5 \times 0.9 \text{ mm}^3$, and $1 \times 1 \times 1 \text{ mm}^3$).

The dataset lesion masks have been obtained by human experts manual delineation. The images were pre-processed using the lesionBrain pipeline from the volBrain platform [Manjón and Coupé, 2016]. First, it included denoising of each modality [Coupé et al., 2008]. Second, an affine registration to MNI space was performed on the T1, then the FLAIR was registered to the transformed T1. Skull stripping and bias correction have been performed on the modalities, followed by a second denoising. Finally, the intensities have been normalized.

Table 2.1: Description of datasets used in this work.

	2D/3D	Site	# Subjects	# Raters	Modalities
ISBI train-set	2D	Mono	5	2	T1, FLAIR, PD, T2
MSSEG'16	3D	Multi	15	7	T1, T1-Gd, FLAIR, PD, T2
In-house	3D	Multi	43	1	T1, FLAIR

2.2.4.4 Datasets Summary

Table 2.1 summarizes the main differences between the 3 datasets. We focused specifically on the resolution of FLAIR due to its known relevance in MS lesion segmentation. To summarize, ISBI train-set contains multiple time points of only five subjects, acquired in a single clinical site with two human expert segmentations. Except for 3D MPRAGE T1, the other three modalities are in 2D.

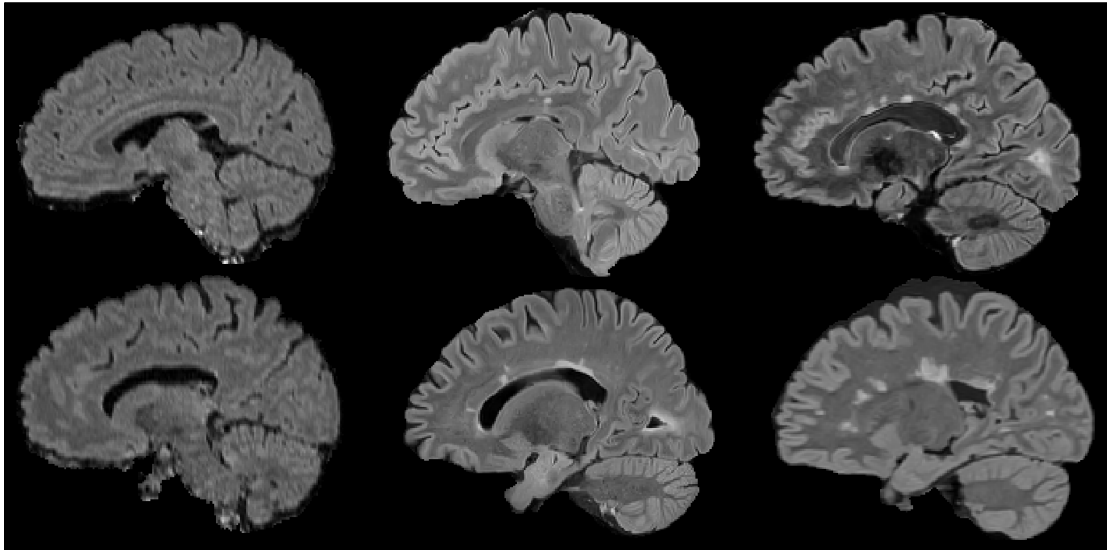


Figure 2.3: FLAIR examples from the considered three datasets in the MNI space and after intensity normalization. From left to right, the two images are from ISBI, MSSEG'16, our in-house data, respectively.

MSSEG'16 dataset is a multi-site database comprising 15 patients, with seven human segmentations. This dataset contains 5 available modalities with 3D FLAIR. Finally, the In-house dataset is the largest dataset with 43 patients, and multi-site 3D modalities, segmented by a single human rater and validated by a second one.

Figure 2.3 shows examples from the three presented datasets, each image represents a sagittal section of the FLAIR modality in the MNI space after intensity normalization. The two images on the left are examples from the ISBI dataset. We notice blurring effects which makes it hard to distinguish precisely brain structures. This blur comes from 2D low-resolution acquisitions. In the middle, the two examples come from the MSSEG'16 dataset. These 3D FLAIR are noticeably of higher resolution than the other images. Therefore, lesion boundaries are more easily delineated and main structures are apparent. The final two images on the right are from our In-house 3D dataset. The 3D resolution enables the differentiation of white matter, gray matter, and shows the lesions clearly. In terms of FLAIR images, we notice that both MSSEG'16 and In-house dataset (3D FLAIR) propose better visual quality than ISBI dataset (2D FLAIR).

2.2.5 Validation Framework

2.2.5.1 Evaluation Metrics

The assessment of a segmentation method is usually measured by a similarity metric between the predicted segmentation and the human expert ground truth.

First, we use several complementary metrics to assess segmentation performance. Namely, we use the Dice similarity coefficient, the Positive Predictive Value (PPV), True Positive Rate (TPR), and Pearson's correlation coefficient (CORR).

$$PPV = \frac{TP}{TP + FP}, \quad TPR = \frac{TP}{TP + FN}, \quad (2.2)$$

$$Dice = \frac{2 \times TP}{(TP + FN) + (TP + FP)}, \quad (2.3)$$

where TP, FN, FP represent respectively true positives, false negatives, and false positives.

Second, recent works (*i.e.*, [Commowick et al., 2018]) question the relevance of classic metrics (Dice) compared to detection metrics, which are used for MS diagnostic and clinical evaluation of the patient evolution. Thus, in addition to the voxel-wise metrics, we also use lesion-wise metrics that focus on the lesion count. Such as Lesion False Positive Rate (LFPR) and Lesion True Positive Rate (LTPR).

$$LTPR = \frac{LTP}{LTP + LFN}, \quad LFPR = \frac{LFP}{LTP + LTN}, \quad (2.4)$$

where LTP, LFN, LFP represent respectively lesion true positives, lesion false negatives, and lesion false positives.

Moreover, [García-Lorenzo et al., 2013] pointed out that even though Dice is commonly used and simplifies method comparison, multiple complementary metrics are needed to provide a better understanding of the performance. Recently, international challenges took into consideration several metrics ([Carass et al., 2017] and [Commowick et al., 2018]). Consequently, we decided to evaluate our methods using Hybrid score proposed by [Carass et al., 2017]. This metric combines voxel-wise segmentation, lesion-wise detection, and volumetric metrics. It is defined as:

$$Hybrid = \frac{Dice}{8} + \frac{PPV}{8} + \frac{(1 - LFPR)}{4} + \frac{LTPR}{4} + \frac{CORR}{4} \quad (2.5)$$

Finally, we also use the ISBI Submission Score (Sub. Score) for the evaluation of ISBI test-set segmentations. [Carass et al., 2017] defined it as the average of the

hybrid scores of all image examples with the different human raters and with inter-rater variability taken into consideration. This score is computed after submitting the segmentation to ISBI’s challenge website ². Obtaining an ISBI score of 90 or higher with a segmentation technique indicates that this method is similar to the human raters.

2.2.5.2 Reference Methods

During experiments, our method was compared to three publicly available state-of-the-art approaches. We performed training and validation for all three compared methods, in the same conditions regarding datasets and preprocessing. The reference methods are **nicMSlesion** by [Valverde et al., 2019], **DeepMedic** by [Kamnitsas et al., 2017], and **2.5D Tiramisu** by [Zhang et al., 2019b]. These methods have been selected for the availability of the authors source code and the relevance of their contributions in the MS segmentation community.

nicMSlesion: This method is based on a cascade of two 3D patch-wise CNNs. The first one is trained to be sensitive to reveal lesion candidates. The second one is trained to reduce the misclassified voxels from the first network. Training is performed on $11 \times 11 \times 11$ patches randomly augmented with flipping and rotations. Therefore, **nicMSlesion** involves classical data augmentation. In the first network, the negative class is under-sampled to the same number of existing lesion voxels. It is composed of patches extracted from all of the available lesion voxels and a random selection of normal-appearing tissue voxels. Afterwards, an evaluation of the first CNN model is computed by performing inferences on the same train-set and identifying negative voxels that have been misclassified as lesions (False Positives). Finally, the second model is trained using a balanced set composed of all the lesion voxels and a random selection from the identified False Positives in the previous step.

DeepMedic: This method is based on an 11-layers deep dual pathway 3D CNN designed for brain lesion segmentation. to incorporate both local and larger contextual information, the dual pathway architecture processes the input images at multiple scales simultaneously. To overcome the computational burden, the authors use a hybrid dense training scheme processing adjacent image patches into one pass through the network. To refine the network segmentation and remove false positives, a 3D fully connected conditional random field is used. The training includes data augmentation with sagittal reflections.

2.5D Tiramisu: This method is based on a fully convolutional densely connected network. The model uses stacked slices from all three anatomical planes to achieve a 2.5D based method. Individual slices from a given orientation provide

²<https://smart-stats-tools.org/lesion-challenge>

global context along the plane and the stack of adjacent slices adds local context. The training also includes flipping and rotations of the 2D patches for data augmentation. Therefore, 2.5D Tiramisu involves classical data augmentation. For each stack of 2D 128×128 slices composed of a center slice and its 2 adjacent slices, the model produces the segmentation of the center slice. Then, the inference results along the different orientations are combined via a majority vote to output the final segmentation.

For both these methods, we use the implementations provided publicly by the authors (see ³ and ⁴).

2.2.5.3 Statistical Test

To assert the advantage of a technique obtaining the highest average score, we conducted a two-sided Wilcoxon test (*i.e.*, paired statistical test) over the lists of hybrid scores measured at image level (for the consistency of the segmentation section we took the lists of dice indices between the two segmentations). The significance of the test is established for a p-value below 0.05. In the following tables, * indicates a significantly better average score when compared with the rest of the other approaches.

2.3 Results

2.3.1 Ablation Study

To demonstrate the impact of each proposed contribution on domain generalization, we measured separately their effects on different metrics. To show both the effect on accuracy improvement and the domain shift robustness, we propose an out-of-domain and in-domain ablation study. First, we trained each method configuration on ISBI challenge train-set, then we validated on both ISBI test-set (see Table 2.2) and MSSEG'16 (see Table 2.3). To ensure a fair comparison, each configuration is trained until convergence. Specifically, we used an early stopping criterion of 50 epochs (*i.e.*, the training stops if the loss function does not improve on the validation set during 50 epochs) with a maximum number of 500 epochs. We verified that none of the configurations reached this maximum number.

Table 2.2 shows the effect of each contribution to segmentation accuracy, when trained on ISBI challenge train-set and tested on ISBI test-set. First, the best performing combination is DLB with HSL and IQDA, it obtained an ISBI Score of 92.849. Second, both the versions of DLB without IQDA and DLB without HSL

³<https://github.com/sergivalverde/nicMSlesions>

⁴<https://github.com/MedICL-VU/LesionSeg>

Table 2.2: Ablation study results with different variants of our approach trained on ISBI challenge train-set and tested on ISBI test-set. DeepLesionBrain (DLB) refers to using our spatially distributed specialized networks, each network in charge of segmenting a sub-volume. The generic network represents the variant of DLB with a single network (without the spatially distributed strategy). Hierarchical Specialized Learning (HSL) indicates that we initialized the “specialized networks” with the “generic Network”. To evaluate the performance of the proposed Data Augmentation, we compared variants with IQDA (previously defined in 2.2.1.2) and without IQDA. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the other approaches using the two-sided Wilcoxon test.

Method	Hybrid	Dice	PPV	TPR	LFPR	LTPR	Sub. Score
DLB with HSL and IQDA	0.747*	0.646	0.888	0.545	0.131	0.486	92.849
DLB with HSL and without IQDA	0.732	0.677	0.849	0.603	0.192	0.489	92.383
DLB without HSL and with IQDA	0.710	0.576	0.892	0.453	0.121	0.360	91.713
DLB with models genesis init. and IQDA	0.718	0.621	0.867	0.513	0.187	0.438	91.885
DLB with AssemblyNet init. and IQDA	0.723	0.628	0.885	0.515	0.140	0.406	92.109
The generic network with IQDA	0.736	0.668	0.859	0.585	0.178	0.489	92.491
The generic network without IQDA	0.688	0.654	0.502	0.869	0.162	0.468	92.425

are less accurate. They obtained respectively ISBI scores of 92.383 and 91.713. The later comparison shows the impact of HSL on the accuracy of segmentations. Moreover, the generic network is less accurate than our spatially distributed approach used in DLB. The variant of generic Network with IQDA obtained a score of 92.491, whereas the variant without IQDA obtained a hybrid score of 92.425. Finally, we compare HSL with other weight initialization strategies. Specifically, HSL is compared with the neighbor transfer learning from AssemblyNet proposed by [Coupé et al., 2020] and models genesis proposed by [Zhou et al., 2021]. Although both variants obtained a better score compared to DLB without HSL, both initialization strategies gave a lower score than DLB with HSL and IQDA

Table 2.3: Ablation study results with different variants of our approach trained on ISBI challenge train-set and tested on MSSEG’16 (see caption of Table 2.2 for details).

Method	Hybrid	Dice	PPV	TPR	LFPR	LTPR
DLB with HSL and IQDA	0.684*	0.639	0.768	0.608	0.319	0.700
DLB with HSL and without IQDA	0.673	0.669	0.728	0.671	0.416	0.725
DLB without HSL and with IQDA	0.648	0.562	0.806	0.489	0.320	0.629
DLB with models genesis init. and IQDA	0.623	0.593	0.737	0.576	0.436	0.665
DLB with AssemblyNet init. and IQDA	0.610	0.541	0.708	0.537	0.466	0.705
The generic Network with IQDA	0.672	0.665	0.721	0.673	0.413	0.727
The generic network without IQDA	0.626	0.625	0.763	0.588	0.449	0.611

Table 2.3 shows the effect of each contribution to domain shift robustness, when trained on ISBI challenge train-set and tested on MSSEG’16. First, the most robust combination is DLB with HSL and IQDA, it obtained a hybrid score of 0.684. Second, both the variants of DLB without IQDA and DLB without HSL are less accurate. They obtained hybrid scores of 0.673 and 0.648 respectively. Moreover, the generic network is less robust than our spatially distributed approach with DLB. The variant of the generic network with IQDA obtained a score of 0.672, whereas the variant without IQDA obtained a hybrid score of only 0.626. The later comparison shows the impact of IQDA on robustness even without the spatially distributed networks. Finally, HSL is compared with Assemblynet [Coupé et al., 2020] and model genesis [Zhou et al., 2021] initialization strategies. The variants with model genesis and AssemblyNet initialization methods obtained respectively hybrid scores of 0.623 and 0.6103.

2.3.2 Cross-dataset Testing

In this section, we assess the cross-dataset robustness and generalization ability of our proposed approach. We chose to compare our method with three state-of-the-art approaches: nicMSlesion [Valverde et al., 2019], DeepMedic [Kamnitsas et al., 2017], and 2.5D Tiramisu [Zhang et al., 2019b].

During the proposed validation, all the methods have been trained on exactly the same dataset (*i.e.*, same preprocessing, same number of modalities, *etc.*) to ensure a fair comparison of method performance. Although reference methods have been originally proposed with a specific number of modalities (*i.e.*, Tiramisu 2.5D and nicMSlesion were tested with 4 and 3 modalities respectively), their implementation is independent of the number of modalities since all modalities are concatenated and fed to the CNN. Besides, their official open-source implementations support the usage of only T1 and FLAIR sequences. Thus, in this evaluation, all methods are trained using only these two modalities. The following cross-dataset testing (cross-domain testing) consists in training each technique on one dataset at each time. Afterward, the obtained models are evaluated on the other datasets which contain unseen domains. We verified the average inference time per image for each method on the same machine and the same preprocessed images: 57.353s for nicMSlesion, 17.547s for DeepMedic, 47.471s for 2.5D Tiramisu, and 38.014s for DLB (this time does not include image preprocessing). Unlike using a single network to segment patches coming from the entire image, DLB uses multiple networks, one network for each sub-volume. These networks are loaded one by one to enable the use of a common GPU hardware solution (*e.g.*, NVIDIA Titan Xp with 12 GB in our setup). Even though DLB requires sequential loading of multiple networks on GPU, the inference time over the whole image is similar to using a single network since network weights loading time is negligible compared to patch

segmentation time. The ISBI score is returned by the challenge website only for ISBI test-set evaluation, and thus this metric is not available (NA) for testing on other datasets.

2.3.2.1 Trained on ISBI

Table 2.4: Results of the different approaches trained on the ISBI training dataset, with T1 and FLAIR modalities. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the three other approaches using the two-sided Wilcoxon test. Red values indicate hybrid scores lower than 0.5 or Dice index below 0.25.

Tested on	Approach	Hybrid	Dice	PPV	TPR	LFPR	LTPR	CORR	Sub. Score
MSSEG'16	nicMSlesion	0.537	0.442	0.614	0.423	0.504	0.629	0.495	NA
	DeepMedic	0.510	0.476	0.542	0.560	0.829	0.850	0.509	NA
	2.5D Tiramisu	0.711	0.664	0.741	0.658	0.284	0.695	0.730	NA
	DLB	0.684	0.639	0.768	0.608	0.319	0.700	0.650	NA
In-house dataset	nicMSlesion	0.419	0.204	0.727	0.129	0.309	0.361	0.158	NA
	DeepMedic	0.523	0.536	0.633	0.499	0.805	0.765	0.549	NA
	2.5D Tiramisu	0.654	0.545	0.871	0.410	0.204	0.476	0.635	NA
	DLB	0.696*	0.675	0.850	0.564	0.342	0.644	0.718	NA

Table 2.4 shows the results of segmentation when training the different approaches using T1 and FLAIR modalities, on the ISBI training dataset (2D resolution FLAIR).

When validating the methods on MSSEG'16, we report that 2.5D Tiramisu obtained slightly better results (not significantly) than DLB, in terms of hybrid score whereas nicMSlesion and DeepMedic performed relatively worse with 0.537 and 0.51 respectively.

On our in-house dataset, DLB performed significantly better with a hybrid score of 0.696 while 2.5D Tiramisu, DeepMedic, and nicMSlesion obtained respectively 0.654, 0.523, and 0.419. We can notice that nicMSlesion offers poor cross-domain performance on 3D FLAIR when trained with a 2D FLAIR dataset.

2.3.2.2 Trained on MSSEG'16

Table 2.5 shows the results of segmentation when training the different approaches on the MSSEG'16 dataset comprising 3D T1 and 3D FLAIR modalities. First, we notice that our approach obtained significantly better hybrid scores for both the ISBI test and the In-house datasets. Second, when validating on ISBI, the obtained submission score is 89.043 for DLB (the closest to human performance), 87.344 for DeepMedic, 87.173 for nicMSlesion, and 86.686 for 2.5D Tiramisu (the farthest from human performance). In the same conditions, 2.5D Tiramisu obtained the

Table 2.5: Results of the different approaches trained on the MSSEG’16 dataset, with T1 and FLAIR modalities. For each metric, the bold values indicate the best result. In hybrid score column, * indicates a significantly better score than the three other approaches using the two-sided Wilcoxon test. Red values indicate hybrid scores lower than 0.5 or Dice index below 0.25.

Tested on	Approach	Hybrid	Dice	PPV	TPR	LFPR	LTPR	CORR	Sub. Score
ISBI test-set	nicMSlesion	0.555	0.398	0.717	0.292	0.368	0.206	0.822	87,173
	DeepMedic	0.547	0.378	0.801	0.265	0.416	0.298	0.717	87,344
	2.5D Tiramisu	0.462	0.165	0.937	0.096	0.075	0.160	0.212	86,686
	DLB	0.618*	0.535	0.697	0.471	0.353	0.373	0.835	89.043
In-house dataset	nicMSlesion	0.669	0.686	0.689	0.705	0.467	0.717	0.737	NA
	DeepMedic	0.597	0.645	0.647	0.670	0.721	0.811	0.650	NA
	2.5D Tiramisu	0.664	0.706	0.766	0.694	0.432	0.801	0.552	NA
	DLB	0.697*	0.746	0.681	0.847	0.478	0.754	0.799	NA

average Dice of 0.165 which indicates a failure of the method and thus a lack of generalization in this scenario (when trained on high-quality 3D FLAIR and tested on low-quality 2D FLAIR). Finally, for the In-house dataset, DLB produced significantly better segmentation than other methods. DLB obtained a hybrid score of 0.697 while nicMSlesion obtained 0.669, 2.5D Tiramisu obtained 0.664, and DeepMedic obtained the lowest score of 0.597.

2.3.2.3 Trained on In-house

Table 2.6: Results of the different approaches trained on In-house dataset, with T1 and FLAIR modalities. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the three other approaches, using the two-sided Wilcoxon test. Red values indicate hybrid scores lower than 0.5 or Dice index below 0.25.

Tested on	Approach	Hybrid	Dice	PPV	TPR	LFPR	LTPR	CORR	Sub. Score
MSSEG’16	nicMSlesion	0.700	0.650	0.822	0.586	0.150	0.607	0.607	NA
	DeepMedic	0.717	0.694	0.750	0.701	0.345	0.782	0.709	NA
	2.5D Tiramisu	0.745	0.665	0.741	0.687	0.164	0.720	0.722	NA
	DLB	0.741	0.719	0.735	0.744	0.209	0.671	0.776	NA
ISBI test-set	nicMSlesion	0.453	0.131	0.644	0.075	0.338	0.050	0.712	84,512
	DeepMedic	0.523	0.385	0.807	0.273	0.388	0.215	0.670	86,810
	2.5D Tiramisu	0.608	0.355	0.938	0.231	0.065	0.160	0.689	89,289
	DLB	0.638*	0.476	0.877	0.348	0.104	0.193	0.787	89.843

Table 2.6 shows the results of segmentation when training on our In-house dataset with 3D FLAIR. First, the obtained results when testing on MSSEG’16 indicates a close segmentation accuracy for DLB and 2.5D Tiramisu in terms of hybrid score (0.745 and 0.741) and slightly lower performance from nicMSlesion

and DeepMedic (0.7 and 0.717). Second, we notice that our approach obtained a significantly higher hybrid score when validating on the ISBI testing dataset, with a submission score of 89.843 compared to 2.5 Tiramisu, DeepMedic, and nicMSlesion with 89.289, 86.810, and 84.512 respectively. In this scenario, nicMSlesion obtained the worst score with a Dice of 0.131 indicating a failure of the method.

2.3.2.4 Cross-dataset Testing Summary

First, it is noteworthy that when our approach obtained a better score, the superiority was statistically significant. On the contrary, when one of the other approaches obtained a higher score, the advantage was not significant using the Wilcoxon test.

Second, it should be pointed out that in all the considered cross-domain cases, DLB did not degenerate not even once while maintaining high scores. We reported for nicMSlesion trained on ISBI and validated on the In-house dataset a hybrid score of 0.419. We also recall the low performance of 2.5D Tiramisu trained on MSSEG'16 and tested on ISBI (0.462 hybrid score). This shows the cross-domain robustness of the proposed strategy.

Table 2.7 sums up cross-dataset experiments results. This table presents the average score estimated over all the images obtained during the three experiments presented in Table 2.4, Table 2.5, and Table 2.6 (61 images for ISBI test-set, 43 images for In-house, 15 images for MSSEG'16). We notice that DLB obtains the highest hybrid score and Dice index by a large margin compared to 2.5D tiramisu, DeepMedic, and nicMSlesion.

Table 2.7: Summary of the cross-dataset experiment. The table represent the average of cross-dataset experiment results (see Table 2.4, Table 2.5, and Table 2.6) based on the number of images for each dataset. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the three other approaches using the Wilcoxon test.

Strategy	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	CORR
nicMSlesion	0.526	0.365	0.695	0.308	0.362	0.338	0.595
DeepMedic	0.554	0.483	0.725	0.429	0.556	0.520	0.649
2.5 D Tiramisu	0.608	0.443	0.870	0.368	0.179	0.402	0.537
DLB	0.663*	0.601	0.775	0.550	0.299	0.484	0.780

2.3.3 Same Domain Validation

Despite the previously mentioned limitations of in-domain validation, we also provide experiments using the same domain as complementary results. First, Table

2.8 shows the results of DLB, nicMSlesion, DeepMedic, and 2.5D Tiramisu on ISBI test-set after being trained on ISBI train-set (same domain), with T1 and FLAIR modalities. The three approaches give close results with submission scores of 92.923 for 2.5D Tiramisu, 92.849 for DLB, and 92.161 for nicMSlesion. DeepMedic comes last with a submission score of 90.866.

Second, Table 2.9 shows the current top-performing methods on the ISBI challenge website. 2.5D Tiramisu [Zhang et al., 2019b] is the best-ranked method with the current highest ISBI Score of 93.21, followed in second place by nnUnet [Isensee et al., 2019] with 93.09. Both approaches rely on 4 modalities (T1, FLAIR, T2, PD). Our approach comes in third place using only 2 modalities, and obtained the ISBI submission score of 92.85. Although DLB uses a lower number of modalities, it obtained better results than IMAGINE [Hashemi et al., 2018], Self-adaptive network [Feng et al., 2019], and Multi-branch [Aslani et al., 2019] that obtained respectively the scores of 92.49, 92.41, and 92.12.

Table 2.8: Results of the different approaches trained on the ISBI training dataset and tested on ISBI test-set, with T1 and FLAIR modalities. For each metric, the bold values indicate the best result. two-step in the hybrid score column, * indicates a significantly better score than the three other approaches using the two-sided Wilcoxon test.

Approach	Hybrid	Dice	PPV	TPR	LFPR	LTPR	CORR	Sub. Score
nicMSlesion	0.724	0.639	0.853	0.541	0.144	0.432	0.863	92.161
DeepMedic	0.649	0.643	0.827	0.557	0.408	0.530	0.873	90.866
2.5D Tiramisu	0.750	0.672	0.865	0.592	0.150	0.513	0.868	92.923
DLB	0.748	0.646	0.888	0.545	0.131	0.486	0.868	92.849

Table 2.9: State-of-the-art published results for the ISBI challenge

Approach	Modalities	CNN type	Sub. Score
2.5D Tiramisu [Zhang et al., 2019b]	T1, FLAIR, T2, PD	2D	93.21
nnUnet [Isensee et al., 2019]	T1, FLAIR, T2, PD	2D and 3D	93.09
DLB [ours]	T1, FLAIR	3D	92.85
IMAGINE [Hashemi et al., 2018]	T1, FLAIR, T2, PD	3D	92.49
Self-adaptive network [Feng et al., 2019]	T1, FLAIR, T2, PD	3D	92.41
Multi-branch [Aslani et al., 2019]	T1, FLAIR, T2	2D	92.12

The high accuracy of the results was expected as both the training and testing sets share the same domain (same acquisition conditions, and same scanner...). By tuning and adapting a method to this specific domain conditions, we can obtain artificially higher performance (*e.g.*, DLB with 4 modalities obtained a score of

92.92, and a 2D version of DLB obtained 93.14⁵). However, in our opinion, results reported in the same domain experiment do not truly reflect methods performances. For instance, the best performing method of this section (2.5D Tiramisu) failed when trained on different datasets (obtained submission scores of 89.043 and 89.289 in Table 2.5 and Table 2.6). The limitation of such a validation strategy is one of the main messages of our work. Hence, we consider that cross-dataset evaluation with diverse images from different domains is a better alternative for method assessment.

2.3.4 Cross-dataset Segmentation Consistency

Finally, a usually under-investigated method property is its cross-dataset segmentation consistency. To assess the consistency of our model segmentation, we decided to compare the segmentation produced by each approach on the same data when the model is trained on different datasets. We compute the Dice between the different segmentations of a method as a similarity index to quantify the prediction consistency. Table 2.10 shows the segmentation consistency for each approach in our cross-dataset setting.

First, we analyzed the segmentations on In-house when the models are trained respectively on ISBI train-set and MSSEG’16. In this case, DLB obtained the best score of 0.647, followed by 2.5D Tiramisu and DeepMedic with 0.6261 and 0.602 respectively. Lastly, nicMSlesion obtained a score of 0.217. Second, we analyzed the segmentations on MSSEG’16 when the models are trained respectively on ISBI train-set and In-house. In this case, we obtained close consistency scores for 2.5D Tiramisu and DLB with Dice scores around 0.72 while DeepMedic and nicMSlesion are less consistent with 0.537 and 0.514 respectively. Finally, we analyzed the segmentations on ISBI test-set when comparing the models trained on ISBI train-set, the models trained on In-house, and the models trained on MSSEG’16. For all settings, DLB was significantly more consistent than both other methods with a Dice ranging from 0.63 to 0.649. 2.5D Tiramisu segmentation consistency index varies from 0.217 to 0.485. DeepMedic consistency index fluctuates from 0.49 to 0.602. nicMSlesion is the least consistent with scores ranging from 0.177 to 0.512.

During our cross-dataset consistency experiment, DLB was the only method capable of ensuring segmentation consistency independent of the training dataset. Both other methods failed several times as indicated with red color in Table 2.10.

Figure 2.4 represents an image from the In-house dataset and the segmentation of the different methods when trained on the ISBI challenge and MSSEG’16 datasets. First, both nicMSlesion and 2.5D Tiramisu fail to segment the majority of lesions when trained on ISBI challenge dataset. This exhibits the limitation

⁵<https://smart-stats-tools.org/lesion-challenge>

Table 2.10: The consistency of the segmentations for each approach in the cross-dataset setting. The consistency index represents the test-set average of Dice values, each Dice is computed between two segmentations produced by the same method when trained on two different train-sets. Higher values indicate better consistency in the segmentations. The bold values indicate the best result and red values indicate consistency lower than 0.5. * indicates a significantly better segmentation consistency score than the three other approaches, using the two-sided Wilcoxon test.

Test-set		In-house	MSSEG'16	ISBI Test-set		
Train-sets	Dataset 1 vs. Dataset 2	ISBI Train-set MSSEG'16	ISBI Train-set In-house	ISBI Train-set MSSEG'16	In-house MSSEG'16	ISBI Train-set In-house
The consistency of the model predictions when trained on Dataset1 vs. Dataset2	nicMSlesion	0.217	0.514	0.512	0.250	0.177
	DeepMedic	0.602	0.537	0.490	0.602	0.496
	2.5D Tiramisu	0.615	0.726	0.217	0.485	0.460
	DLB	0.647	0.719	0.630*	0.637*	0.649*

of the robustness of these methods to domain shift, especially for 2.5D Tiramisu currently considered as the state-of-the-art approach on the ISBI challenge. Second, DLB detects almost all the lesions in the same conditions. Third, although DeepMedic also detects most of the lesions, it is more prone to false positives compared to the other methods. Finally, when choosing MSSEG'16 as a training dataset, DLB produces the most similar segmentation to expert annotation.

Figure 2.5 represents an image from MSSEG'16 dataset and the segmentation of the different methods when trained on ISBI challenge and In-house datasets. First, when trained on ISBI dataset, the segmentations of 2.5D Tiramisu, DeepMedic, and DLB are more accurate than nicMSlesion segmentation, although all techniques missed a large portion of the central lesion (False Negative) located around the midsagittal plane. These common voxels misclassification can result from the subjectivity of raters between training and testing datasets. Second, when trained on In-house, DLB delineates successfully most of the lesions. Especially in the case of small lesions, DLB missed only one lesion, whereas both nicMSlesion and 2.5 D Tiramissu missed four lesions, and DeepMedic misses two lesions.

Figure 2.6 represents an image from the ISBI challenge and the segmentation of the different methods when trained on MSSEG'16 and In-house datasets. From the four methods, DLB had the most consistent segmentation across different conditions of training domains. In this case, nicMSlesion produced a decent segmentation for this example only when trained on MSSEG'16. Likewise, 2.5D Tiramisu produced better segmentation when trained on In-house than on MSSEG'16. Although DeepMedic is consistent for this case, the produced segmentation was less precise and prone to false positives compared to the other methods.

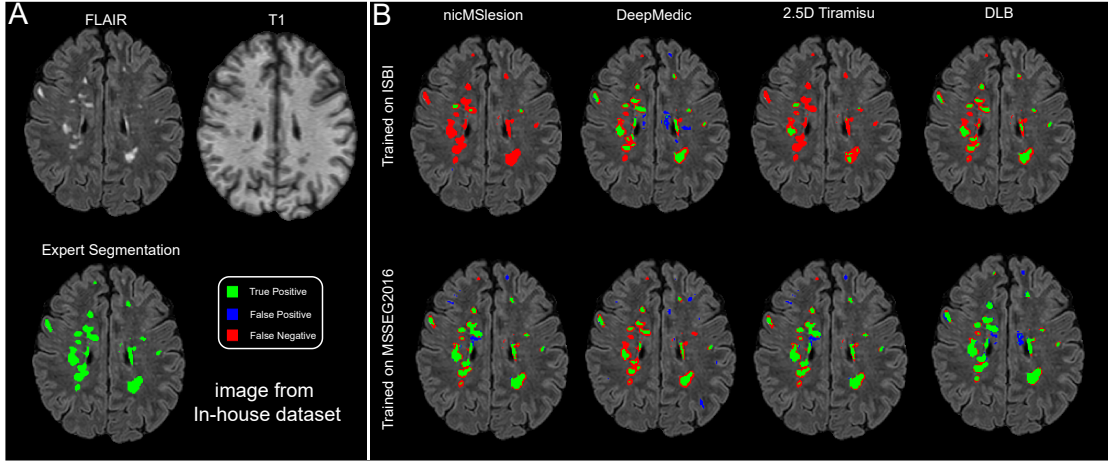


Figure 2.4: Part A (left) axial sections of multi-modal MRI (T1 and FLAIR) from In-house dataset, and its respective expert consensus segmentations for MS lesion segmentation. Part B (right) cross dataset segmentation of the image section shown in Part A. The first and second rows illustrate the segmentations of methods when trained respectively on ISBI dataset, and MSSEG’16 datasets. The first, second, third, and fourth columns represent respectively the segmentations of nicMSlesion, DeepMedic, 2.5D Tiramisu, and DeepLesionBrain.

2.4 Discussion and Conclusion

2.4.1 Discussion

Deep learning-based segmentation models can be prone to generalization failure due to domain shift between training and testing data. Such a domain shift may be caused by hardware and preprocessing diversity, the difference in acquisition protocol or annotation protocol, that results in a difference between the distributions of training and testing datasets. Besides, we also have to acknowledge the subjectivity of raters in training datasets. Indeed, the disagreement between expert segmentations, both in the same dataset and across different datasets, can make it difficult to train a generalizing model. Our experiments showed the limited generalization capability of state-of-the-art approaches, whereas DLB was able to adapt across different domains. Our study emphasizes the importance of cross-dataset validation, particularly when considering the clinical application of machine learning.

DLB uses a group of several separately trained networks, each network is specialized in a particular sub-volume of the brain. In the ablation study (see Tables 2.2 and 2.3), our spatially distributed networks strategy showed better generalization and higher accuracy than using a single model.

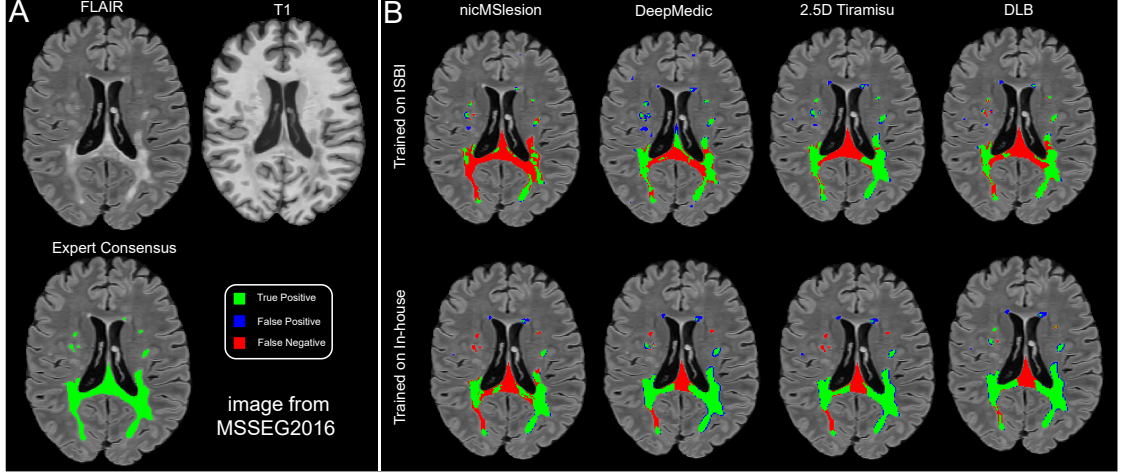


Figure 2.5: Part A (left) axial sections of multi-modal MRI (T1 and FLAIR) from MSSEG'16 dataset, and its respective expert consensus segmentations for MS lesion segmentation. Part B (right) cross dataset segmentation of the image section shown in Part A. The first and second rows illustrate the segmentations of methods when trained respectively on ISBI challenge, and In-house datasets. First, second, third, and fourth columns represent respectively the segmentations of nicMSLesion, DeepMedic, 2.5D Tiramisu, and DeepLesionBrain.

In our work, we considered both specialized local features, and generic global features of MS lesions. The hierarchical specialization learning proposes an alternative to network cascades (*i.e.*, [Valverde et al., 2017]). Instead of using cascades that are prone to error propagation, we suggested a logical hierarchy during learning based on data selection and transfer learning. The ablation study (see Tables 2.2 and 2.3) exhibits the contribution of HSL to accuracy and domain generalization compared to DLB without HSL.

In this chapter, the proposed method was validated using an out-of-domain cross-dataset evaluation. This strategy ensures that the performance obtained is not biased by the training dataset domain information. Indeed, the use of testing and training images from the same domain is questionable and does not reflect the generalization ability. The community should start considering this issue for both the validation and the comparison of proposed methods.

Automated MS lesion segmentation should be able to render the most accurate segmentation with the minimum number of modalities, to be efficiently adopted in clinical conditions and to limit inter-modality dependence. Many experts agree that FLAIR is the most important modality for MS lesion delineation. Moreover, T1 modality can provide complementary information for better white-matter, gray-matter, and cerebrospinal-fluid distinction. FLAIR and T1 are the most avail-

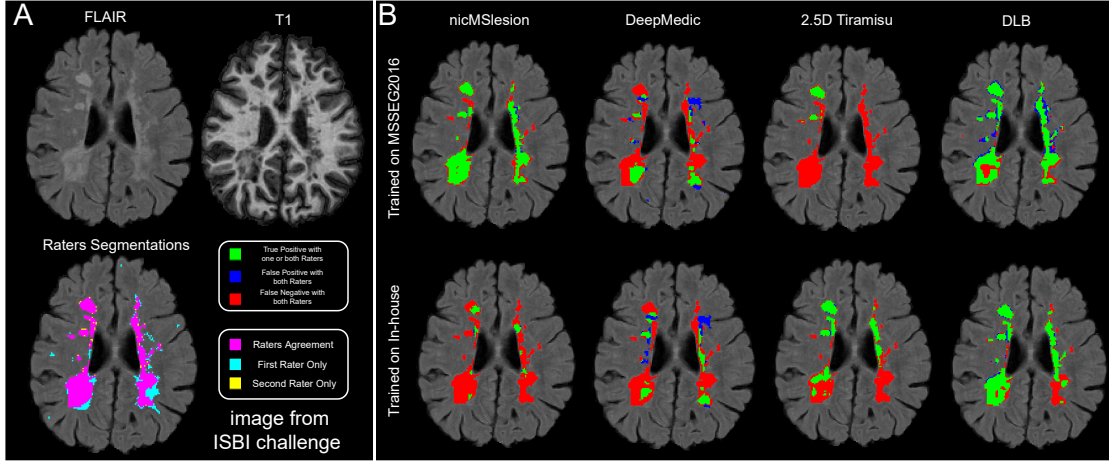


Figure 2.6: Part A (left) axial sections of multi-modal MRI (T1 and FLAIR) from ISBI challenge dataset, and its respective raters segmentations for MS lesion segmentation. Part B (right) cross dataset segmentation of the image section shown in Part A. The first and second rows illustrate the segmentations of methods when trained respectively on MSSEG’16, and In-house datasets. First, second, third, and fourth columns represent respectively the segmentations of nicMSlesion, DeepMedic, 2.5D Tiramisu, and DeepLesionBrain.

able modalities for MS patients and in all MS benchmark datasets. Our method achieved a competitive performance using these two modalities even on unseen domains.

In this chapter, we proposed a novel data augmentation technique to reduce domain shift introduced by the variability of image resolution and quality. IQDA simulates different acquisition conditions to reduce covariate shift. Our ablation study (see Tables 2.2 and 2.3) showed IQDA as a solid contribution to segmentation accuracy and cross-domain generalization. Indeed, while other methods (nicMSlesion, DeepMedic, and 2.5D Tiramisu) involve usual data augmentation (rotation and flipping), such simple strategies failed to ensure good generalization on unseen datasets.

Both domain generalization and adaptation are concerned with reducing dataset bias. The difference between these strategies is that for domain adaptation, some unlabeled data or even a few labeled data from the target domain are exploited to capture properties of the target domain for model adaptation. However, in domain generalization, no samples of any kind are used from the target domain. Domain generalization has been proposed to address the problem of unavailability of target domain samples by leveraging the labeled data to learn a universal representation to generalize for any target domain and without any prior insight from that domain.

In this work, we emphasize on testing the domain generalization of our approach with cross-dataset evaluation. Unlike domain adaptation such as one-shot domain adaptation (*i.e.*, [Valverde et al., 2019]), DLB does not need expert segmentation from the target domain. Our testing conditions draw a clear distinction between training data containing source domains and testing data containing unseen target domains.

In section 2.3.2, we reported that the best performances of DLB have been obtained when using high-resolution 3D FLAIR datasets and multi-rater consensus ground truth for training. The resulting model can render more accurate segmentations for both 2D and 3D image resolution data, even across unseen domains. This observation led us to believe that to efficiently train 3D CNN-based models for domain generalization, it may be desirable to optimize the model using high-resolution training data.

With current available hardware, it is unfeasible to exploit 3D CNNs with equivalent depth and kernel size as state-of-the-art 2D CNNs. Consequently, many neuroimaging automated pipelines are still using 2D CNNs despite processing 3D data. Our results suggest that using multiple compact networks can approximate a larger and more stable model since the sum of features extracted by the group of specialized networks and the features of a hypothetical big network may be equivalent in terms of relevant information for MS lesion segmentation. In our work, we have chosen to break down the complexity of the task spatially, based on the sub-volume division of the whole brain volume. One other advantage of this distribution is the ability to train networks in parallel since network weights and images of each region are independent. It is possible to use several GPUs for parallel training.

Our full pipeline including the preprocessing, MS lesion segmentation with DLB, and an easy-to-read report is available on our repository ⁶.

2.4.2 Conclusion

DeepLesionBrain is a deep learning framework for MS lesion segmentation designed for domain generalization. First, we use a spatially distributed strategy of multiple compact 3D CNNs with large overlapping receptive fields, to produce consensus-based segmentation robust to domain shift. Our method is trained using hierarchical specialization learning to efficiently incorporate both generic and specialized features. Second, we propose a novel image quality data augmentation to increase training data variability in a realistic way. Finally, we use only T1 and FLAIR modalities to propose a method compatible with a large number of datasets.

⁶<https://github.com/volBrain/DeeplesionBrain>

The ablation study showed the impact of each contribution on segmentation accuracy and domain generalization. The out-of-domain cross-dataset testing is suggested as an alternative for method evaluation in areas that are sensitive to domain bias (*i.e.*, medical imaging). Our validation showed the generalization ability of our method and its robustness to domain shift. We also proved experimentally that DLB produces consistent segmentations compared to other state-of-the-art approaches regardless of the training data domain.

On the same topic, we proposed an extension of this work using semi-supervised learning (see Appendix A). The method uses unlabeled data to compensate for the scarcity of annotated images and the lack of method generalization to unseen domains. The method combines consistency regularization and pseudo-labeling in a complementary fashion and uses a proximity graph to select data from the easiest to the more difficult ones, therefore limiting confirmation bias.

Chapter 3

New MS Lesions Segmentation and Detection

This chapter corresponds to the following publication:

Reda Abdellah Kamraoui, Boris Mansencal, José V Manjon, and Pierrick Coupé. Longitudinal detection of new MS lesions using deep learning. *Frontiers in Neuroscience* [2].

The detection of new multiple sclerosis (MS) lesions is an important marker of the evolution of the disease. The applicability of learning-based methods could automate this task efficiently. However, the lack of annotated longitudinal data with new-appearing lesions is a limiting factor for the training of robust and generalizing models. In this work, we describe a deep-learning-based pipeline addressing the challenging task of detecting and segmenting new MS lesions. First, we propose to use transfer-learning from a model trained on a segmentation task using single time-points. Therefore, we exploit knowledge from an easier task and for which more annotated datasets are available. Second, we propose a data synthesis strategy to generate realistic longitudinal time-points with new lesions using single time-point scans. In this way, we pretrain our detection model on large synthetic annotated datasets. Finally, we use a data-augmentation technique designed to simulate data diversity in MRI. By doing that, we increase the size of the available small annotated longitudinal datasets. Our ablation study showed that each contribution lead to an enhancement of the segmentation accuracy. Using the proposed pipeline, we obtained the best score for the segmentation and the detection of new MS lesions in the MSSEG2 MICCAI challenge.

3.1 Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system. The pathology is characterized by inflammatory demyelination and axonal injury, which can lead to irreversible neurodegeneration. The disease activity, such as MS lesions, can be observed using magnetic resonance imaging (MRI). The detection of new MS lesions is one of the important biomarkers that allow clinicians to adapt the patient’s treatment and assess the evolution of this disease.

Recently, the automation of single time-point MS lesion segmentation has shown encouraging results. Many techniques showed performance comparable to clinicians in controlled evaluation conditions (see [Carass et al., 2017], and [Commowick et al., 2016]). These methods use a single time-point scan to segment all appearing lesions at the time of the image acquisition. However, these cross-sectional techniques are not adapted to the longitudinal detection of new lesions. Indeed, using these methods requires repeatedly running the segmentation process for each time-point independently to segment MS lesions before detecting new ones. Unlike the human reader, these methods are not designed to jointly exploit the information contained at each time point. Consequently, single-time MS lesion segmentation methods performance is not optimal for the detection of new lesions between two time-points. Moreover, inconsistencies may appear between segmentations of both time-points since they are processed independently.

To specifically address this detection task using both time-points at the same time, some detection methods have been proposed. In one of the earliest works, [Bosc et al., 2003] used a nonlinear intensity normalization method and statistical hypothesis test methods for change detection. [Elliott et al., 2013] used a bayesian tissue classifier on the time-points to estimate lesion candidates followed by a random-forest-based classification to refine the identification of new lesions. [Ganiler et al., 2014] used image subtraction and automated thresholding. [Cheng et al., 2018] integrated neighborhood texture in a machine learning framework. [Salem et al., 2018] trained a logistic regression model with features from the image intensities, the image subtraction values, and the deformation field operators. [Schmidt et al., 2019] used lesion maps of different time-points and FLAIR intensities distribution within normal-appearing white matter to estimate lesion changes. [Krüger et al., 2020] used a 3D convolutional neural network (CNN) where each time-point is passed through the same encoder. Then, the produced feature maps are concatenated and fed into the decoder.

Training learning-based methods for the task of new lesions detection require a dataset specifically designed for the task. The most obvious form of the training data would be a longitudinal dataset of MS patients (with two or more successive time-points) with new appearing lesions carefully delineated by experts in the field. However, the construction of such a dataset is very difficult. To begin, new

lesions may take several months or even years to appear and be visible in a patient’s MR image. Moreover, a time-consuming and costly process is necessary for several experts to annotate new lesions from the two time-points and to obtain an accurate consensus segmentation. Although the organizers of the MICCAI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge (MSSEG2-challenge [MICCAI, 2021]) provided such a dataset, the training set is severely impacted by class imbalance (see Section 3.2.5.3 for more details) due to the difficulty of finding new lesions in the follow-up scan. This under-representation of new lesions in longitudinal datasets is limiting the training of state-of-the-art deep learning algorithms from scratch on this complex task. Besides, achieving generalizing results on unseen domains (see [Mårtensson et al., 2020, Bron et al., 2021, Omoumi et al., 2021]) may requires more data diversity.

Several studies tackled the problem of training data scarcity. First, transfer learning is a strategy used to create high-performance learners trained with more widely available data from different domains when the target domain/task data are expensive or difficult to collect (see [Torrey and Shavlik, 2010, Weiss et al., 2016]). Second, synthetic data generation is performed by using a model able to simulate realistic artificial data that can be used during training (see [Tremblay et al., 2018, Tripathi et al., 2019, Khan et al., 2021]). Third, data-augmentation is a set of techniques used to handle the variability in real-world data by enhancing the size and quality of the training dataset (see [Shorten and Khoshgoftaar, 2019]). Recently, [Zhang et al., 2020b] showed that applying extensive data augmentation during training also enhances the generalization capability of the methods.

In this chapter, we propose an innovative strategy integrating these three strategies into a single pipeline for new MS lesion segmentation to tackle data rarity for our task. First, we use transfer-learning to exploit the larger and more diverse datasets available for the task of single-point MS lesion segmentation which does not require longitudinal data. Second, we propose a novel data synthesis technique able to generate two realistic time-points with new MS lesions from a single FLAIR scan. Third, we use a data-augmentation technique to simulate a large variety of artifacts that may occur during the MRI acquisitions. This technique aims to enhance both the variability and size of the training data and to improve the generalization of our model.

3.2 Method and Material

3.2.1 Method Overview

To deal with data rarity for new MS lesion segmentation, we proposed a three stage pipeline as shown in Fig. 3.1. **In Stage One**, an encoder-decoder network

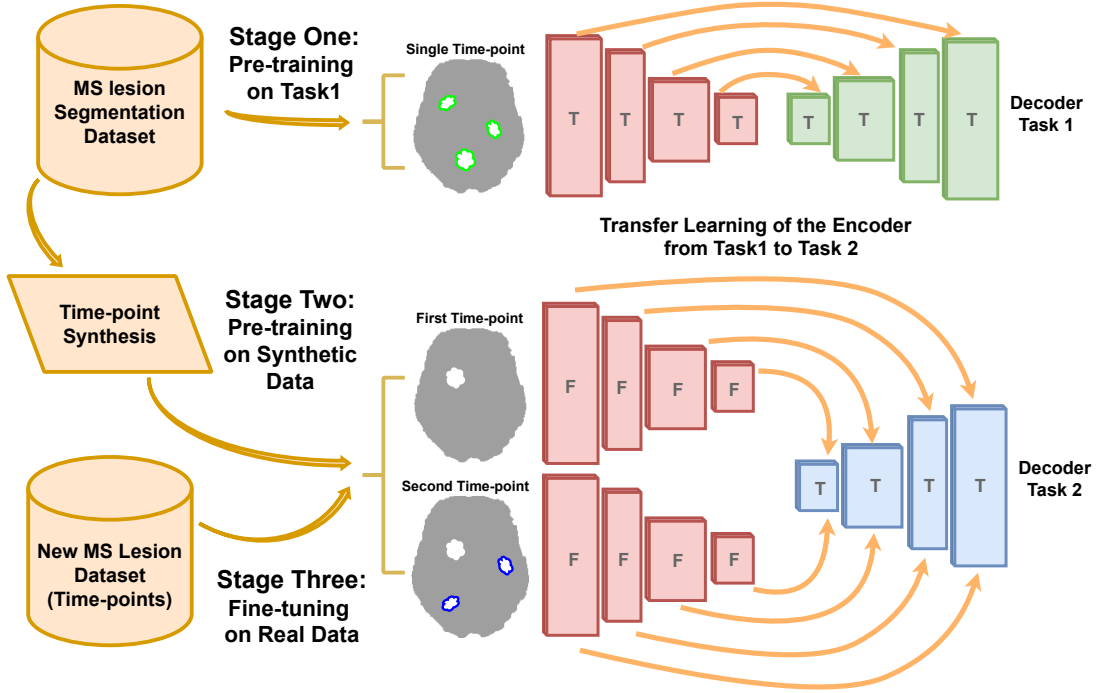


Figure 3.1: The pipeline of our new MS lesion segmentation method. The three stages include: First, the pre-training on the task of single-time-point MS lesion segmentation (Task 1). Second, pre-training on the task of new MS lesions segmentation (Task 2) with synthetic data. Third, fine-tuning the model with real data. The encoder weights are trained (T) in Stage One and frozen (F) in Stage Two and Stage Three.

is trained on the task of single time-point MS lesions segmentation. This step aims to train the encoder part of the network to extract relevant features related to MS lesions that can be used in the next steps. Stage One enables to indirectly use large datasets dedicated to single time-point MS lesion segmentation for the task of new lesions segmentation. This stage is detailed in Section 3.2.2. **In Stage Two**, the new lesions segmentation model composed of the previous task encoder is pretrained with synthetic data. To this end, we trained external models able to generate two realistic time-points from a single image also taken from single time-point MS datasets. It combines the effects of lesion inpainting and lesion generating models to simulate the appearance of new lesions. This strategy is detailed in Section 3.2.3. **In Stage Three**, the decoder is fine-tuned with real longitudinal data from the new MS lesion training-set of the MSSEG2 MICCAI challenge.

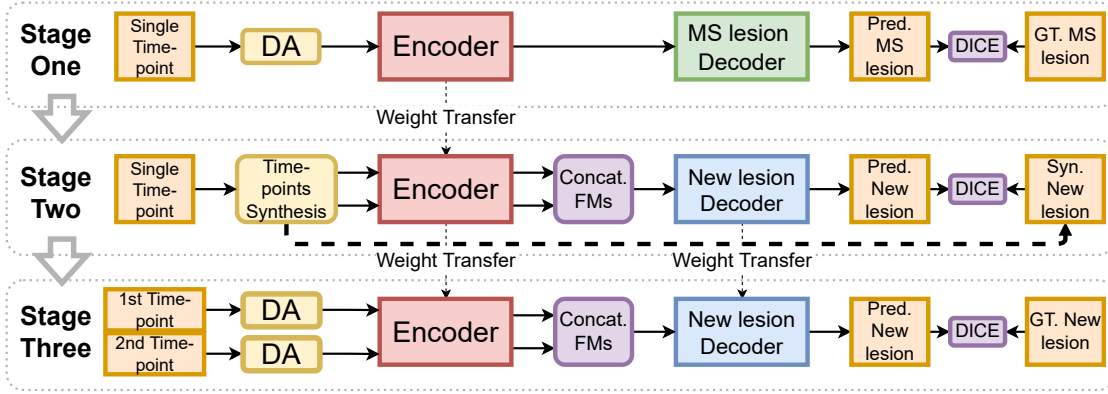


Figure 3.2: The diagram represents our training method. Input images are augmented with the proposed method (DA). The encoder trained in Stage One is used in Stage Two and Stage Three to extract feature maps (FMs) of the two-time points. The aggregation block (Concat. FMs) is used to combine features

3.2.2 Transfer-learning from single time-point MS lesion segmentation task

The encoder used for new MS lesion segmentation is first trained on single time-point lesion segmentation (see Fig.3.2, from Stage One to Stage Two). This choice is motivated by two reasons. First, we consider that datasets for MS lesion segmentation with lesion mask segmentation by experts are more diverse and larger than available datasets for new lesions segmentation (which requires a longitudinal study). Second, the task of MS lesion segmentation is tightly close to the one of new MS lesion segmentation. By learning to segment lesions, the model implicitly learns the concept of a lesion, either the lesion is considered new or was already existing in the first time-point. To conclude, since there is a proximity between the two tasks, there is likely a gain from exploiting the large amount of training data of the first task to improve the second task’s performance.

3.2.2.1 Model Architecture Design

Our method is based on the transfer learning from the task of “Single time-point MS lesion segmentation” to the task of “new lesions segmentation from two time-points”. Thus two different architectures are used but with the same building blocks for each task. For the first task, a 3D U-Net shape architecture is used, as shown in Fig. 3.3 Part A. This kind of architecture has been very effective and robust for MS lesion segmentation [2, Isensee et al., 2021]. It is composed of an encoder and a decoder linked with one another by skip connections.

For the second task, a siamese-encoder followed by a single decoder is used, as

shown in Fig.3.3 Part B. The shared-weights encoders are chosen to extract the same set of features from both time points. Then, these features resulting from the different levels of both encoder paths are aggregated (see Fig. 3.3 Part B). The aggregation module is composed of concatenation and a convolution operation. Feature maps are first concatenated by channels (*i.e.*, result channel size is twice the original size), then the convolution operation aggregates the information back to the original channel size. Finally, the aggregated features are passed through the decoder.

3.2.3 Time-points Synthesis

The data synthesis method is based on the simulation of new MS lesions between two time-points using single time-point FLAIR images. As shown in Fig. 3.4, our pipeline generates “on the fly” synthetic 3D patches that represent longitudinal scans of the same patient with an evolution in their lesion mask. The synthetic data is generated in three steps. In the first step, a 3D FLAIR patch and its MS lesion segmentation mask are randomly sampled from different MS lesion segmentation datasets (see Section 3.2.5.1). Then, the patch and lesion mask are randomly augmented with flipping and rotations. A copy of the FLAIR patch is performed to represent the two time-points. Then, both identical patches are altered with the described data augmentation (see Section 3.2.4) to differentiate the two patches. At this point, the lesion masks of the two synthetic time-points are still identical. Thus, there are no new lesions. In the second step, a connected component operation is used to separate each independent lesion from the lesion mask. Each lesion is either inpainted (*i.e.*, removed) from one of the two time-points or both of them, or it can be kept in both of the time-points. The lesion inpainting model is used to inpaint the lesion region with hallucinated healthy tissue (see Section 3.2.3.1). Next, the new lesion mask is constructed from lesion regions that have been kept in the second time-point but not the first one. In the third step, the lesion generator model is used to simulate new synthetic lesions at realistic locations (using white/gray matter segmentation and a probabilistic distribution of MS lesions on the brain in the MNI space). Synthetic lesions are generated for one of the time-points or both of them (see Section 3.2.3.2). Similar to the previous step, the new lesion mask is updated to include only the generated lesions on the second time-point.

3.2.3.1 Lesion Inpainting Model

The lesion inpainting model is trained, independently and priorly to our proposed pipeline, with randomly selected 3D FLAIR patches which do not contain MS lesions or white matter hyperintensities. Similarly to [Manjón et al., 2020], A 3D

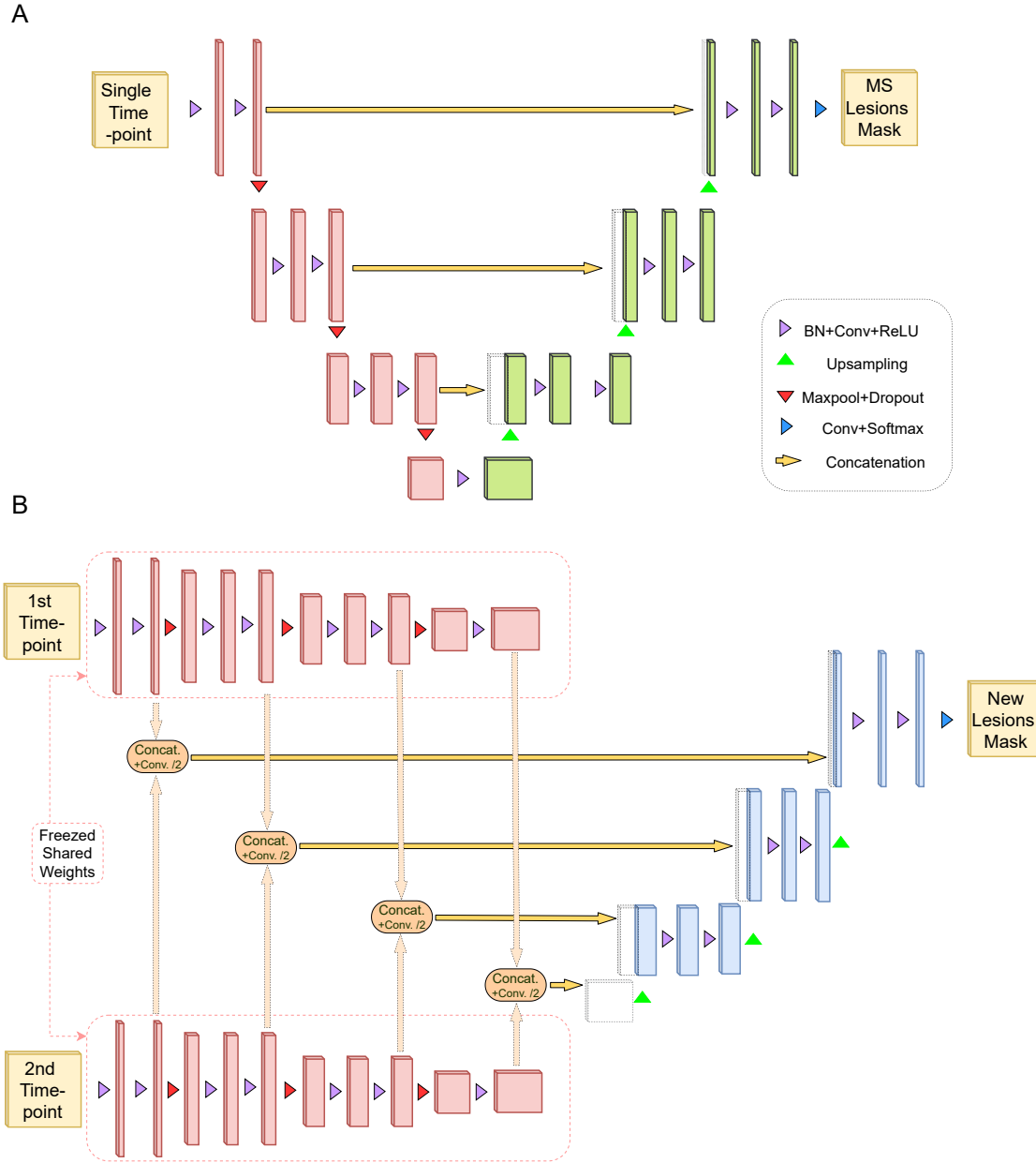


Figure 3.3: Part A represents U-Net like architecture composed of an encoder (in red) and a decoder for the task of MS lesion segmentation (in green). This task requires a single time-point as input and produce the MS lesion mask. Part B shows a siamese-encoder (in red) to extract the same sets of features from the two time-points. Same-level features are aggregated with a combination module and are forwarded to a decoder for the task of new lesions segmentation (in blue).

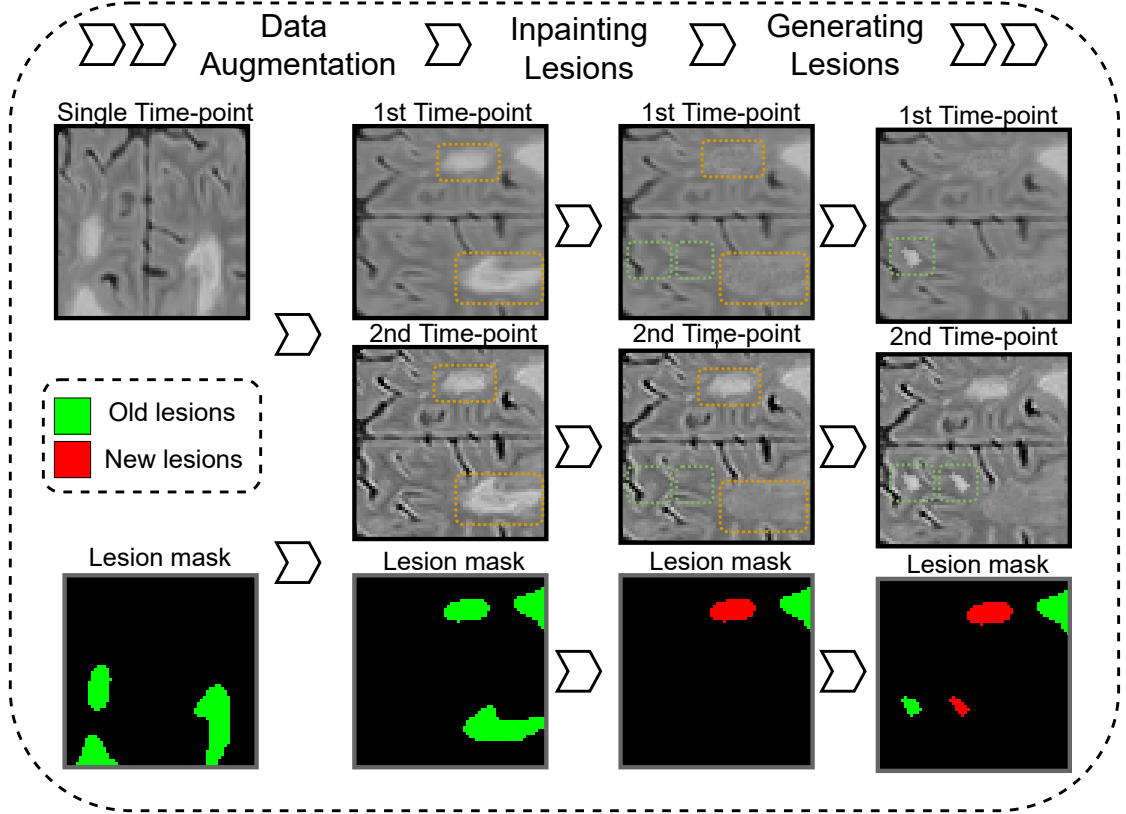


Figure 3.4: Synthetic time points with new MS lesion generation pipeline. Dashed orange and green rectangles on images represent areas where lesion are inpainted or generated

U-Net network is optimized to reconstruct altered input images. Specifically, the input patch is corrupted with Gaussian noise (*i.e.*, with a mean and a standard deviation of the image intensities) in lesion-like areas at random locations. When the model is trained, it can be used to synthesize healthy regions in lesion locations that are replaced with random gaussian (see [Manjón et al., 2020] for details).

3.2.3.2 Lesion Generator Model

The lesion generator is trained before our proposed pipeline to simulate realistic lesions. The generator is a 3D U-Net network with two input channels and one output channel. The first input channel receives an augmented version of 3D FLAIR patches containing MS lesions where lesions are replaced with random noise. The second input channel receives the MS lesion mask of the original 3D FLAIR patch. The output channels predict the original 3D FLAIR patch with lesions. Thus, the trained model can simulate synthetic MS lesions from a 3D

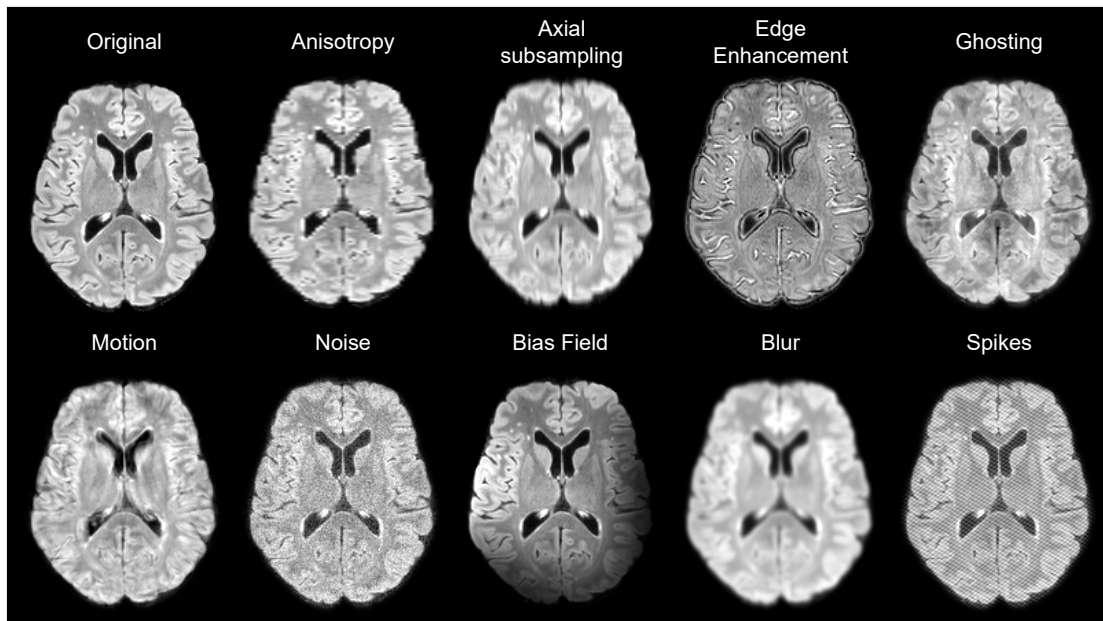


Figure 3.5: Examples of data augmentation applied on FLAIR images

patch of FLAIR and its corresponding lesion mask.

3.2.4 Data Augmentation

The quality of the MRI greatly varies between datasets. The quality of the images depends on several factors such as signal-to-noise ratio, contrast-to-noise ratio, resolution, or slice thickness. Since our training set is limited, it does not reflect the diversity of real-world images. To make our training stages robust to the large variety of artifacts that may occur during the MRI acquisitions, an extensive Data Augmentation (DA) is used (see "DA" in Fig. 3.2 and "Data Augmentation" in Fig. 3.4). Such DA technique also helps to better oversample the scarce samples with new lesions (see 3.2.5.3).

We use an improved version of the data augmentation strategy proposed in [2], which simulates MRI quality disparity. During training, we simulate “on the fly” altered versions of 3D patches. We randomly introduce a set of alterations in the spatial and frequency space (k-space): Blur, edge enhancement, axial subsampling distortion, anisotropic downsampling, noise, bias-field variation, motion effect, MRI spike artifacts, and ghosting effect. Figure 3.5 shows augmentation samples.

For the blur, a gaussian kernel is used with a randomly selected Standard Deviation (SD) ranging between $[0.5, 1.75]$. For edge enhancement, we use unsharp

Table 3.1: Summary of the used datasets. For each dataset, the object count (Obj. Count) and the total volume (Tot. Vol. cm^3) represent respectively the total number and the total volume in cm^3 of lesions or new lesions (depending on the task).

Task	Dataset	# Patients	# Time-point	# Raters	Obj. Count	Tot. Vol. (cm^3)	Clinical Site / Scanners
MS Lesion Segmentation	ISBI	5	4-5	2	514	243	Single-site
	MSSEG' 16	15	1	7	512	367	Multi-site: 3 sites
	In-house	43	1	2	2391	1313	Multi-site
New MS Lesion Segmentation	MSSEG2	40	2	4	123	23	Multi-site:
	Training-set						15 MRI scanners
	MSSEG2 Test-set	60	2	4	174	60	(GE scanners only in Test-set)

masking with the inverse of the blur filter. For axial subsampling distortion, we simulate acquisition artifacts that can result from the varying slice thickness. We use a uniform filter (a.k.a mean filter) along the axial direction with a size of $[1 \times 1 \times sz]$ where $sz \in 2, 3, 4$. For anisotropic downsampling, the image is downsampled through an axis with a random factor ranging between $[1.5, 4]$ and upsampled back again with a B-spline interpolation. For noise, we add to the image patch a Gaussian noise with 0 mean and an SD ranging between $[0.02, 0.1]$. Bias-field variation is generated using the work of [Sudre et al., 2017] that considers the bias field as a linear combination of polynomial basis functions. Motion effect has been generated based on the work of [Shaw et al., 2018]. The movements are simulated by combining in the k-space a sequence of affine transforms with random rotation and translation in the ranges $[-5, 5]$ degrees and $[-4, 4]mm$ respectively. Both MRI spike artifacts and ghosting effect have been generated with the implementation of [Pérez-García et al., 2021].

3.2.5 Data

Different datasets are used for the training and validation of the two tasks (see Table 3.1).

3.2.5.1 Single Time-point Datasets

For time-points synthesis (see 3.2.3) and encoder pretraining (see 3.2.2), we jointly used three datasets containing cross-sectional FLAIR and lesion masks, corresponding to a single point in time for each subject. First, the ISBI [Carass et al., 2017] training-set contains 21 FLAIR images with expert annotation done by two raters. Although the dataset is composed of longitudinal time-points from 5 patients, the provided expert annotations focus on the lesion mask of each time-point independently from the others and do not provide new lesion masks. Thus, we use

the 21 images independently. Second, the MSSEG'16 training-set [Commowick et al., 2016] contains 15 patients from 3 different clinical sites. Each FLAIR image is along with a consensus segmentation for MS lesions from seven human experts. Third, our in-house [Coupé et al., 2018] dataset is composed of 43 subjects diagnosed with MS. The images were acquired with different scanners and multiple resolutions and their lesion masks have been obtained by two human experts.

All images were pre-processed using the lesionBrain pipeline from the volBrain platform [Manjón and Coupé, 2016]. First, it includes image denoising [Manjón et al., 2010]. Second, an affine registration to MNI space is performed using the T1w modality, then the FLAIR is registered to the transformed T1w. Skull stripping and bias correction have been performed on the modalities, followed by the second denoising. Finally, the intensities have been normalized with kernel density estimation.

3.2.5.2 Two Time-points Datasets

The dataset provided by the MSSEG2-challenge [MICCAI, 2021] is used to train our method. The challenge dataset features a total of 100 MS patients. For each patient, two 3D FLAIR sequence time-points have been acquired spaced apart by a 1 to 3 years period. The dataset has been split into 40 patients for training and 60 patients for testing. A total of 15 different MRI scanners were used for the acquisition of the entire dataset. However, all images from GE scanners have been reserved only for the testing set to see the generalization capability of the algorithms. Reference segmentation on these data was defined by a consensus of 4 expert neuroradiologists.

For preprocessing, the challenge organizers proposed a docker ¹ built with the Anima scripts. It includes bias correction, denoising, and skull stripping. In addition, we added a registration step to the MNI space using a FLAIR template (*i.e.*, the training and inference are performed in the MNI space, then the segmentation masks are transformed-back to the native space for evaluation).

Before challenge day, the testing set (the 60 patients) was not publicly available. Thus to test our methods (see Section 3.3.1.1), we defined an internal validation subset from the 40 challenge training data. From the 40 patients, 6 cases containing confirmed new lesions were kept out from the training-set and were used as an internal test-set. For the challenge evaluation (see Section 3.3.2), the model submitted to the challenge organizers was trained on the entire MSSEG2 training-set.

¹<https://github.com/Inria-Empenn/lesion-segmentation-challenge-miccai21/>

3.2.5.3 Dataset Class Imbalance

Anomaly detection/segmentation tasks, such as MS lesion segmentation, suffer from class imbalance where the positive class is scarce (see [Johnson and Khoshgof-taar, 2019]). Herein, the MSSEG2-challenge [MICCAI, 2021] dataset is composed of 100 patients (40 for training and 60 for test) and all the MS Lesions Segmentation datasets combined account for 64 patients and 79 images. Therefore, the number of image is similar. However, the class imbalance is highly different when evaluating the class imbalance using the number of objects to detect/segment (which represent MS lesions for the first task and new lesions for the second one) and their total volume for each dataset (see Table 3.1). Indeed, we see that the MSSEG2-challenge datasets (especially training-set) suffer from more severe under-representation of the positive class. Consequently, it will be more difficult to train a model for New MS lesion segmentation than for the task of single time-point MS lesion segmentation. Furthermore, it shows that MS lesion segmentation datasets could significantly enrich the training of New MS lesion segmentation models.

3.2.6 Implementation Details

First, all models are trained on 3D image patches of size $[64 \times 64 \times 64]$. For the two time-points new lesion model, an ensemble of 5 networks (different training/validation data-split) is used. During inference, the consensus (prediction average) of the ensemble segmentation is taken. For each voxel, the two classes output probabilities of the 5 networks are averaged, and the class with the highest probability is picked (new lesion voxel or not).

Second, the Dice-loss (soft DICE with probabilities as continuous values) is used as a loss function for the training of the single time-point MS lesion segmentation and the two time-points new lesion models. The mean-squared error is used as a loss function to train time-point synthesis models (inpainting and lesion generator models).

Finally, the experiments have been performed using PyTorch framework version 1.10.0 on Python version 3.7 of Linux environment with NVIDIA Titan Xp GPU 12 GB RAM. All models were optimized with Adam [Kingma and Ba, 2014] using a learning rate of 0.0001 and a momentum of 0.9.

3.2.7 Validation Framework

3.2.7.1 Evaluation Metrics

The assessment of a segmentation method is usually measured by a similarity metric between the predicted segmentation and the human expert ground truth.

First, we use several complementary metrics to assess segmentation performance. Namely, we use the Dice similarity coefficient, the Positive Predictive Value (PPV or the precision), true positive rate (TPR, known as recall or Sensitivity).

Second, recent works (*i.e.*, [Commowick et al., 2018]) question the relevance of classic metrics (Dice) compared to detection metrics, which are used for MS diagnostic and clinical evaluation of the patient evolution. Thus, in addition to the voxel-wise metrics, we also use lesion-wise metrics that focus on the lesion count. We use the Lesion Detection F1 (LesF1) score defined as

$$LesF_1 = \frac{2 \times S_L \times P_L}{(S_L + P_L)}, \quad (3.1)$$

where S_L is lesion sensitivity, *i.e.*, the proportion of detected lesions and P_L is lesion positive predictive value, *i.e.*, the proportion of true positive lesions. For result harmonization with challenge organizers and participants, the same evaluation tool is used, *i.e.*, animaSegPerfAnalyzer [Commowick et al., 2018]. All lesions that are smaller in size than $3mm^3$ are removed. For S_L , only ground-truth lesions that overlap at least 10% with segmented volume are considered positive. For a predicted lesion to be considered positive for P_L it has to be overlapped by at least 65% and do not go outside by more than 70% of the volume.

Finally, to jointly consider the different metrics (*i.e.*, segmentation and detection performance), it would be convenient to aggregate them into a single score. Thus, we propose the average of DICE and $LesF_1$ (Avg. Score) as an aggregation score for comparing different methods.

3.2.7.2 Statistical Test

To assert the advantage of a technique obtaining the highest average score, we conducted a Wilcoxon test (*i.e.*, paired statistical test) over the lists of metric scores. The significance of the test is established for a p -value below 0.05. In the following tables, * indicates a significantly better average score when compared with the rest of the other approaches.

3.3 Results

Several experiments were conducted on the model trained with the proposed training method, which we will be referring to as Longitudinal DLB, including an ablation study and the comparison with state-of-the-art methods in competition during the challenge evaluation.

Table 3.2: The internal validation results for the ablation study. ✓ and ✗ symbolize using or not each contribution. Bold values indicate the best result for a metric and * indicates that the advantage is statistically significant (Wilcoxon test).

Transfer Learning	Time-point Synthesis	Data augm.	Avg. Score	DICE	$LesF_1$	TPR	PPV
✓	✓	✓	0.543*	0.514*	0.573*	0.500*	0.546
✓	✗	✗	0.483	0.480	0.486	0.461	0.532
✗	✓	✗	0.501	0.461	0.541	0.384	0.602*
✗	✗	✓	0.477	0.464	0.488	0.406	0.565
✗	✗	✗	0.469	0.449	0.489	0.413	0.534

3.3.1 Internal Validation

3.3.1.1 Ablation Study

To evaluate each contribution of our training pipeline, Table 3.2 compares our full method with a baseline and other variations of our method on the internal validation dataset. The baseline in this experiment was trained with real time-points only and by using a classic data augmentation composed of orthogonal rotations and mirroring.

First, when using only transfer learning on the top of the baseline, we measured an increase in DICE and TPR compared to the baseline but approximately the same $LesF_1$ and PPV. Second, when using only time-point synthesis pretraining on the top of the baseline, we obtained a significantly higher $LesF_1$ compared to the baseline and an increase in DICE. This variation also obtained the highest PPV at the expense of the lowest TPR. Third, when comparing the use of the proposed data augmentation, we see an increase in DICE and PPV but approximately the same $LesF_1$. Finally, when combining the transfer learning, time-point synthesis pre-training, and the proposed data-augmentation, we obtained the highest Avg. Score, DICE, $LesF_1$, and TPR.

3.3.1.2 The Impact of Longitudinal Dataset Size

Figure 3.6 shows the performance of our method when trained with different longitudinal dataset sizes. From the 34 patients available for the training with two time-points in Internal Validation settings (refer to 3.2.5.2), we tested the performance of our model when training on 34, 36, 17, 8, and 0 patients. In the case of 0 patients, our method performance was obtained using synthetic data only (*i.e.*, Stage Two where only cross-sectional MS segmentation databases were used as described in Table 3.1). For the rest of the experiments, the reported number of patients with two time-points were used for the fine-tuning step (*i.e.*, Stage Three).

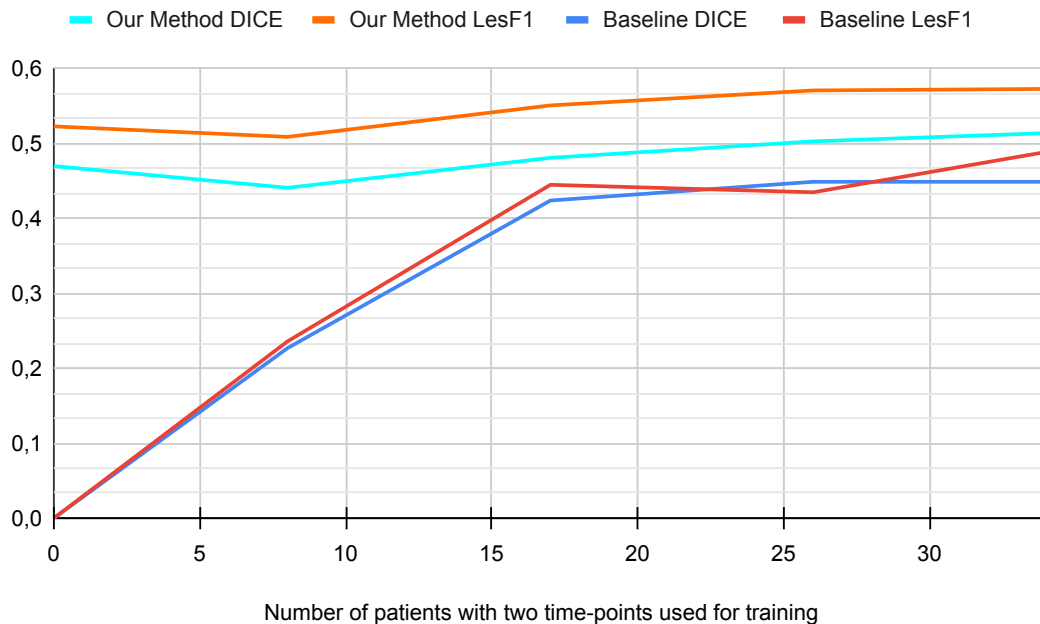


Figure 3.6: The performance in the internal validation of our method and the baseline based on the number of patients used for training (from MSSEG2 Training-set)

First, for the baseline version (*i.e.*, with neither pre-training nor data augmentation), the graph can be separated into two phases. From 0 to 17 patients, the graph shows an increase in both metrics. From 17 to 34 patients, metrics of baseline versions reach a plateau. Since the baseline is trained from scratch, its performance improves with the increase in dataset size. However, the performance increase is less significant for the second phase since it is more difficult to improve metrics when approaching their optimal value.

Second, for our method, the graph shows two phases. From 0 to 8 patients, the performance decreases slightly. From 8 to 34 patients, the graph shows a slow increase in metrics until plateauing. Since we use transfer learning and pretraining on synthetic data for our method, its performance does not depend only on the number of patients from MSSEG2 Training-set. The drop in performance in the first phase can be explained by the fact that using 8 patients for fine-tuning is less effective than using the model trained on synthetic data only.

3.3.2 Challenge Evaluation

To evaluate our method on the challenge dataset, Table 3.3 compares it to the leader-board state-of-the-art methods. Results of the top performing methods

were reported from challenge-day results.

Besides the top-performing methods, Table 3.3 also includes the expert raters performance to give an insight into the human performance. Their performance is measured compared to each other, contrary to the top methods that are evaluated using consensus segmentation. Raters x Vs. y means that we evaluate the performance of rater x when considering rater y segmentations as ground truth. Indeed, we consider that such a strategy can be more meaningful than the consensus segmentation in our case since the expert consensus already encodes the raters segmentation and thus is unfair when comparing to other strategies that did not participate in the consensus.

First, from the Top 5 best-performing methods, LaBRI-IQDA [9] (our team’s submission during the challenge-day) obtained the best score during the challenge. This method was similar to the proposed baseline with data augmentation. Second, Longitudinal DLB (results obtained after challenge-day) obtained the highest $LesF_1$ and Average score. Moreover, these both scores are significantly better than all the listed state-of-the-art methods. The DICE score obtained by MedICL was not significantly better than the one obtained by our method. Third, all but one (Empenn) leader-board automatic method obtained better DICE than raters segmentation. Our proposed method, LaBRI-IQDA, and MedICL even surpassed all raters in Average Scores.

Figure 3.7 shows the segmentation of new lesions of our proposed method. As a ground-truth reference, we compare the segmentation with the consensus segmentation of raters. We also compare each rater segmentation against their consensus. From the five segmentation, we see that our segmentation is the most accurate with the consensus. Each of the human experts Rater 2, Rater 3, and Rater 4 missed one or multiple lesions when segmenting this sample. While Rater 1 did not miss any lesions, we see that our segmentation is the closest to the consensus compared to his/her.

Overall, our method obtained the best result in the MSSEG2 challenge evaluation (during the challenge and after). Moreover, the result of the experiments showed that our segmentation is objective and can produce more accurate segmentations than human raters.

3.4 Discussion

The transfer-learning from single time-point MS lesion segmentation task is an effective method to train the model for the task of two time-points new MS lesion segmentation even with a small dataset. Indeed, it enables to exploit the large available MS cross-sectional datasets compared to longitudinal datasets. In our case, the encoder for the first task was compatible with the siamese-encoder of

Table 3.3: Results of MSSEG2-challenge [MICCAI, 2021] evaluation. From top to bottom, the table shows the challenge raters agreement on the segmentation compared to each other, the leader-board results of the challenge-day top methods, and the result of the method described in this chapter (obtained after challenge-day). For automatic methods, bold values indicate the best result for a metric and * indicates that the advantage is statistically significant (Wilcoxon test).

	Experiment	Avg. Score	DICE	$LesF_1$
	Raters 1 Vs. 2	0.466	0.426	0.507
	Raters 1 Vs. 3	0.499	0.434	0.564
	Raters 1 Vs. 4	0.434	0.382	0.486
Challenge-day	LaBRI-IQDA [9]	0.507	0.498	0.515
	MedICL [Zhang et al., 2021b]	0.503	0.506	0.5
	SNAC [Cabezas et al., 2021]	0.496	0.484	0.513
	Mediaire-B [Dalbis et al., 2021]	0.489	0.436	0.541
	Empenn [Masson et al., 2021]	0.478	0.423	0.532
	Longitudinal DLB	0.523*	0.495	0.550*

the second task and thus was used to extract MS-relevant features from the two time-points. Additionally, we used a learnable aggregation module for time-points feature combination. Besides, by freezing the encoder weights after the transfer-learning from the first to the second task, we ensure that the extracted features in the second task are dataset-independent from the second task dataset (smaller dataset). This independence ensures that the high performance of the proposed method is stable and generalizing.

Longitudinal time-points synthesis is an original approach on how to augment data diversity. It can be extended to other change detection tasks where longitudinal data are hard to acquire. According to the results of our experiments, this strategy turns out to be very effective when used as pretraining. Indeed, when the model is first pretrained with time-point synthesis, it is subject to a wider range of diversity, which aims to constrain the model to extract more generalizing features.

The proposed data augmentation method is an effective technique to make our learning process less dependent on MRI quality and acquisition artifacts. It simulates different acquisition conditions to enhance generalization and helps to better over-sample the available new lesions examples. Our data-augmentation comparison (see Table 3.2) showed the proposed augmentation method contributes to segmentation accuracy in both internal validation and challenge evaluation (*i.e.*,

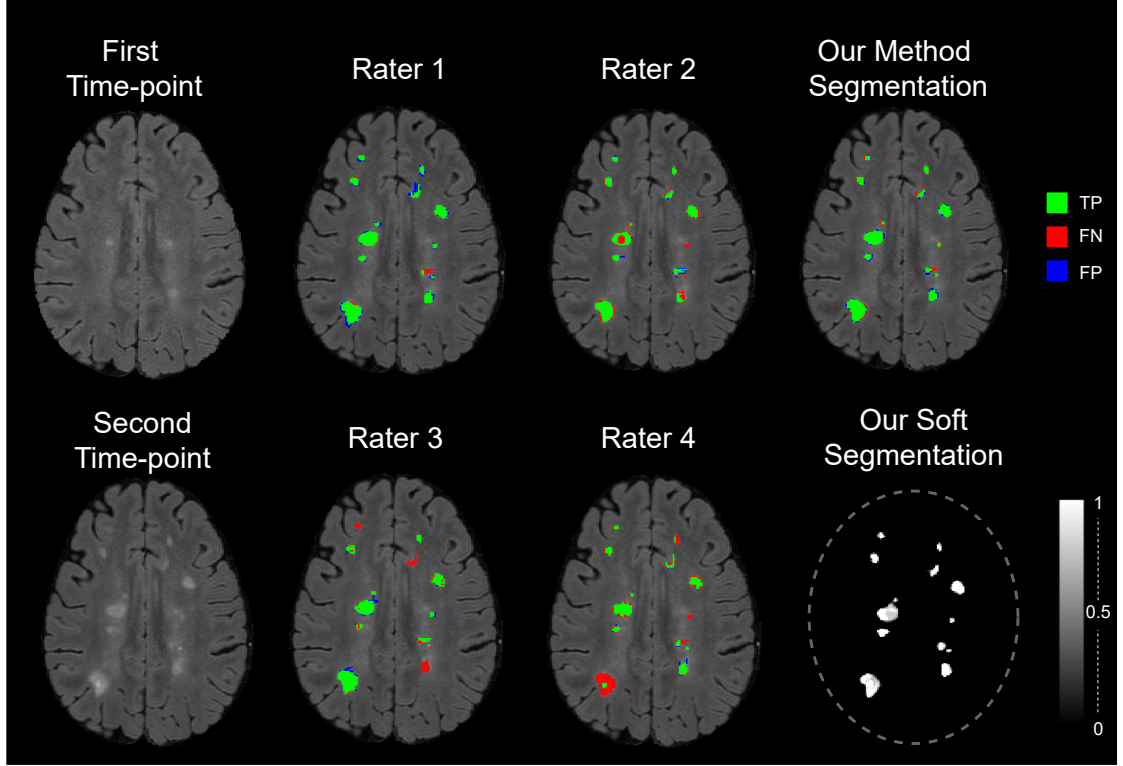


Figure 3.7: The segmentation of the proposed method and the expert rater on a sample image from MICCAI 2021 - Longitudinal Multiple Sclerosis Lesion Segmentation Testing Dataset. The segmentations are compared against the consensus of the 4 raters using the colors: green, red, and blue to symbolize TP, FN, and FP regions of new lesions.

MRI from scanners not seen during training).

The ablation study performed using the internal validation process showed that each contribution, taken separately, enhanced the segmentation accuracy. It also showed that when combining all contributions, we achieved the best results. Similarly, the challenge evaluation showed that the proposed method achieved better results than the best-performing methods of the challenge.

Our experiment in section 3.3.1.2 has shown interesting behavior of our method when trained on only 8 patients (minor performance decrease compared to using synthetic data only). The fine-tuning and optimization by selecting the best weights combination based on a very limited validation set has foreseeably led to overfitting. Thus, it is advised that the number of samples and their quality (containing enough new MS lesions) are sufficient so the fine-tuning step could enhance the performance. If the labeled dataset is not sufficient, combining both synthetic

and real data could also be explored.

Our work explored the possibility of using a similar task such as MS lesion segmentation to better train new MS lesion segmentation models. Transfer learning has led to satisfactory results. However, other methods for instance multi-task learning and consistency regularization should be explored likewise. Other of our experiments (that have not been covered in our study) investigated such strategies on both single time-point MS and new MS lesion segmentation. Unfortunately, it is difficult to deal with the different class imbalances and complexities of both tasks which makes optimizing jointly over single time-point MS and new MS lesion segmentation harder. We believe that a training-set containing both the segmentation of new lesions and the segmentation of other lesions contained in both time-points could lead the community to propose better segmentation/detection models.

Although it is sometimes difficult for experts to agree upon whether a lesion is new or not, their consistency in the segmentation of new lesions is even more difficult. This inconsistency, despite being mitigated by the consensus of several experts, will have repercussions on the quality of the segmentation accuracy. Thus we believe that if there is interest in the quantification of new lesions volume, the output of models trained only on one modality (FLAIR) and for the task of new lesion segmentation should be taken with precaution. Combining the outputs of this model with another one trained on a single time-point with several modalities (T1w and FLAIR) could lead to better and more accurate segmentation.

Besides the detection of new lesions, another interesting biomarker for MS clinicians is the measurement of disappearing lesions. Our proposed method could potentially be used for this task by inverting the time-point order. However, it has not been validated in our work and requires the appropriate expert annotations.

3.5 Conclusion

In this chapter, we propose a training pipeline to deal with the lack of data for new MS lesion segmentation from two time-points. The pipeline encompasses transfer learning from single time-point MS lesion segmentation, pretraining with time-point synthesis, and data-augmentation adapted for MR images. Our ablation study showed that each of our contributions enhances the accuracy of the segmentation. Overall, our pipeline was very effective for new MS lesions segmentation (Best score in MSSEG2-challenge [MICCAI, 2021]) and can be extended to other tasks that suffer from longitudinal data scarcity.

Chapter 4

Automatic EDSS Estimation based on MRI and Clinico-demographic Data using Deep Learning

This chapter corresponds to a journal paper submitted to Artificial Intelligence in Medicine by the following authors:

Reda Abdellah Kamraoui, Boris Mansencal, Ismail Koubiyr, Bruno Brochet, Aurélie Ruet, Thomas Tourdias, Pierrick Coupé, and on behalf of OFSEP investigators [4].

MS is an inflammatory and degenerative disorder of the central nervous system that causes functional impairments and disabilities. Measurement of dysfunctional severity helps to monitor and choose therapeutic interventions and to measure the progression of disability of MS patients. The Expanded Disability Status Scale (EDSS) is a widely used clinician-administered assessment scale evaluating the patient's disability state. However, EDSS suffers from inter-rater and intra-rater variability, the processing of large cohorts is time-consuming, and EDSS requires the presence of the patient and thus can not be performed retrospectively to assess previous states of the subject. We propose a novel method for the automatic estimation of the EDSS based on MRI and clinico-demographic information using Deep Learning. To achieve this, we first extract relevant neurodegenerative and neuroinflammatory biomarkers. Second, we propose a CNN architecture that effectively combines image-based and tabular information. Moreover, we present novel multi-phase learning and data augmentation to mitigate the data imbalance effect that makes learning from real-world distributions difficult. Overall, the results of the experiments suggest that our Deep Learning method obtains competitive performance compared to state-of-the-art methods. Finally, model interpretability shows consistency with works associating MS functional impairments with eloquent brain structures.

4.1 Introduction

MS is an autoimmune condition in which the myelin sheath, an insulating layer that covers axons and helps the propagation of action potentials, is damaged in the brain and spinal cord. Demyelination causes a disruption in the ability of the nerves to conduct electrical impulses to and from the brain thus leading to motor, sensorial, and cognitive impairments depending on the location of the inflammatory attacks. These inflammatory attacks, forming focal lesions, can impact focally the brain but can also disconnect distant areas of the brain through the interruption of long-range WM tracts.

MS is an inflammatory-demyelinating disorder of the Central Nervous System (CNS) that also has a strong neurodegenerative component. For instance, [Bermel and Bakshi, 2006] showed that brain atrophy is a clinically relevant component of disease progression in MS. Indeed, studies investigating the association of brain atrophy with clinical deficits suggested that permanent tissue damage is a more significant hallmark of disease progression than what can be explained by standard assessments of inflammatory lesions. Furthermore, the work of [Jacobsen et al., 2014] showed that patients with disability progression (over 5 years) exhibited a significant loss of the whole brain, cortical, and putamen volumes compared to patients without disability progression.

It has been heavily disputed, but it is still unclear whether neurodegeneration in MS is independent or related to neuroinflammation [Louapre and Lubetzki, 2015], [Hutchinson, 2015], [Koudriavtseva and Mainero, 2016]. Most clinical findings support the existence of a strong relationship between inflammation and neurodegeneration in MS [Fisniku et al., 2008], [Confavreux et al., 2003], [Leray et al., 2010], [Scalfari et al., 2010]. On the other hand, the presence of early cortical damage independent of inflammatory WM lesions has been observed using advanced neuroimaging techniques [Louapre and Lubetzki, 2015]. Moreover, neurodegeneration is not only associated with late-stage MS but also early stages when affecting all gray matter compartments including the cortex and the thalamus [Lucchinetti et al., 2011, Wylezinska et al., 2003]. Inflammatory and neurodegenerative components might therefore both contribute more or less independently to the disease severity.

Several measurements have been proposed to scale MS impairment, evaluate the effectiveness of therapeutic interventions, and measure the progression of disability in MS patients. The Expanded Disability Status Scale (EDSS) is arguably the most popular and used in MS. EDSS was developed by [Kurtzke, 1983] as an improvement of the Disability Status Scale. EDSS is a clinician-administered assessment scale evaluating the functional systems of the CNS. EDSS is an ordinal rating system ranging from 0 (normal neurological status) to 10 (death due to MS) in 0.5 increments interval (when reaching EDSS 1). The EDSS score measures

4. Automatic EDSS Estimation based on MRI and Clinico-demographic Data using Deep Learning

several functions: visual, brainstem, pyramidal, cerebellar, sensory, bowel/bladder, cerebral and ambulation (*i.e.*, walking).

Although EDSS is widely used, multiple studies raised concerns about its use and clinical relevance. First, the EDSS score is prone to inter-rater variability [Noseworthy et al., 1990]. Second, the same interval difference in EDSS from one value to another has different meanings depending on the initial value [Barker-Collo, 2006]. An EDSS change from 0 to 1 is not equivalent in terms of severity compared to a change from 5 to 6. Third, the EDSS has different rates of change according to the stage of the pathology [Ravnborg et al., 2005]. The worsening occurs at higher rates for patients with low EDSS baseline.

Similarly to other automatic tasks related to MS, such as MS lesion segmentation [Zeng et al., 2020, Shoeibi et al., 2021, 2], detection of new lesions [Comowick et al., 2021, 1], or segmentation of brain structures [González-Villà et al., 2016, Coupé et al., 2020], the use of Machine Learning (ML) and Deep Learning (DL) can be explored for the estimation/prediction of EDSS. Indeed, the automated estimation of EDSS offers several advantages including getting rid of inter-rater variability, processing large cohorts, obtaining an estimation for the previous states of the patient, or obtaining an estimation when the patient is not available. Moreover, achieving EDSS estimation with an automatic method opens door to anticipating (predicting) EDSS at later time points. This can be even more relevant for identifying patients at risk and choosing the best medication in terms of the balance of benefits/side-effects.

The EDSS could be estimated from different types of data that are correlated with the patient’s impairment level. For instance, some works have investigated the use of medical reports in a textual form to infer the EDSS score. [Alves et al., 2022] validated a machine learning model to estimate EDSS scores at specific time points for MS patients using available text-based clinical notes from a real-world data source. The EDSS estimation model uses XGBoost gradient-boosting regression models to perform this task. The authors have shown the capacity of a learning-based model to fit a clinical score such as the EDSS with high accuracy. Moreover, the paper showcases the need for an automatic and objective solution when processing a large cohort of data. However, this method requires clinical notes that already contain keywords such as "wheelchair", "pain", "spasm", and other textual indications of the patient’s state thus limiting the method use cases.

Other works have explored the use MRI based information to estimate the EDSS. As highlighted earlier, functional impairments and disabilities related to MS are linked to neurodegeneration (neuronal damage, brain atrophy) and neuroinflammation (focal MS demyelinating lesions). For instance, [Pontillo et al., 2021] proposed an ML method using clinico-demographic and MR imaging-derived variables to estimate EDSS score. MRI-based information is composed of a set of

features derived from volumetric, connectivity, and texture analyses. A subset of features was first selected using LASSO, then a regression model was trained with these features. According to their study, the most informative variables were: age, secondary-progressive course, a subset of texture features extracted from the prefrontal cortex, subcortical GM, and cerebellum. In their experiment, the authors have shown the feasibility of using regression models to estimate EDSS from MRI-derived information. However, although the EDSS values of the studied population were not evenly distributed, the authors did not analyze their results based on different EDSS ranges nor proposed measures to deal with this phenomenon well known as data imbalance. Besides, the texture-based features represent generic hand-crafted features that may not be suited to EDSS estimation, unlike DL method which guides the extraction of features based on the target task. Moreover, the volumetric-based features were not normalized using the normative lifespan trajectory to model the deviation of volumes and measure the atrophy.

In another attempt to model EDSS from MRI-based information, [Roca et al., 2020] proposed an algorithm that combines multiple ML techniques, including DL, to predict the EDSS score of MS patients at two years using age, sex, and FLAIR MRI data. The authors reported an important characteristic of the used dataset [Vukusic et al., 2020] which is the high imbalance of EDSS scores at two years (lower than high EDSS scores) and addressed it by using a metric that averages the score over each EDSS group for the analysis of the results. Although this study is one of the first trying to predict future EDSS scores from baseline information, it suffers from limitations which for the most are related to the limited access to data on the study population. The major drawback is that most of the MS patients in this study did not have any evolution between the baseline and 2 years follow-up since a large part of them were having a Disease Modifying Treatment (DMT) and the period was relatively short to observe a significant dynamic in EDSS. Thus, it is difficult to conclude if [Roca et al., 2020] results indicate that their strategy was able to predict future EDSS or if it was predicting EDSS at a baseline that was close or even identical to the follow-up one. Moreover, the authors did not use T1 sequences that better highlight the atrophied brain structures and the neurodegenerative impact of MS.

In this chapter, we address the limitations of the previous automatic methods by proposing a novel Deep Learning model for EDSS estimation using MRI and clinico-demographic data. First, we used T1 and FLAIR to extract both neuroinflammation-based and neurodegeneration-based features preserving spatial information. Specifically, the extracted features include the brain segmentation of 133 GM structures and WM regions, spatial representations for volumetric deviations of the brain structures, MS lesion map, and the disconnectome of 64 WM tracts. Moreover, the proposed DL method integrates clinico-demographic in-

formation (tabular data) to image-based information using a modulation strategy that enables to combine efficiently both information types and better guide feature extraction. Finally, we proposed a methodological solution to deal with data imbalance during model training using targeted data augmentation and multi-phase learning.

4.2 Method and Material

4.2.1 Dataset

4.2.1.1 Study Population

The dataset used in this study includes 2225 subjects provided by the "Observatoire français de la sclérose en Plaques" (OFSEP), a nation-wise prospective cohort, coming from 66 different sites. Each subject has between 1 and 11 (with an average of 1.93 and a median of 1) visits recorded in this dataset. This provides a total of 3951 medical records (which will be used independently in this study) composed of T1 and FLAIR sequences, age, sex, the type of DMT if any, and the EDSS score. The dataset was divided into training and test sets with a ratio of 80% and 20% respectively. The split is performed in a stratified way to keep both demographic/clinical data (see Table 4.1) and EDSS distribution (see Figure 4.2), as close as possible to the distribution of the complete dataset. During the split, we ensured that all records of the same patient are only in one set to avoid data leakage.

4.2.1.2 Preprocessing

T1 and FLAIR sequences are preprocessed with [2] pipeline. First, denoising is applied to both sequences [Manjón et al., 2010]. Second, an affine registration to the MNI space is performed on the T1, then the FLAIR is registered on the transformed T1 [Avants et al., 2011]. Brain localization is performed on the T1 then the mask is applied to both modalities (skull-stripping) [Manjón et al., 2014]. Then, bias correction [Tustison et al., 2010] is applied on the T1 and FLAIR. Finally, the intensities are normalized with the kernel density estimation.

4.2.1.3 Biomarker Extraction

As mentioned above, both neurodegenerative and neuroinflammation components can contribute to clinical performance, and cause functional impairments and disabilities. The impact of these components can be observed and quantified using MRI modalities such as T1 and FLAIR. From both sequences, several Biomarker

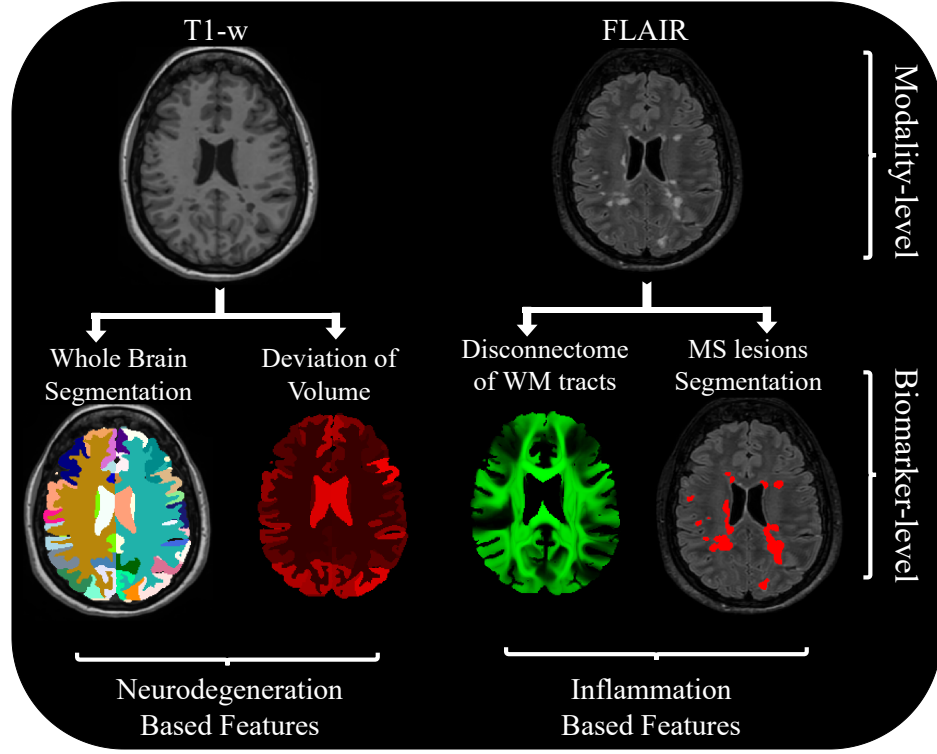


Figure 4.1: The image-based features used for the estimation of EDSS. Biomarker-level features are extracted from the MRI sequences. For the deviations of volume, higher color intensity indicates a higher deviation from the normal volume based on age and sex. For the disconnectome of WM tracts, higher color intensity indicates a higher probability that a WM tract is impacted by MS lesions.

Level Features (BLF) relevant to MS are extracted (see Figure 4.1): whole brain segmentation, deviation of brain structures volumes from normative lifespan trajectory, MS lesion segmentation, and the disconnectome of WM tracts. The first two BLF are associated with neurodegeneration while the latter two BLF are linked to neuroinflammation. These biomarkers quantify and localize brain abnormalities caused by MS that can be correlated with disease impairments. Thus, the proposed model is trained to estimate EDSS from these biomarkers and complementary clinico-demographic information.

Whole Brain Segmentation

In order to estimate structure volume, the whole brain segmentation is performed with Assemblynet [Coupé et al., 2020] pipeline. Assemblynet is an ensemble method based on a large number of CNNs processing different overlapping brain

areas. The framework is made of two assemblies of U-Nets where the second assembly, at a higher resolution, refines the decision taken by the first one. Assemblynet includes a sharing of knowledge among neighboring U-Nets and a final decision obtained by majority voting. The framework reliability was demonstrated using a comparison with state-of-the-art methods, a scan-rescan consistency analysis, and an analysis of robustness against disease presence. To ensure the accurate segmentation of structures, a lesion inpainting [Manjón et al., 2020] step was performed on T1 using lesion masks (see section 4.2.1.3) before Assemblynet segmentation.

Deviation of Volume

In order to quantify volume anomalies, the deviation from normative value is performed with lifespan modeling as proposed by [Coupé et al., 2017]. Here, a 3D map is computed to indicate the deviation of the volume of brain structures from normal development and aging of a typical healthy brain. The deviation from this normal distribution may be caused by neurodegeneration. As discussed earlier, neurodegeneration is a strong component of MS. The atrophy of several brain regions, such as the cortex, putamen, and thalamus, is associated with MS.

To compute the deviation of the volume map, we use Assemblynet whole brain segmentation and lifespan models of each structure (see [Coupé et al., 2017, Planche et al., 2022]). These normality bounds were automatically estimated from the dataset presented by [Coupé et al., 2017] composed of 2944 healthy subjects. First, the volume of each brain structure is extracted from Assemblynet segmentation and normalized compared to the total Intra-Cranial Volume (ICV).

Second, depending on the age and sex of the patient, we compute the deviation of each structure from the normative lifespan model using the following formula:

$$Abnormality(vol) = Sigmoid(\alpha * Deviation(vol) - \beta), \quad (4.1)$$

$$Deviation(vol) = \frac{(vol - mean_dist)^2}{std_dist^2}, \quad (4.2)$$

$$Sigmoid(x) = \frac{1}{1 + \exp(-x)}, \quad (4.3)$$

where vol is the relative volume of the structure, $mean_dist$ and std_dist correspond to the mean and standard deviation of the normal volume distribution. α and β are coefficients to calibrate the impact of the deviation. In this work, we used $\alpha = 2$ and $\beta = 5$.

Third, the volume abnormality value of each structure is assigned to its corresponding anatomical region using the Assemblynet segmentation to construct a 3D map that shows brain structures that deviate from normal volumes. Using such a

Table 4.1: Dataset metadata (statistics) for the entire dataset, training and testing dataset

Data	Patients	Samples	Male (%)	On DMTs (%)	Age mean (std)	EDSS mean (std)
All Data	2225	3951	1109 (28.1%)	2829 (71.6%)	41.55 (11.6)	2.37 (2.1)
Training Set	1780	3163	884 (27.9%)	2282 (72.1%)	41.35 (11.4)	2.37 (2.0)
Testing Set	445	788	225 (28.5%)	547 (69.4%)	42.37 (12.1)	2.35 (2.1)

strategy also allows us to combine age and sex information spatially image-based information.

MS Lesion Segmentation

The MS lesion mask is computed with DeepLesionBrain [2] pipeline. This pipeline is specifically trained for generalization and to perform well on unseen datasets. The MS lesion segmentation is based on the consensus of a large group of compact 3D CNNs predictions to produce a robust prediction. Moreover, the training of this model included a strong data augmentation that simulated real-world diversity and MRI artifacts to reduce dependency on training data. DeepLesionBrain generalization was validated in multiple cross-dataset experiments with state-of-the-art protocols.

Disconnectome of the WM tracts

Recent studies indicate that WM disconnection might be a better predictor of brain dysfunction and recovery compared to lesion location [Thiebaut de Schotten et al., 2014, Herbet et al., 2016]. Inspired by the work of [Thiebaut de Schotten et al., 2020], the disconnectome of WM tracts is used to indicate the potential disconnections of the WM tracts resulting from MS lesions. The disconnectome is computed using DeepLesionBrain MS lesion mask and the HCP1065 Population-Averaged Tractography Atlas [Yeh, 2022]. First, T1 is used for the non-linear registration of the MS lesion mask in the WM tract atlas space. Then, by overlaying the registered MS lesion mask on the WM tract probabilistic atlas, it is possible to compute a 3D map indicating the probability of disconnection for each WM tract. Finally, the probability map of WM tract disconnections is put back on the original space with the inverse transform.

4.2.2 Multi-phase Training and targeted augmentations

Data imbalance occurs in a dataset when the target value has an uneven distribution of observations. Depending if the target is categorical or continuous, the target value refers either to classes or values ranges. As shown in Figure 4.2a,

4. Automatic EDSS Estimation based on MRI and Clinico-demographic Data using Deep Learning

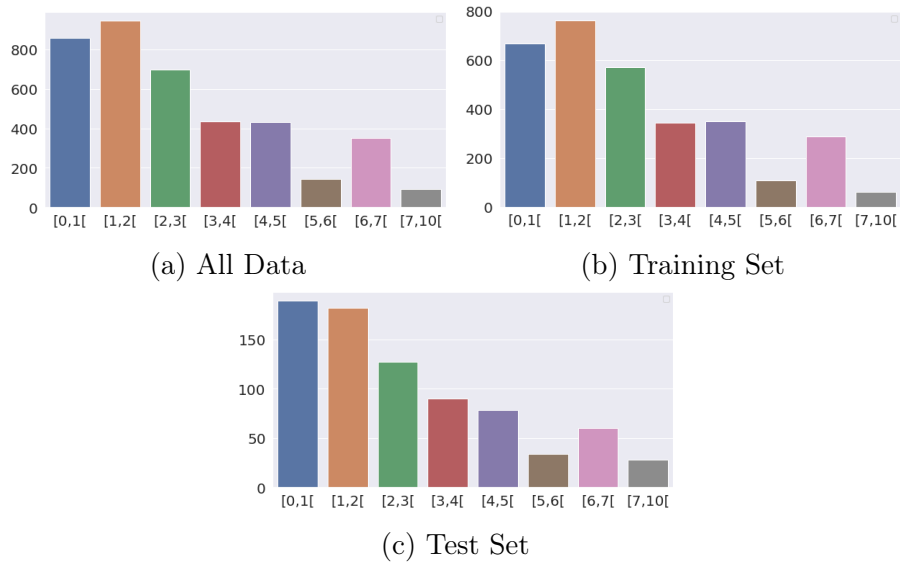


Figure 4.2: Histograms by EDSS ranges of the entire dataset, the training set, and the testing set.

the distribution of EDSS is unbalanced between the different EDSS ranges. For example, the number of samples having an EDSS in the ranges $[0, 1[$ or $[1, 2[$ is significantly higher than samples in the ranges $[5, 6[$ or $[7, 10[$. Although this EDSS distribution may accurately reflect the MS population statistics, it is difficult to train ML and DL models with unbalanced datasets. To mitigate this problem, several strategies have been studied (see [Johnson and Khoshgoftaar, 2019] for details).

First, weighting-based techniques aim to balance the impact of overrepresented and underrepresented data on the optimization process. This is usually performed by using a weighting function that gives more importance to the minority values and gives less importance to the majority values. The intuitive approach would be to define the weighting function based on the histogram of observations. However, weighting-based techniques were initially proposed for ML and may not be beneficial to DL depending on an optimization algorithm, network architecture, and loss function [Byrd and Lipton, 2019].

Second, data-centric techniques change the distribution of the training data set to decrease the problem of imbalance. Random under-sampling and random over-sampling are well-known methods to deal with data imbalance. Under-sampling discards data from the majority target values, reducing the total amount of information for training the model. Over-sampling will cause an increased training time due to the duplication of the minority values and can cause over-fitting [Johnson and Khoshgoftaar, 2019]. For classification, [Chawla et al., 2002] introduced

the Synthetic Minority Oversampling TEchnique (SMOTE), a method that uses interpolation between existing minority samples and their same-class neighbors to produce artificial minority samples. SMOTER is an extension of SMOTE for regression proposed by [Torgo et al., 2013] where the new target is a weighted average of the target values used during interpolation. Similarly, SMOGN [Branco et al., 2017] combines SMOTER with the introduction of gaussian noise.

Third, Multi-phase learning methods limit data-imbalance problems by training successively the model using different data distributions. Indeed, ML/DL are based on the assumption that test distribution comes from the same distribution as the training dataset. Thus, if the model is trained on a balanced distribution that does not reflect the real-world target distribution, this assumption will not be respected. Both [Lee et al., 2016] and [Havaei et al., 2017] proposed a similar two-phase learning procedure. The DL model is first pretrained with a balanced subset using random under-sampling and then fine-tuned using a more representative distribution of the data. [Pouyanfar et al., 2018] used a dynamic sampling technique that over-samples the low-performing classes and under-samples the high-performing classes.

In this chapter, we propose to combine both data-centric and multi-phase learning to deal with data imbalance for the training of our EDSS estimation DL model. Inspired by SMOTE and SMOTER, we propose to use MixUp augmentation [Zhang et al., 2020a] to combine image-based biomarkers samples and clinico-demographic vectors of candidates that have similar EDSS. This strategy allows us to target underrepresented EDSS ranges and augment them significantly. Additionally, our DL model was trained with the dynamic sampling strategy adapted to our task. Training batches are composed of observations sampled from the training set using a sampling function. The sampling function defines the probability of sampling from each EDSS range and is initialized first with uniform probability. Then after each epoch, the sampling function is adapted to have a higher sampling probability for the ranges that perform worse on the validation set. Finally, once the model learned to extract relevant features for each EDSS range, we freeze the encoder convolution layers and fine-tune the rest of the model on the original training set distribution.

4.2.3 Dual path DL model

To estimate the EDSS from image-based and clinico-demographic data we use the following architecture (See Figure 4.3). The choice of the dual-branch encoder was motivated by two reasons. First, several studies in MS showed that neurodegeneration is often related to neuroinflammation, but in some cases the two phenomena are independent [Hutchinson, 2015, Louapre and Lubetzki, 2015, Koudriavtseva and Mainero, 2016]. Thus, the dual-branch encoder enables learning separately

4. Automatic EDSS Estimation based on MRI and Clinico-demographic Data using Deep Learning

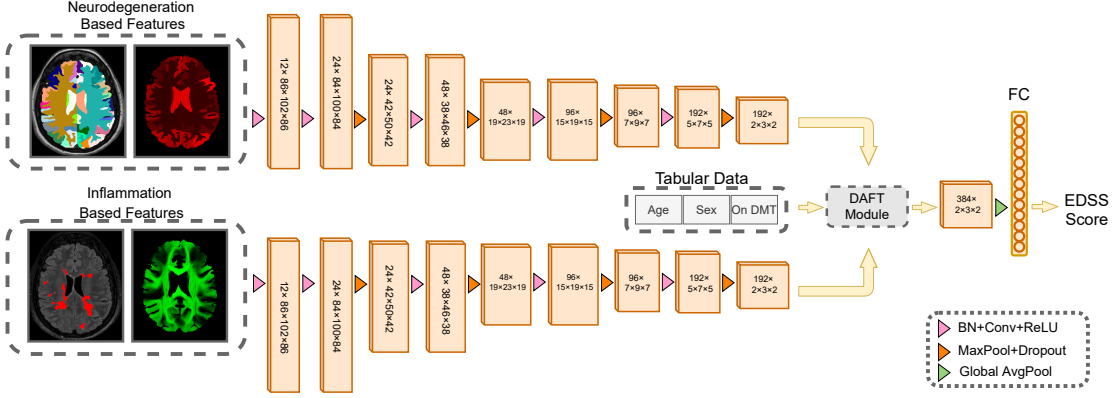


Figure 4.3: DL Architecture used for EDSS estimation

from each type of feature, and later, the combination module allows learning by exploiting the dependencies. Second, the concatenation of all input maps may not be the best strategy as pointed out in [Aslani et al., 2019]. Alternatively, the dual downsampling branches enable the model to encode information efficiently from each branch input.

Therefore, in this study, we propose a dual-branch encoder to separate feature extraction from neurodegeneration-based and neuroinflammation-based input maps. Then, an aggregation module is used to combine the dual-branch outputs (image-based information) and clinico-demographic data (tabular data). This will be discussed in-depth in Section 4.2.3.1. Afterward, the resulting feature maps are flattened and passed through fully connected layers, which in the end estimate the EDSS score.

4.2.3.1 Combining image and tabular data for DL

In addition to the image-based features, other information related to the patient and relevant to MS, such as demographic data or treatment, can be used to better estimate EDSS. These pieces of information are usually represented in tabular form (vector), and thus raise the question of how they can be combined with image data and jointly used in CNN. Most commonly, DL approaches integrate tabular data with images by fusing latent image representation with tabular information before the last fully connected layer (see [Hao et al., 2019], or [Roca et al., 2020]). However, this combination strategy limits the interaction of image spatial information with tabular data.

[Wolf et al., 2022] proposed the Dynamic Affine Feature Map Transform (DAFT) module that can be incorporated into any type of CNN to combine feature maps with tabular information. DAFT module learns scaling factors and offsets to dy-

namically apply an affine transformation to the feature maps. The scaling and offset parameters result from applying a "Squeeze-and-Excitation" operation (see [Hu et al., 2018]) on the concatenation of global-averaged-pooled feature maps and tabular data.

The DAFT module is used in our architecture to learn the best way to combine image-based features and clinico-demographic data while taking into account the spatial nature of image-based features.

4.2.4 Implementation Details

The model estimating EDSS is an ensemble of 5 networks trained separately with different training/validation data-split. During the test, the model estimation is the average of all networks predictions.

Next, each network is trained on 3D image-based inputs of size $[88 \times 104 \times 88]$ and tabular-based inputs of 3 elements. Adam [Kingma and Ba, 2014] was used for optimization with a learning rate of 0.0001 and a momentum of 0.9. Mean Squared Error (MSE) is used as a loss function for the training.

Moreover, each phase of the proposed multi-phase learning, including a sampling distribution based on training performance on each EDSS range followed by training on the real data distribution, is trained until convergence. Specifically, we used an early stopping criterion of 50 epochs (*i.e.*, the training stops if the loss function does not improve on the validation set during 50 epochs) with a maximum number of 500 epochs.

Furthermore, the experiments have been performed using PyTorch framework version 1.10.0 on Python version 3.7 of Linux environment with NVIDIA Titan Xp GPU 12 GB RAM.

4.2.5 Evaluation Metrics

Due to the data imbalance aspect of our dataset, we use different metrics to assess the EDSS estimation models. First, Mean Squared Error (MSE) is used to measure the model performance for the overall distribution. Second, we define Average Range-based Error (ARE) as the average of mean squared errors by EDSS-range:

$$ARE = \frac{1}{R} \sum_{r \in Ranges} MSE(y_true_r, y_pred_r), \quad (4.4)$$

where *Ranges* are the following EDSS ranges: $\{[0, 1[; [1, 2[; [2, 3[; [3, 4[; [4, 5[; [5, 6[; [6, 7[; [7, 10[\}$, y_true_r are the ground truth EDSS observations that belong to the EDSS range r , and y_pred_r are the model predictions respective to y_true_r . This metric ensures that equivalent importance is given to all EDSS ranges when

estimating the error. Third, to measure the correlation between the true EDSS values and the estimated EDSS, we use Pearson and Spearman correlations. Finally, to consider all the metrics mentioned above, we define *MixError* as follows:

$$MixError = \frac{ARE + MSE}{|Pearson| + |Spearman|}, \quad (4.5)$$

where a lower *MixError* represents a higher similarity between the estimated and ground-truth EDSS scores. In the following experiments, *MixError* will be primarily used to evaluate the performance since it better estimates the prediction error compared to each metric taken individually.

4.2.6 Reference Methods

In section 4.3.2, the proposed EDSS estimation model is compared against reference methods.

First, the proposed method is compared to the state-of-the-art ML method Random Forest Regressor (RFR) trained BLF. To address the data imbalance, we also combine SMOTER [Torgo et al., 2013] with RFR+BLF.

Second, similarly to the work of [Pontillo et al., 2021], our method is compared to ridge regressor on features selected using LASSO regression (Ridge+Feat-selec.). The selected features were composed of the normalized volumes of CSF, Inf. lateral ventricle, the lesion burden, the volume of juxtacortical lesions, and the number of cerebral lesions. Besides, the volume anomalies of several structures were selected such as the 3rd ventricle, the parietal operculum, the anterior insula, the medial orbital, and the postcentral gyrus. Additionally, multiple features representing WM tracts disconnections were selected such as the corticopontine, optic radiation, reticulospinal, uncinate, occipital, and medial longitudinal fasciculus.

Finally, to better analyze the results of the proposed method and the other state-of-the-art methods, we propose two baselines: RFR using age as the only feature and RFR on random values.

4.3 Results

4.3.1 Ablation Study

The proposed EDSS estimation method combines several image-based information extracted from MRI modalities and tabular data containing demographic and treatment information. To evaluate the importance of feature components and their relevancy for the EDSS estimation task, Table 4.2 shows the results of an ablation study. The possible input maps of our CNN model are organized into

Table 4.2: Results of our experiment on different combinations of features as input for the estimation model. Each experiment is performed using Multi-Phase learning and uses tabular data (clinico-demographic information) except for the last row. For input composed of only neurodegenerative features, only neuroinflammatory features, or one of the two modalities single branch encoder is used.

Features/ Input Maps	MSE	ARE	Pearson	Spearman	MixError
All features	3.17	4.52	0.55	0.53	7.18
Modality level features	3.15	4.59	0.55	0.54	7.15
T1	3.08	4.51	0.56	0.54	6.92
FLAIR	3.21	4.87	0.54	0.53	7.61
Biomarker level features (BLF)	2.91	3.84	0.58	0.56	5.95
Neurodegeneration	3.10	3.95	0.57	0.56	6.35
Structure Volumes	3.32	3.94	0.57	0.55	6.55
Normative Modeling	3.17	4.67	0.54	0.52	7.47
Neuroinflammation	3.45	4.77	0.52	0.51	7.93
MS lesions	3.35	4.53	0.51	0.50	7.78
Disconnectome	3.68	4.61	0.51	0.50	8.17
BLF without tabular data	3.16	4.34	0.54	0.53	7.01

modality level features which regroup T1 and FLAIR sequences and BLF which are extracted using these MRI sequences. BLF are also categorized into neurodegenerative (*i.e.*, whole brain segmentation and volume deviation) and neuroinflammatory (*i.e.*, MS lesions and WM tracts disconnectome) features.

First, the method that uses all possible input maps obtained worse performance than that that uses only BLF. The use of MRI modalities as input for our model obtained a decent performance when considering that no additional expert knowledge or external tools have been utilized. However, BLF showed the best performance compared to all the other combinations.

Second, using neurodegeneration-based maps obtained better performance than using neuroinflammation maps as input for our model. Our experiment suggests that neurodegeneration-based maps are more important for EDSS estimation. The inflammation-based maps showed their relevancy only when combined with neurodegeneration-based maps (MS lesion and disconnectome obtained the worst performance).

Finally, when comparing BLF without tabular data and BLF, using tabular data (composed of age, sex, and the use or not of DMT) had a substantial improvement on model estimation.

4. Automatic EDSS Estimation based on MRI and Clinico-demographic Data using Deep Learning

Table 4.3: Results of our comparative study of the proposed EDSS estimation model with state-of-the-art methods. Results are organized, from top to bottom, to include the proposed method and its variation without multi-phase learning, state-of-the-art methods, and the performance of baseline

Method	MSE	ARE	Pearson	Spearman	MixError
The Proposed Method	2.91	3.84	0.58	0.56	5.95
The Proposed Method without Multi-phase Training	3.14	4.60	0.55	0.54	7.10
RFR+BLF	3.22	4.85	0.53	0.52	7.69
RFR+BLF+SMOTER [Torgo et al., 2013]	3.39	4.39	0.52	0.51	7.55
Ridge+Feat-selec. [Pontillo et al., 2021]	3.31	4.85	0.51	0.50	8.08
RFR+Age Only	3.73	5.61	0.05	0.04	103.78
RFR+Random Only	5.94	9.03	0.01	0.01	748.50

BLF= Biomarker Level Features, RFR= Random Forest Regressor, Feat-selec= feature selection

4.3.2 Method Comparison

To evaluate the performance of our EDSS estimation model, we compare its performance with reference methods. Table 4.3 shows the results of a comparison study that includes the proposed method and variations of our method, state-of-the-art strategies, and two baselines using age only or random values only.

First, the use of multi-phase learning, the strategy we proposed to deal with data imbalance, resulted in a better overall performance (*MixError*). In addition to the predictable reduction of ARE when using multi-phase learning, the results showed an improvement in MSE, Pearson, and Spearman correlations. This suggests that training our model with multi-phase learning improved even the performance over the real dataset distribution (population statistics).

Second, when comparing our method with RFR+BLF(+SMOTER), the CNN obtained a substantially better performance for all metrics. The use of SMOTER [Torgo et al., 2013] to mitigate the data imbalance effect improved ARE but not MSE compared to the one without SMOTER, unlike the results of using multi-phase learning for CNN which improved all metrics.

Third, similarly to the work of [Pontillo et al., 2021], we compared our method to ridge regressor on features selected using LASSO regression. Ridge+Feature-selection performed worse than RFR+BLF.

Finally, when comparing all automatic methods with the baselines RFR+Age Only and RFR+Random Only, all metrics, and particularly *MixError* that aggregates them, show that the automatic methods perform better than a model relying only on age or trained to fit random values.

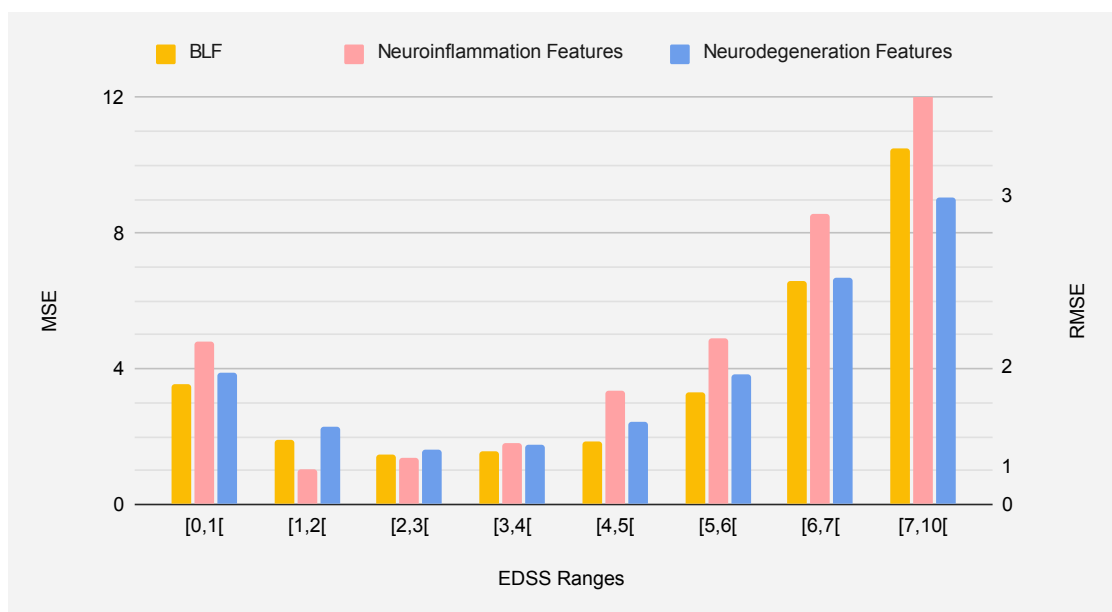


Figure 4.4: The performance of the EDSS estimation model on each EDSS range when using BLF channels (the proposed method), neuroinflammation, and neurodegeneration-based channels. The performance is measured with MSE on the left vertical axis, and Root Mean Squared Error (RMSE) on the right vertical axis

4.3.3 Model Performance based on EDSS ranges

Figure 4.4 shows the performance of the EDSS estimation model on each EDSS range, measured with MSE on the left vertical axis, and with the Root Mean Squared Error (RMSE) on the right vertical axis.

First, the figure shows a low EDSS error for the EDSS values ranging from 0 to 6 ($0 \leq \text{EDSS} < 6$). However, the estimation model seemed more prone to error for the rest of the EDSS values ($6 \leq \text{EDSS} < 10$). Indeed, although the use of multi-phase learning to mitigate data imbalance, the results suggest that more training samples are necessary, at the higher end of the EDSS spectrum, to better estimate high EDSS values.

Second, when comparing the models trained with inflammation-specific and neurodegenerative-specific channels, the results showed inflammation-specific channels to be more efficient for low EDSS ranges while neurodegenerative-specific showed better results for high EDSS ranges. This pattern is interesting since early forms of MS are characterized by an inflammatory response (lesions) and long-term MS frequently results in neurodegeneration (atrophy of brain structures).

Third, the model trained with BLF did not always obtain the best results, but it

obtained the best average performance over all EDSS ranges. This result suggests that using BLF is a trade-off between using only neurodegenerative-specific channels that perform well for high EDSS ranges and using only inflammation-specific channels which work better for low EDSS ranges.

4.3.4 Model interpretability by EDSS range

Due to their increasing complexity, DL models are often regarded as black-box tools since it is difficult to understand how the network layers operate consecutively and how the different parts of an input are relevant to the final decision. In this section, we try to better understand which brain regions contribute the most to the prediction of our EDSS estimation model using Gradient-weighted Class Activation Mapping (GradCAM) [Selvaraju et al., 2017].

For the GradCAM target layer, we have chosen the concatenation of the last convolution layers of the dual-path encoder. As discussed in section 4.2.4, the EDSS estimation model prediction is the average of 5 networks trained on different splits of the training set. Figure 4.5 represents the resulting visualization in the 3 MRI views (*i.e.*, axial, coronal, and sagittal) for each EDSS range. Each range visualization represents the average of GradCAM for all testing set patients within this specific range, averaged across the 5 networks.

We first notice the progressive shift of the importance from the forebrain-subcortical region for low EDSS values, to the parietal-occipital region for high EDSS values.

Moreover, we distinguish three phases where the importance is focused differently: $0 \leq \text{EDSS} < 3$, $3 \leq \text{EDSS} < 5$, $5 \leq \text{EDSS} < 10$. Brain structures with high intensity for each phase are summarized in Table 4.4. For $0 \leq \text{EDSS} < 3$ patients, the estimation model focused on accumbens area, basal forebrain, subcallosal area, caudate, amygdala, thalamus, hippocampus, entorhinal area, anterior insula, frontal operculum, superior parietal lobule. For $3 \leq \text{EDSS} < 5$, we see that the importance is not particularly focused on a specific region. For $5 \leq \text{EDSS} < 10$, the most significant structures are the cuneus region, supplementary motor cortex, post/pre/pre-medial central gyri, and superior occipital gyrus. the calcarine cortex, inferior and fusiform occipital gyri.

As we can see in the higher scale visualization of Fig 4.5, the most highlighted regions are the motor function, sensory, somatosensory, and visual areas. On the other hand, in lower scales, the most highlighted regions are subcortical structures which have a strong association with cognitive dysfunction [Chiaravalloti and DeLuca, 2008, Small, 2014]. As discussed in [Minagar et al., 2013, Houtchens et al., 2007], the atrophy of deep gray nuclei, such as the thalamus, can be detected in the earliest phases of MS. The involvement of these regions in MS is associated with clinical manifestations including cognitive decline, fatigue, painful

Table 4.4: Brain structures with high GradCAM density by EDSS ranges (refer to Figure 4.5).

$0 \leq \text{EDSS} < 3$	$3 \leq \text{EDSS} < 5$	$5 \leq \text{EDSS} < 10$
Accumbens Area, SCA, Basal Forebrain, Putamen, Thalamus, Hippocampus, Caudate, Amygdala, Ventral DC, AIns, Pallidum, Ent, SPL, POrG, MOrG, FO	No regions with high values. The importance is distributed.	SPL, OCP, Cun, PCu, Calc, MPoG, MPrG, SOG, AnG, PoG, MOG, IOG, OFuG.

SCA= subcallosal area, FO= frontal operculum, SPL= superior parietal lobule, Ent= entorhinal area, AIns= anterior insula, Cun= cuneus, PCu= precuneus, SMC= supplementary motor cortex, Calc= calcarine cortex.

Gyrus: MPoG= medial postcentral, MPrG= medial precentral, SOG= superior occipital, AnG= angular, PoG= postcentral, MOG= middle occipital, PrG= precentral, IOG= inferior occipital, OFuG= occipital fusiform, POrG= posterior orbital, MOrG= medial orbital

syndromes, and ocular motility disturbances.

4.4 Discussion

In this work, we showed that it is possible to estimate EDSS from MRI-based and clinico-demographic information.

In our experiments, the obtained results suggested that, overall, neurodegenerative features such as whole brain segmentation map and the deviation from the normal volume better contribute to EDSS estimation compared to neuroinflammation features that include MS lesion mask and disconnectome of WM tracts. In more details, the inflammatory features impact more the low values of EDSS while the neurodegenerative features become progressively more and more important with higher EDSS values. This result fits well with the history of the disease that is known to be more inflammatory at the early stage (with good response to immunomodulatory medications) while neurodegeneration (possibly driven by the initial phase of inflammation) will become more and more prominent. Furthermore, the brain region affected by atrophy and the intensity of neurodegeneration could explain or predict functional impairments.

In this chapter, we raised the problem of dataset imbalance and the need for multiple metrics to assess the performance of our EDSS estimation model. The proposed *MixError* which aggregates several metrics allows us to track a single value. However, the selection of metrics should be well-considered since some metrics are mathematically related to each other. This and other difficulties related to metric aggregation have been pointed out in the work of [Reinke et al., 2021].

Our results showed that the proposed multi-phase learning, including a sampling distribution based on training performance on each EDSS range followed by training on the real data distribution, enhanced all metrics. This result suggests that training the model directly on an unbalanced dataset leads to a sub-optimal solution. Thus, we consider that the proposed multi-phase learning is a way of finding a trade-off between efficient learning on the balanced dataset and learning

from the distribution that better describes real-world data.

DL interpretability results need to be taken with care since Explainable/Interpretable AI has some known limitations [Rudin, 2019]. Contrarily to other works that use an interpretability strategy on a single image and a single network to understand its prediction, each visualization presented in section 4.3.4 was constructed using 5 different models that have been initialized differently and trained on different data. Besides, each visualization results from the average prediction of several dozen to a few hundred images. Thus, the visualizations are stable and our interpretation is more confident. Overall, our interpretation is consistent with the characteristics of the EDSS score and the state-of-the-art studies that associate functional impairments of MS with brain structures.

Future work should explore the possibility of predicting patient future EDSS based on MRI, clinico-demographic data, and the baseline EDSS score. Indeed, studies showed the correlation of some brain structure volume with disease progression [Jacobsen et al., 2014]. This leads us to hypothesize that EDSS prediction could be achieved similarly to our strategy. However, this is challenging since it would require a wide longitudinal dataset with patients having multi-point homogeneous data, spaced over a long period of time.

4.5 Conclusion

In this chapter, we proposed a novel method for the estimation of EDSS from MRI-based and clinico-demographic information with a Deep Learning model. We first propose to extract neurodegenerative and neuroinflammatory biomarker maps from T1 and FLAIR modalities that will serve as input to our estimation model. These Biomarker Level Features (BLF) are composed of the whole brain segmentation, the deviation from the normal volume of structures, MS lesion segmentation, and the disconnectome of WM tracts. Next, we propose a novel dual path encoder to efficiently extract neurodegenerative and neuroinflammatory based features from two separate pathways before aggregating them alongside tabular information. The latter aggregation uses a module suited for learning how to combine spatial features and tabular information. To mitigate the data imbalance effect that makes learning from real-world distributions difficult, we propose a novel multi-phase learning that balances training sampling based on the model performance and a novel data-augmentation technique that combines SMOTE and MixUP principles. During validation, our Deep Learning method surpassed conventional state-of-the-art methods. Moreover, the interpretability of our model showed consistency with works associating MS functional impairments to brain structures.

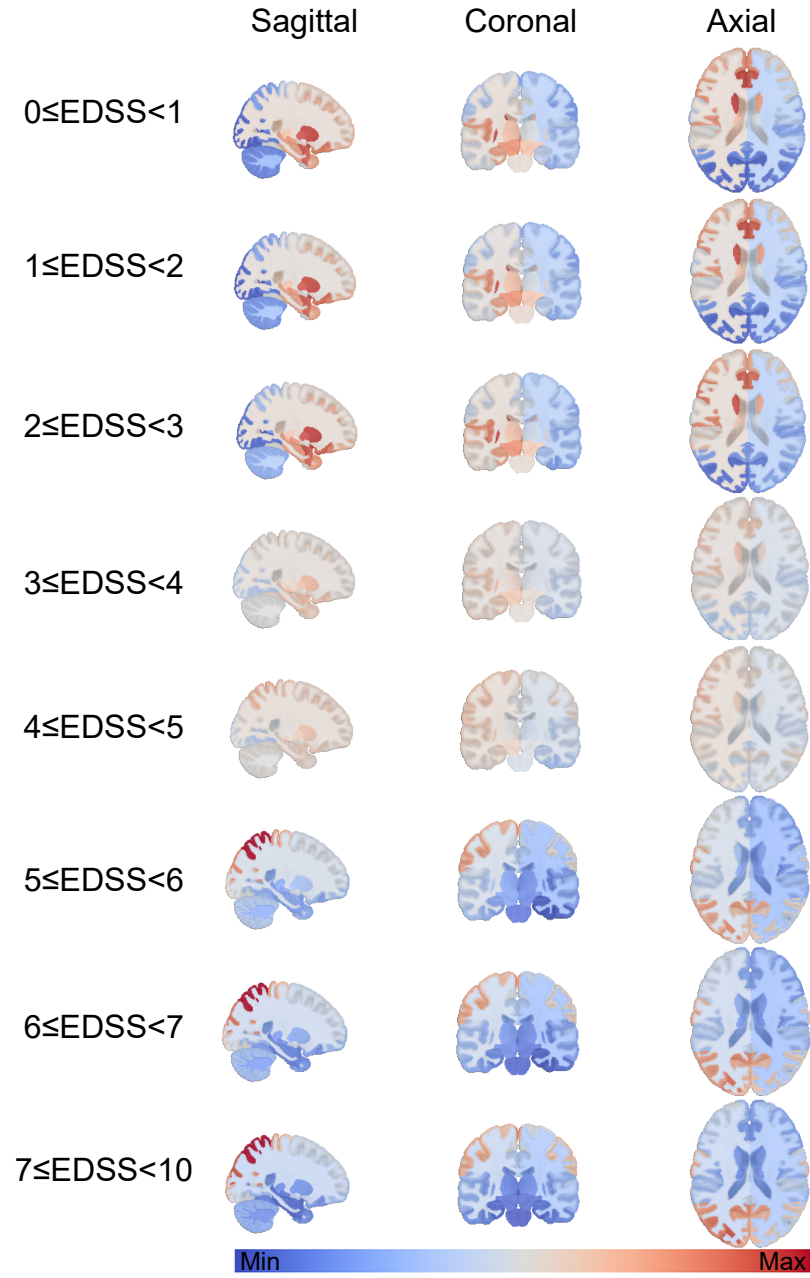


Figure 4.5: Visualizations of brain regions that contribute the most to the estimation of EDSS with the proposed model. For each EDSS range, the heat-map highlights high importance with warm colors (*i.e.*, dark-red is the maximum), and low values with cold colors (*i.e.*, dark-blue is the minimum).

Chapter 5

MRI Analysis Pipelines for MS

5.1 Introduction

The majority of MR image analysis pipelines are available as packages that must be downloaded, installed, and configured. Installation steps might be challenging, necessitating the experimented people who are not always available in research laboratories or clinical settings. Additionally, computing resources must be allocated to run the program and users must be educated to use it. The usage of these packages may be complicated by these needs, particularly the most modern and sophisticated ones as they frequently have demanding hardware requirements. Furthermore, the user community must be provided with multi-platform versions and support.

In this section, we describe how the pipelines proposed in this thesis have been made easily available and can be freely used by the MS community.

5.2 Pipeline Accessibility

Each pipeline proposed during this thesis is containerized and is either available on the volBrain platform, at the time of redaction, or will be added shortly.

5.2.1 Containerization

To deal with environment and software heterogeneity, it is possible to distribute a containerized version of the pipeline. Containers are standard software unit that packages code and all of its dependencies to speed up and reliably run applications from one computing environment to another. Containers ensure the application works uniformly despite differences between development and production environments. Although containerization solves the problem of environment compatibility

and installation, it does not solve the hardware resource problem and the problem of inexperienced users.

5.2.2 Hosting Platform: volBrain

To solve the remaining problems including the infrastructure and the expertise to install and run the pipeline, our pipelines are hosted by volBrain. volBrain is a web interface, proposed by [Manjón and Coupé, 2016], for hosting and running MR image analysis pipelines¹. volBrain does not require any installation, configuration, training, or hardware resources. Instead, the users interact through a web interface using a SaaS (Software as a Service) model to select a pipeline, upload MRI and provide demographic information, and then start execution. Once a job is submitted, the platform starts an instance (container) of the pipeline from the image (Docker) of the selected pipeline, on the combined resources of the Polytechnic University of Valencia and the University of Bordeaux. The container will receive user input, execute MR image analysis, and provide for the submitted case easy-to-read reports (*i.e.*, PDF) and additional complementary outputs (*i.e.*, zip archive). Since 2015, this platform has processed more than 400.000 MRIs for more than 7000 users around the world.

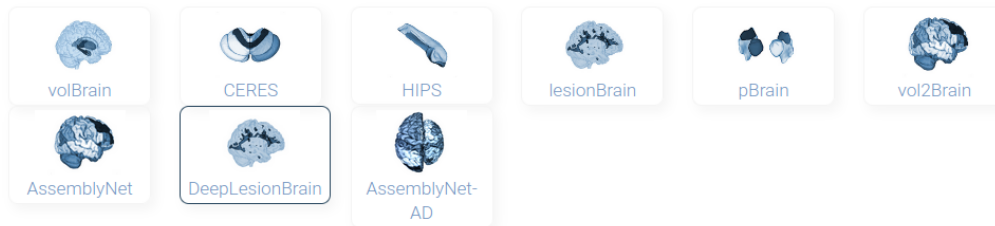
5.3 Pipeline Description

During this thesis, we developed containerized versions of our pipelines compatible with the volBrain interface. First, when the user chooses a task using the web-based interface, he/she is asked to upload the required anonymized MRI modality/modalities in compressed Nifti format and provide subject information. For instance, Figure 5.1 shows that the user has chosen DeepLesionBrain which requires the uploading of T1 and FLAIR modalities and optionally providing sex and age for further analysis. Afterward, a job is launched, and the container is running. The job takes a few minutes depending on the task (9 minutes for DeepLesionBrain) and several jobs can be processed in parallel. Finally, when the pipeline is complete, the user can easily retrieve the results. For instance, Figure 5.2 shows that the user can download an easy-to-read report in PDF format, visualize the provided segmentation using the web interface, or download an archive containing all the results.


In the following, we will detail the reports of DeepLesionBrain and Longitudinal DLB, two pipelines described in Chapter 2 and Chapter 3 of this thesis. Then, we will detail the update of DeepLesionBrain with EDSS estimation described in Chapter 4. It should be mentioned that the following reports also include a

¹<https://volbrain.net/>

Select a pipeline



DeepLesionBrain

Provides segmentation of intracranial cavity, brain tissues (GM, WM, CSF) and hyperintense white matter lesions using T1w and FLAIR MR images. 

The DeepLesionBrain pipeline requires a 3D T1w MRI and a 3D FLAIR MRI as input to process your case (see [description](#)). You have to provide **anonymized** and **compressed** Nifti (.nii.gz) files. If you supply gender and age, the expected bounds for each tissue/structure will be included in the report (see [tutorial](#)).

N.B: DeepLesionBrain has been designed to deal with standard T1w and FLAIR images without any preprocessing (e.g., skull stripping, registration, denoising...). Pipeline failures are expected for other image types (Gd-enhanced T1w, T2w, etc) or preprocessed T1w and FLAIR MRIs.

Mandatory

Optional

Sex Age

Figure 5.1: volBrain pipeline selection. The user has chosen DeepLesionBrain which requires uploading T1 and FLAIR modalities and optionally providing sex and age for further analysis

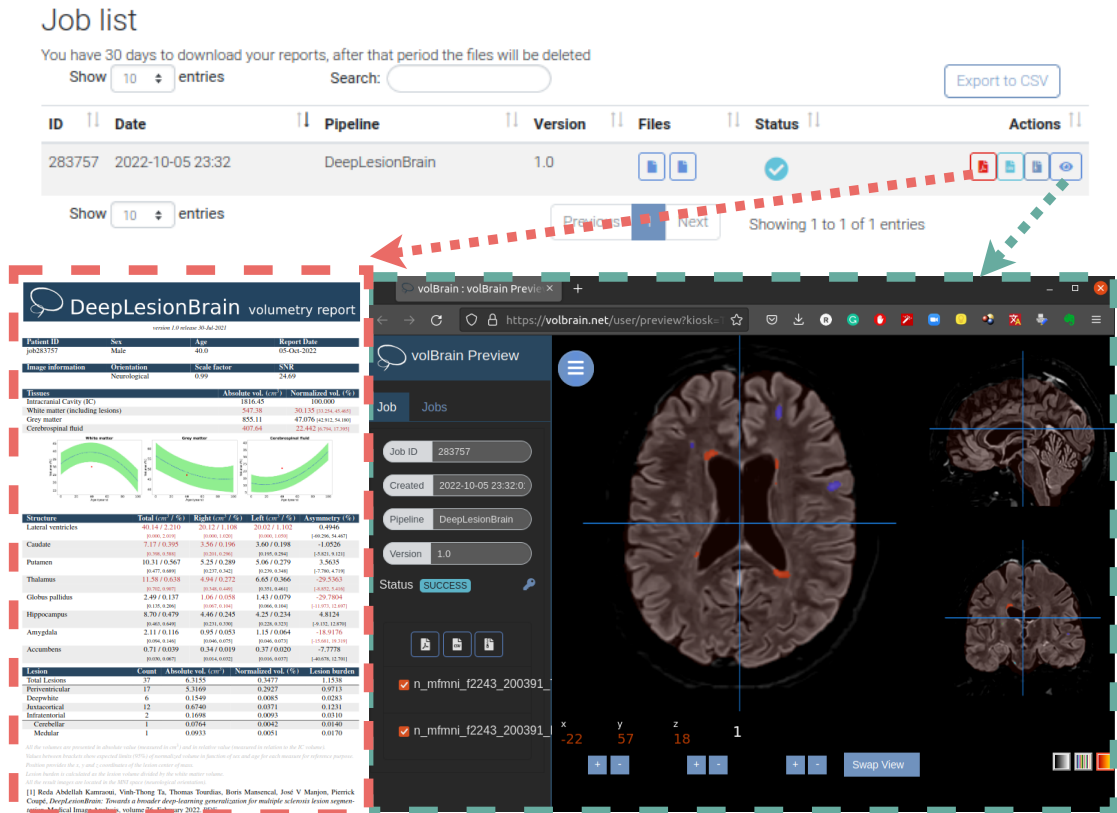


Figure 5.2: volBrain web interface to visualize results. When the task is complete, actions buttons from left to right allow the user to download the PDF report, download a CSV version of the report, download an archive (.zip) with all the results, and visualize the segmentation with the web interface

volumetric analysis of whole brain structures obtained with AssmeblyNet [Coupé et al., 2020].

5.4 Pipeline Report

In this section, we describe the report automatically generated by DeepLesionBrain and Longitudinal DLB pipelines for MS analysis. The design of the report focuses on readability and compactness, providing information and analysis relevant to MS and organizing them by priority. Since the report can be used for multiple purposes (*i.e.*, the diagnosis and the monitoring of MS patients' states, or specific clinical research related to MS), the report starts with a summary of the analysis and then details each section.

5.4.1 DeepLesionBrain for MS lesion Segmentation

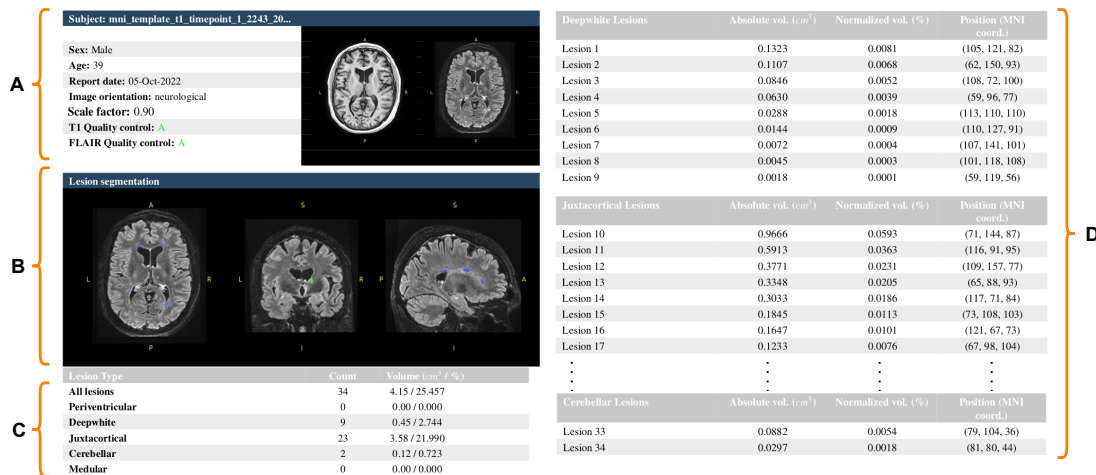


Figure 5.3: MS lesions analysis of DeepLesionBrain report. A: Patient and MRI information. B: MS lesion segmentation. C: Summary of MS lesions. D: Details about each lesion.

DeepLesionBrain report is a single PDF document aggregating several MR image biomarkers relevant to MS showing neuroinflammation and neurodegeneration activities. The information is organized in an easy-to-read and efficient format.

First, the report includes a section on the analysis of lesions (see Figure 5.3). This is helpful for dissemination in space (see Section 1.1.3.2) analysis and shows the inflammatory activity. *i)* An analysis that includes patient information such as sex and age, and MRI information such as the scale factor, image orientation,

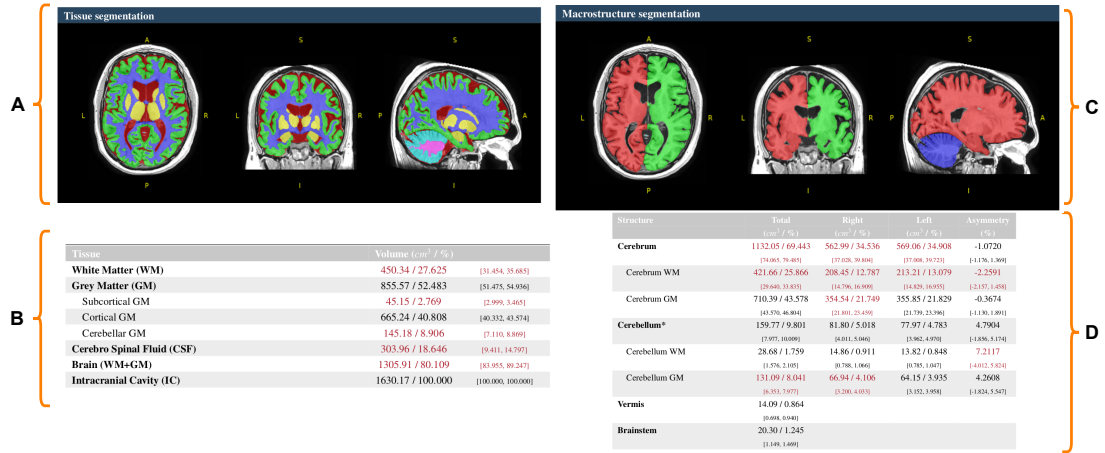


Figure 5.4: Tissues and Macro-structures of DeepLesionBrain report. A: Tissue segmentation. B: Volumetric details of tissue segmentation. C: Macro-structures segmentation. D: Volumetric details of Macro-structures segmentation.

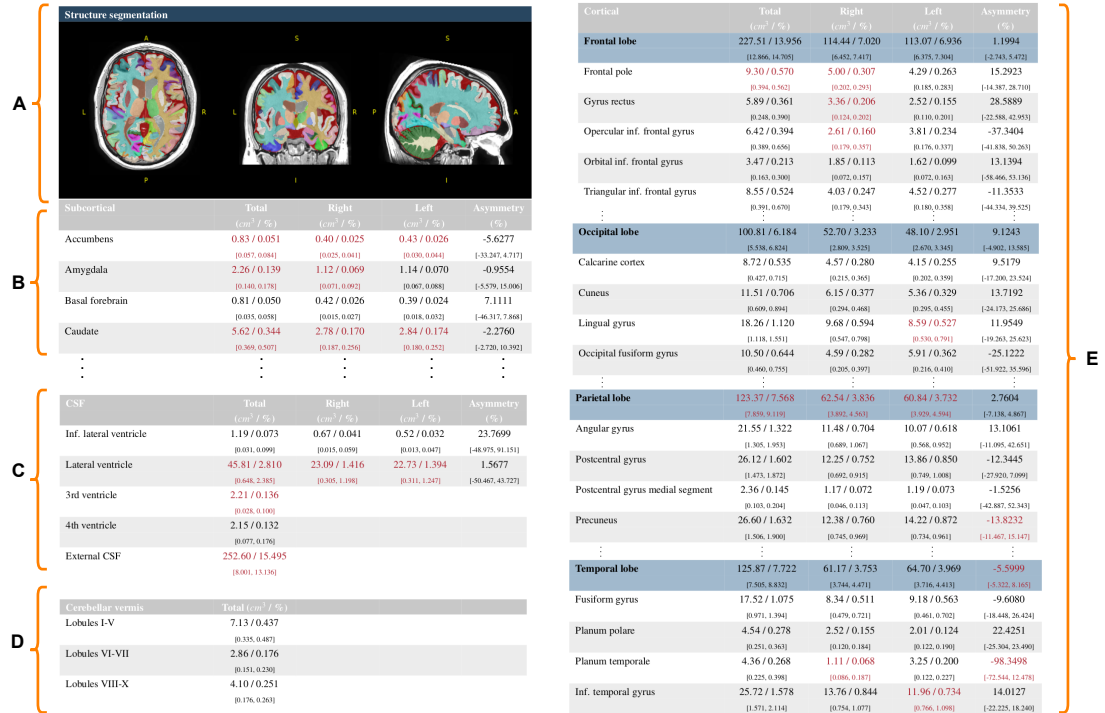


Figure 5.5: Whole brain segmentation of DeepLesionBrain report. A: Whole brain segmentation. B, C, D, and E represent respectively volumetric details of subcortical, CSF, cerebellar vermis, and cortical structures.

and automatic modality quality control is first presented (see Figure 5.3 A). *ii*) Visual representation of MS lesion segmentation is available, where different lesions according to their location are represented in different colors (see Figure 5.3 B). *iii*) A summary of MS lesions showing the lesion and the volume of each lesion type is also provided (see Figure 5.3 C). *iv*) An in-depth analysis of each lesion that encompasses absolute and normalized volumes, and the position of the lesion in the MNI space (see Figure 5.3 D).

Second, the report contains tissue and macrostructures analysis (see Figure 5.4). It helps monitor neurodegeneration and quantify WM, GM, and main brain components atrophy using lifespan model of normative volumes. *i*) The analysis includes a visual representation of tissue segmentation showing WM, GM (for cortical, subcortical, and cerebellar), and CSF (see Figure 5.4 A). *ii*) Shows volumetric details of tissue segmentation indicating absolute and normalized volumes, and normal volumes for each tissue type if age and sex are provided (red values highlight abnormal tissue volumes) -see Figure 5.4 B. *iii*) Macro-structure visualization provide the brain segmentation into the cerebrum, cerebellum, vermis, and brainstem (see Figure 5.4 C). *iv*) This part of the analysis offers volumetric details including absolute and normalized volumes of the total macro-structure, the left or right part, and the asymmetry (see Figure 5.4 D). Normal values of each macrostructure are displayed on the report if age and sex are provided (red values highlight the abnormal macrostructure volumes).

Third, the report contains the whole brain segmentation in 133 structures and regions (see Figure 5.5). This helps to get a detailed analysis highlighting the atrophy level for each brain structure. *i*) a whole brain segmentation visualization (see Figure 5.5 A). *ii*) B, C, D, and E represent respectively volumetric details of subcortical, CSF, cerebellar vermis, and cortical structures. The volumetric analysis includes absolute and normalized volumes of the total structure. If the structure is present in both hemispheres, the analysis also contains left and right part volumes, and the asymmetry. Normal values of each structure are displayed on the report if age and sex are provided (red values highlight the abnormal volumes).

5.4.2 Longitudinal DLB for the Detection of New MS lesions

Longitudinal DLB pipeline is an extension of DeepLesionBrain that highlights changes between two time-points such as new MS lesions, MS lesions that disappeared at follow-up, and longitudinal changes in tissues and structures.

Longitudinal DLB report is structured similarly to the DeepLesionBrain one and should be reviewed jointly with the DeepLesionBrain reports of both time-points for in-depth analysis. Thus, Longitudinal DLB pipeline generates auto-

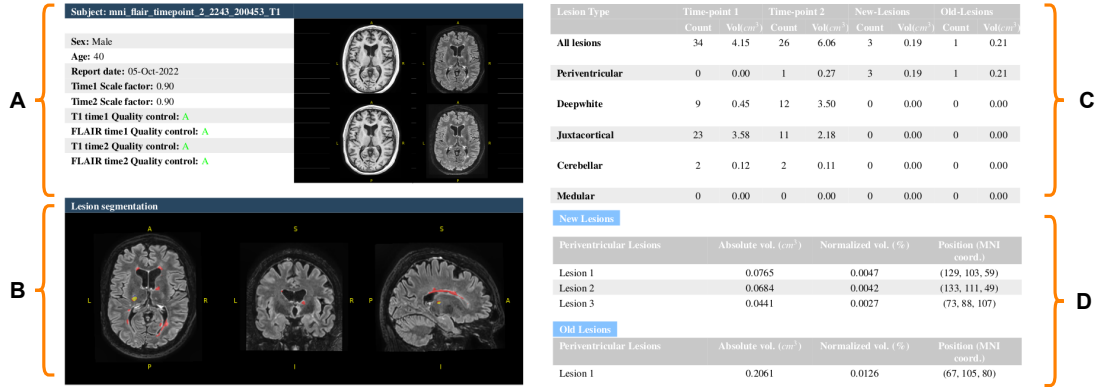


Figure 5.6: MS lesional changes analysis of Longitudinal DLB report. A: Patient and MRI time-points information. B: MS lesion segmentation, showing new, old, and persistent lesions in different colors. C: Summary of MS lesions of the two time-points and the changes. D: Details about each new lesion and old lesion.

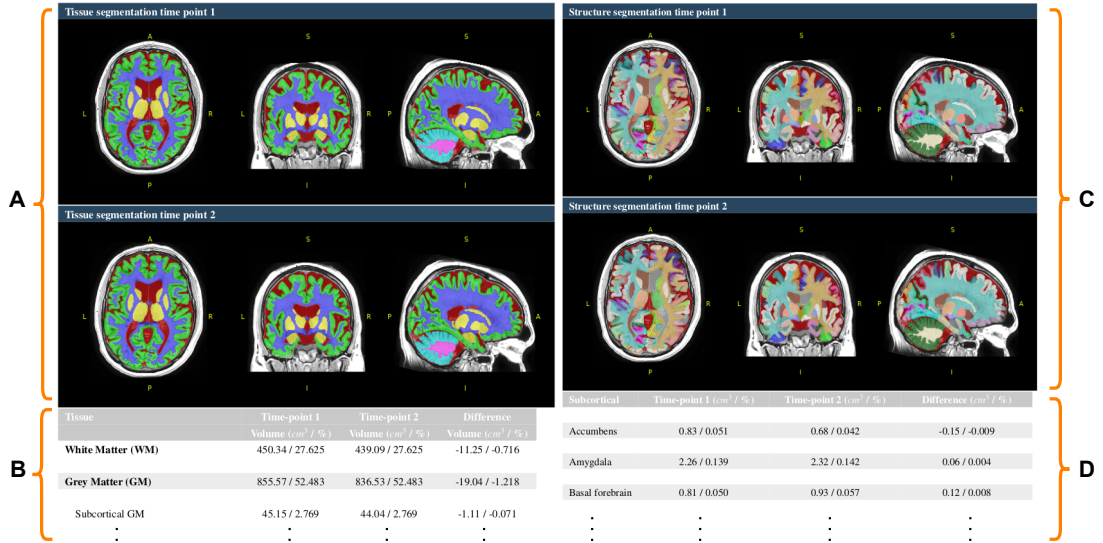


Figure 5.7: Tissues and structures changes of Longitudinal DLB report. A: Two time-points tissue segmentation. B: Volumetric details of the longitudinal changes in tissue segmentation. C: Two time-points whole brain segmentation. D: Volumetric details of the longitudinal changes in structures.

matically reports for both time-points in addition to the longitudinal reports that shows the main changes.

First, the report includes lesional changes (see Figure 5.6). This helps to monitor the inflammation activity by showing the evolution of the lesions and highlighting the dissemination in time (see Section 1.1.3.2). *i)* Patient and MRI information including scale and automatic quality control for each modality of both time-points is presented (see 5.6 A). *ii)* MS lesion segmentation, showing new, old, and persistent lesions in different colors (see 5.6 B). *iii)* Shows a summary of MS lesions detected in each time-point and the changes indicating new and disappearing lesions for each lesion type (see 5.6 C). *iv)* The report lists each new lesion and old one (the lesions that disappeared during the follow-up) and indicates their absolute and normalized volumes and their spatial position in the MNI space (see 5.6 D).

Second, the report highlights longitudinal changes in tissue (WM, GM, and CSF) and whole brain segmentation (see Figure 5.7). This is helpful in measuring subtle intra-subject volume changes that may be caused by neurodegeneration. *i)* The report shows the two time-points tissue segmentation and whole brain segmentation (see 5.7 A and C). *ii)* The report shows volumetric details of the longitudinal changes in tissues and whole brain segmentation. Absolute and normalized volumes of each time-point are mentioned besides the time-point difference to measure atrophy (see 5.7 B and D).

5.4.3 DeepLesionBrain update for EDSS estimation

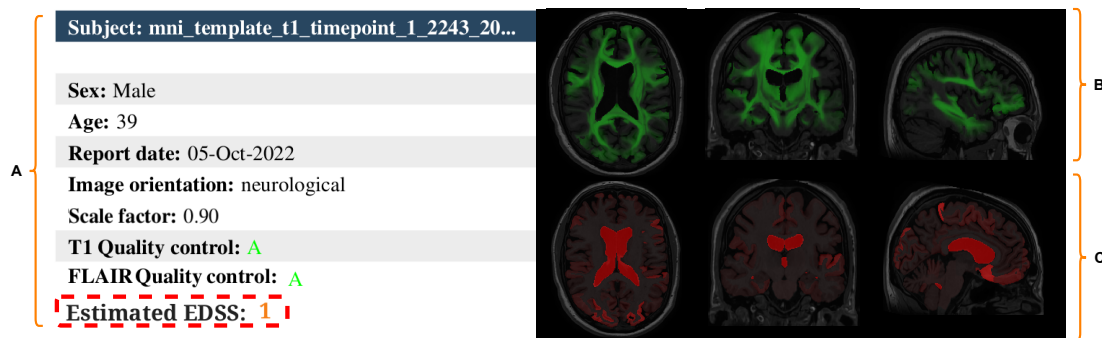


Figure 5.8: Enrichment of DeepLesionBrain by including estimated EDSS in A, disconnectome of WM tracts in B, and structure volume deviation from normal distribution in C.

Our DeepLesionBrain report will be enriched in its future releases with additional information related to the estimated EDSS from MR images (See Figure

5.8). It will also include the disconnectome of WM tracts and structure volume deviation from normal distribution both described in Section 4.

5.5 Conclusion

In the final chapter of this thesis, we focused on how each contribution detailed in the previous chapters will serve the MS imaging community by offering MRI biomarker extraction and MS analysis pipelines. First, we motivated our choice to use containerization to keep our pipelines self-contained and independent of hardware and operating system constraints. Then, we explained why hosting our pipeline in web-based platforms such as volBrain would be beneficial for our users. Moreover, we showcased how our users can easily and freely execute our pipelines and obtain their results. Finally, we detailed the reports generated by our pipelines that offer compact, organized, and easy-to-read information.

Chapter 6

General Conclusion and Perspectives

6.1 Conclusion

In this thesis, we presented original DL methods to automate neuroimaging tasks related to MS.

In Chapter 2, we presented DeepLesionBrain, a DL framework for MS lesion segmentation designed for domain generalization. DeepLesionBrain uses spatially distributed 3D CNNs with large overlapping receptive fields, to produce consensus-based segmentation robust to domain shift. Our method is trained using hierarchical specialization learning to efficiently incorporate both generic and specialized features. Moreover, the proposed image quality data augmentation increases training data variability in a realistic way. Our validation showed the generalizability of our method, its robustness to domain shift, and the consistency of segmentation regardless of the training data.

In Chapter 3, we proposed Longitudinal DLB, a deep learning method for new MS lesion detection and segmentation from two time-points. The pipeline encompasses transfer learning from single time-point MS lesion segmentation, pretraining with time-point synthesis, and data augmentation adapted for MR images. This pipeline was very effective and obtained the best score in the MICCAI 2021 MSSEG2-challenge for the detection of new lesions.

In Chapter 4, we proposed a novel method for EDSS estimation from MRI data and clinico-demographic information using DL. The proposed architecture effectively combines neurodegenerative and neuroinflammatory image-based biomarkers and tabular information. The estimation model is trained with novel multi-phase learning and data augmentation to mitigate the data imbalance effect. During validation, our deep-learning method outperformed conventional state-of-the-

art methods. Furthermore, DL interpretability visualizations showed consistency with works associating MS functional impairments to brain structures.

In Chapter 5, we focused on how each contribution detailed in the previous chapters can benefit the MS imaging community by offering MRI biomarker extraction and MS analysis pipelines. We propose containerized pipelines freely available for execution via volBrain, an open-access and free web-based platform. Doing so allows our users to run the pipelines without caring about installation and hardware resources. The intuitive interface of volBrain helps to facilitate execution and retrieval of results. At last, the easy-to-read generated report offers a compact and organized representation of MRI biomarkers and MS analysis.

During my Ph.D. and across these chapters we raised several interesting points in both methodological computer-vision aspects and clinical neuroimaging standpoints.

Computer vision. Deep learning is able to automate neuroimaging tasks with high efficiency. Nevertheless, this tool does have a number of limitations. DL models are complex and resource intensive, thus they should be used only when simpler alternatives are significantly less efficient or importantly slower. Deep Learning methods do not always generalize well and may perform worse on input data domains that have not been seen during training. For this reason, it is of great importance to design an adequate validation procedure, preferably, with out-of-domain cross-dataset testing. Deep Learning methods are associated with the need for huge amounts of data to be trained, but we found that data quality and data balance are often more critical than data quantity. Moreover, as we have shown in [6] (see Appendix A), semi-supervised learning can be used to limit problems related to data scarcity.

Neuroimaging. Both neurodegeneration and neuroinflammation are strong components of MS. For this reason, acquiring multimodal MRI is necessary, with ideally at least T1-w and FLAIR. Unfortunately, for cost and time constraints, sometimes only FLAIR is acquired although T1 helps to obtain better lesion segmentation and is required to estimate atrophy of brain structures. As shown in Chapter 4, neurodegeneration is useful to track patient’s disability evolution. Using modality synthesis [Manjón et al., 2021] could be useful when only FLAIR is provided. Moreover, inflammation activity leads to systematic inaccuracies in GM structures segmentation, particularly for the subcortical structures as shown in [Dadar et al., 2021]. For this reason, we found during our experiments that whole brain segmentation should always be preceded by inpainting using the MS lesion mask to avoid segmentation inaccuracies.

6.2 Perspectives

Future works can be regrouped into two categories. The first represents works that are directly consistent with this thesis outline. They focus on offering information automatically extracted from MRI, MS-relevant analysis, and prediction of the patient's state. The second category of future works is linked but not exclusive to MS. It focuses on proposing methodological improvements to enhance the use of DL in automating neuroimaging tasks.

6.2.1 Extensions related to this Ph.D.

First, we would like to extend our proposed EDSS estimation model to predict future EDSS scores based on baseline information such as current EDSS, clinico-demographic data, and multi-modal MRI. Predicting future EDSS will help clinicians administer better-suited treatments and allow patients to take appropriate dispositions.

Second, similarly to the aforementioned point, we want to explore the possibility of estimating and predicting MS-related impairments using MRI data. For example, we want to separately quantify the disability or dysfunction for each of the visual, brainstem, pyramidal, cerebellar, sensory, bowel/bladder, cerebral, and ambulation functions.

Third, we want to investigate strategies to predict patient response to Disease Modifying Treatment (DMT) based on multimodal MRI. So far, there is still a high variance in how patients respond to the same DMT. Therefore, predictive models for the DMT response would allow efficient personalized MS therapy.

Moreover, we want to investigate the use of functional MRI (fMRI) both in the estimation and prediction of EDSS and MS-related impairments. Although fMRI lacks spatial resolution, the analysis of fMRI-biomarkers jointly with structural and clinico-demographic data could lead to better understanding and prognostic of MS.

Furthermore, we want to investigate the use of Diffusion Tensor Imaging (DTI) to extract further information related to WM tracts for MS. In fact, DTI is an MRI modality that allows the observation of water diffusion and extracting directional information that can be used to follow WM tracts. Thus, DTI could improve our EDSS estimation model, which currently uses an estimation of the WM tracts based on an average atlas registered using T1 sequence.

Finally, we want to extend DLB and Longitudinal DLB to the spinal cord and other pathologies of WMH such as vascular dementia, ischemia, or micro-hemorrhages.

6.2.2 General perspectives

In terms of general perspectives, we would like first to work on loss function-related problems. Indeed, we would like to design a specific loss function that accounts not only for voxel-wise similarity in terms of volume but also for structural features that are relevant to the task, such as object count and topology. For instance, in the case of MS, lesion count is an important feature when assessing the quality of automatic segmentation. Thus, incorporating these concepts in the loss function will most likely improve the optimization of a learning-based predictive model. However, designing such a loss function is not straightforward, since most optimization algorithms require formulating the loss function with differentiable operations. Hence, the challenge is to express structural features with these operations or find an approximation to do so.

Second, we want to investigate further how to improve the quality of data and design efficient data-centric guidelines for neuroimaging. Indeed, there is a growing debate about Model-centric *vs.* Data-centric Artificial Intelligence (AI). Model-centric AI focuses on using the right set of learning and optimization algorithms and the design of the neural network architecture. This approach has resulted in great advancement these recent years with the emergence of Attention Mechanisms, Transformers, Generative models, and Graph Neural Networks. On the other hand, data-centric AI focus on providing the right kind of data which leads to building high-quality, high-performance, and generalizable predictive models. Current state-of-the-art models such as U-net for segmentation and ResNet for classification have shown their stability and reproducibility for neuroimaging task automation. We started investigating this aspect in [6] using semi-supervised learning, data augmentation [2, 1], and dealing with data imbalance [4]. However, there is still a lot of work to be done on this topic. We believe that future works should focus on improving preprocessing and data cleaning, enforcing consistency on expert annotation, and designing better datasets for accurate data representation.

Appendix A

POPCORN: Progressive Pseudo-labeling with Consistency Regularization and Neighboring

This appendix corresponds to the following publication [6]:

Kamraoui, Reda Abdellah, et al. "POPCORN: Progressive Pseudo-labeling with Consistency Regularization and Neighboring." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.

A.1 Abstract

Semi-supervised learning (SSL) uses unlabeled data to compensate for the scarcity of annotated images and the lack of method generalization to unseen domains, two usual problems in medical segmentation tasks. In this work, we propose POPCORN, a novel method combining consistency regularization and pseudo-labeling designed for image segmentation. The proposed framework uses high-level regularization to constrain our segmentation model to use similar latent features for images with similar segmentations. POPCORN estimates a proximity graph to select data from easiest ones to more difficult ones, in order to ensure accurate pseudo-labeling and to limit confirmation bias. Applied to multiple sclerosis lesion segmentation, our method demonstrates competitive results compared to other state-of-the-art SSL strategies.

A.2 Introduction

Semi-Supervised Learning (SSL) is a promising field which aims to exploit unlabeled data in order to enhance the performance achieved using only labeled data. SSL is explored to mitigate both problems of the limited availability of labeled data and the lack of model generalization to unseen domains. Among SSL works proposed for medical image segmentation tasks, we can distinguish three main categories:

Consistency Regularization (CR) constrains the model to give consistent predictions for the same unlabeled input under different perturbations. [Bortsova et al., 2019] constrained the model to produce similar segmentation when applying different elastic transformations to the same unlabeled images. Similarly, [Perone and Cohen-Adad, 2018] used a mean teacher strategy where the consistency loss constrained teacher and student predictions to be consistent. [Orbes-Arteaga et al., 2019] designed an adversarial loss to minimize the amount of information for a specific domain and to maximize segmentation consistency. CR offers interesting consistency properties on the learned features but it is usually trained under unrealistic scenarios (e.g., using the same input data under different perturbations). Such oversimplification does not guarantee a good generalization of the learned features. Besides, some works showed that consistency regularization using perturbation on input data is not adapted for segmentation [French et al., 2020, Ouali et al., 2020].

Pseudo-Labeling (PL) strategies automatically assign labels to unlabeled data in order to use them during training in combination with labeled data. Pseudo-labels are generally assigned by a model trained on labeled data. Uncertainty can be used to measure the confidence of the predictions. For example, [Sedai et al., 2019] employed prediction uncertainty for estimating segmentation confidence on soft labels. [Cao et al., 2020] considered an uncertainty aware temporal ensembling strategy. [Xia et al., 2020] used uncertainty-weighted mechanism for the pseudo-label fusion of multiple networks predictions. PL is a simple way to use unlabeled data. PL is nevertheless prone to confirmation bias (*i.e.*, error propagation) [Arazo et al., 2020]. So far, this is the main limitation of PL.

Auxiliary Tasks (AT) are secondary objectives combined with the main segmentation task which do not require ground truth annotations. Using unlabeled data, in such a way, implicitly extracts relevant features for the primary segmentation task. [Li et al., 2020] proposed the prediction of surface distance maps to capture more effectively shape-aware features. [Kervadec et al., 2019] predicted the size of the target segmentation as an intermediate task. Alternatively, [Chen et al., 2019] combined supervised segmentation and unsupervised input reconstruction. Finally, [Luo et al., 2020] proposed to predict geometry-aware level set representation of the transformed ground truth annotations. AT demonstrated good

performance, but the choice of the AT highly depends on the addressed problem which limits the method generalization for other segmentation tasks.

In this work, our main contribution is threefold:

- We propose a novel framework that combines consistency regularization and pseudo-labeling for segmentation.
- We propose a consistency regularization strategy that ensures proximity in latent space of images with similar segmentations. This allows us to produce meaningful feature representation and accurate predictions.
- We propose a new pseudo-labeling strategy which selects progressively unlabeled samples according to their similarity with training data, in order to limit confirmation bias.

A.3 Method

A.3.1 Method overview

The proposed strategy is a PrOgressive Pseudo-labeling with COnsistency Regularization and Neighboring (POPCORN) for semi-supervised learning in segmentation (see Fig.A.1). First, the training is performed with a new CR ensuring that: *i*) augmented versions of the same image have identical feature maps, and *ii*) images with similar segmentations have similar feature maps. Second, PL of the unlabeled data is performed gradually. At each selection step, the proximity graph is used to select new unlabeled samples. The pseudo-labels of the chosen data-points are estimated with the current segmentation model and incorporated in the training set.

The main intuition is that our segmentation model is able to produce more accurate pseudo-labels for images similar to our training set. Since our CR ensures close features for similar data, features extracted from the model are used to select new samples.

A.3.2 Bottleneck consistency regularization

In POPCORN, the model architecture is based on 3D U-Net composed of an encoder and a decoder, linked by a bottleneck and skip connections at different scales (see Fig.A.1). For an image X , $F(X)$ represents the prediction of the segmentation network, and $h(X)$ represents the latent features of X extracted at the bottleneck level. Our method is based on a dual/hybrid loss ensuring segmentation quality and consistency relevance.

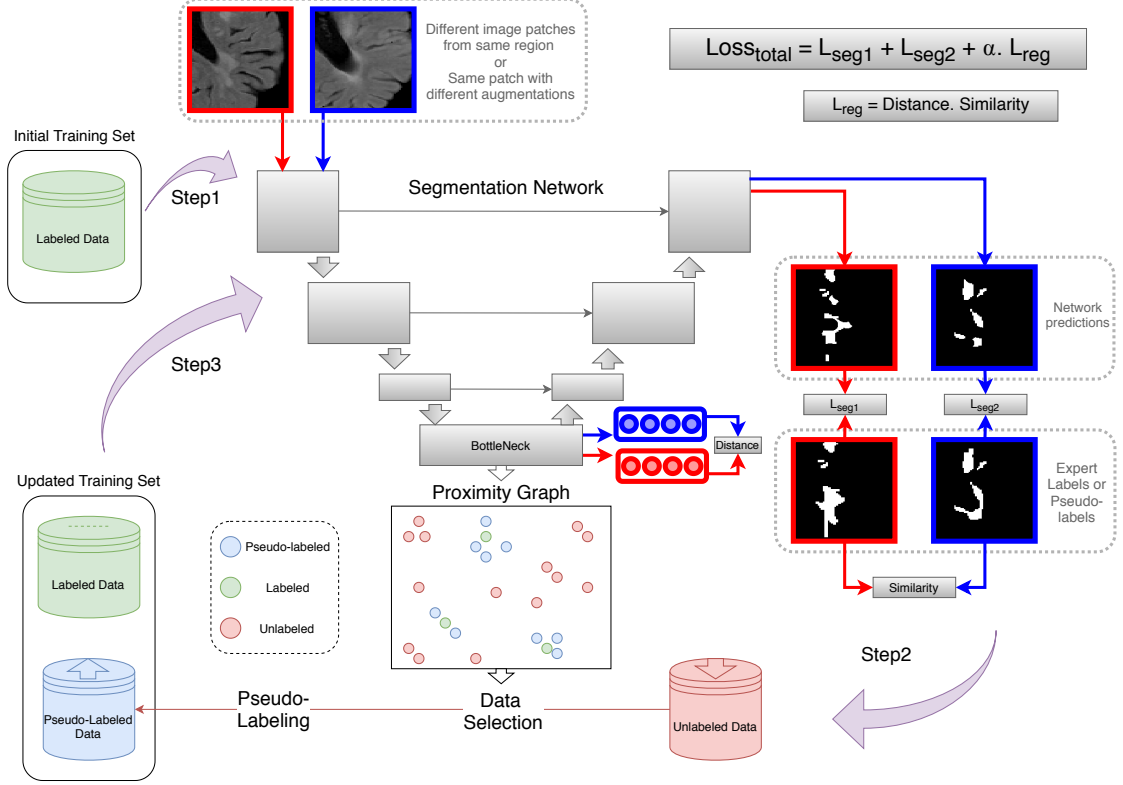


Figure A.1: The training process of POPCORN

A.3.2.1 Segmentation loss:

As traditionally done in supervised learning, we use the Dice similarity loss as the first element of our global loss. This loss ensures the similarity of the produced output with the expected one.

$$L_{\text{seg}}(F(X), y) = 1 - \text{Dice}(F(X), y), \quad (\text{A.1})$$

where y is either the expert segmentation of X when available, or the pseudo-label otherwise.

A.3.2.2 Consistency regularization loss:

Alongside L_{seg} , a regularization loss on the bottleneck is used:

$$L_{\text{reg}}[(X_i, y_i), (X_j, y_j)] = \text{Distance}(h(X_i), h(X_j)) \times \text{Similarity}(y_i, y_j), \quad (\text{A.2})$$

where X_i, X_j are image patches randomly selected as either different augmented versions of the same patch, or patches from different images extracted from the

same region and with the same orientation. Moreover, let us define:

$$\text{Similarity}(y_i, y_j) = e^{-mse(y_i, y_j)}, \quad (\text{A.3})$$

$$\text{Distance}(h_i, h_j) = \frac{2||h_i - h_j||^2}{||h_i||^2 + ||h_j||^2}, \quad (\text{A.4})$$

where mse is the mean squared error, and $||\cdot||$ is the euclidean distance.

The total loss is a combination of (1) and (2) with a weighting coefficient α :

$$\begin{aligned} L_{total}[(X_i, y_i), (X_j, y_j)] &= L_{seg}(F(X_i), y_i) + L_{seg}(F(X_j), y_j) \\ &+ \alpha \cdot L_{reg}[(X_i, y_i), (X_j, y_j)]. \end{aligned} \quad (\text{A.5})$$

A.3.3 Pseudo-labeling data selection

Curriculum learning [Bengio et al., 2009] showed that presenting data with an increasing difficulty can lead to a better learning process. We consider unlabeled data close to the training data as easy examples to be incorporated first in the training process, whereas distant samples are considered more challenging. Indeed, the latent distance between unlabeled and training data can be viewed as a measure of similarity. Thus, pseudo-labeling using the trained model is more accurate for unlabeled samples which are similar to training data.

Our data selection is performed in three steps (see Fig.A.1). Step1: the training set is limited to labeled data. Once the segmentation model is trained until convergence, it is used to extract latent space representation for each unlabeled datapoint. Step2: the proximity graph is used to select K unlabeled data that guarantee a smooth learning (as described in A.3.4). For each selected unlabeled datapoint, a pseudo-label (segmentation) is assigned by the trained model. Step3: the model is trained for N epochs with the new training set containing both labeled and pseudo-labeled data. Step two and three are repeated every N epochs by picking each time K new data points, their respective pseudo-labels are being computed with the newly updated segmentation model. The process is maintained until all unlabeled data are integrated into the training set.

A.3.4 Proximity graph

The proximity graph represents the euclidean distance between the training and the unlabeled samples latent representations:

$$P_{i \in U, j \in T} = ||h(X_i) - h(X_j)||^2 \quad (\text{A.6})$$

where T and U represent respectively the training set (labeled and pseudo-labeled), and unlabeled data. For data selection, we propose the following criteria to select

K elements of U close to T . For each datapoint of U , the proximity with T is defined by the sum of the p closest elements of T to the datapoint:

$$[i_1^*, i_2^*, \dots, i_K^*] = \underset{j \in \text{Argmin}^p(P_i)}{\text{Argmin}^K} \left(\sum P_{i,j} \right), \quad (\text{A.7})$$

where $[i_1^*, i_2^*, \dots, i_K^*]$ represent the K indices of the data selected. $\text{Argmin}^k(V)$ returns the indices of the k smallest values of the vector V .

A.4 Experiments

A.4.1 Dataset

Labeled Data: For labeled data, the ISBI training dataset [Carass et al., 2017] is used. It consists of 21 longitudinal multimodal images (including FLAIR modality) from only five different subjects with Multiple Sclerosis (MS). The images have been acquired on the same MRI scanner. MS lesions were delineated by two expert raters. This dataset has limited image quality diversity (all the images were acquired with the same protocol on a single site) and inter-subject variability (only 5 subjects).

Unlabeled Data: The unlabeled dataset consists of 2901 FLAIR MRI (large inter-subject variability) with white matter hyperintensities. It does not only contain MS, which increases pathology diversity. MRI have been collected across multiple acquisition sites based on different manufacturers, 1.5T and 3T scanners, 2D and 3D sequences. This dataset covers a large diversity in terms of image quality, pathology and inter-subject variability.

Testing Data: For assessing our results, the dataset described in [Coupé et al., 2018] is used. It contains 3D multimodal MRI from 43 subjects diagnosed with MS. The images have been acquired with three different scanners and different acquisition protocols. Consequently this dataset proposes a larger diversity than the labeled dataset. Lesion masks have been obtained by expert manual delineation.

All images have been pre-processed using the same pipeline [Coupé et al., 2018].

A.4.2 Reference Methods

POPCORN is compared to state-of-the-art strategies [Chen et al., 2019], [Sedai et al., 2019] and [Bortsova et al., 2019]. The following strategies have been implemented based on their published works and adapted to MS lesion segmentation. First, the multi-task attention-based SSL [Chen et al., 2019] is an AT strategy. It combines supervised segmentation and unsupervised reconstruction objectives.

The reconstruction task uses an attention mechanism to predict input image regions of different classes. Second, the uncertainty guided pseudo-labeling [Sedai et al., 2019] is a PL strategy. The teacher model, trained only on labeled data, generates soft segmentation (pseudo-labels) and uncertainty maps for all the unlabeled data at once. The uncertainty is used for estimating segmentation confidence of the generated segmentation when training the student model. Finally, the semi-supervised transformation consistency [Bortsova et al., 2019] is based on CR. In addition to the primary loss, a consistency loss ensures that the prediction of the same images under transformations are consistent.

A.4.3 Implementation details

The method hyperparameters were chosen empirically according to the size of labeled and unlabeled datasets. First, 200 from the $M = 2901$ unlabeled images were chosen after each training cycle that ran for 2 epochs ($K = 200$, $N = 2$) to limit computational burden. Second, the number of neighbors $p = 5$ was selected considering the initial training data of 21 labeled images. We suggest that this value is a good compromise in order to consider relevant near neighbors while avoiding far neighbors which mislead data selection.

In addition, we used the architecture proposed by [2] with a patch size of $[64 \times 64 \times 64]$ and a threshold of 0.5 to obtain the binary segmentation. Moreover, image quality data augmentation was used to introduce realistic perturbations, where blur, edge enhancement, and other augmentations simulated image quality heterogeneity [2]. Furthermore, the coefficients for the regularization part of the loss have been set to 0.2 ($\alpha = 0.2$). Finally, the experiments have been performed with Keras 2.2.4 [Chollet et al., 2015] and Tensorflow 1.12.0 [Abadi et al., 2016] on Python 3.6. The model was optimized with Adam [Kingma and Ba, 2014] using a learning rate of 0.0001 and a momentum of 0.9.

A.4.4 Statistical Analysis

To assert the advantage of a technique obtaining the highest average score, we conducted a Wilcoxon test over the lists of Dice scores measured at image level. The significance of the test is established for a p-value below 0.05.

Table A.1: The table represents an ablation study of the key components of POPCORN. The table details the impact of each contribution: the consistency regularization (CR), the proximity graph, and using labeled/pseudo-labeled (lab/pseudo) data. "Ours with half the data" indicates the performance of our method when half of the selection steps are passed ($M = 1400$). Best result is displayed in bold, and the second best result is underlined.

Strategy	Trained on	CR on	Dice	Precision	Sensitivity
Our method	Lab + Pseudo	Lab + Pseudo	73,09%	73,33%	74,29%
Ours with half the data	Lab + Pseudo	Lab + Pseudo	70,59%	68,26%	75,91%
Ours without CR	Lab + Pseudo	None	69,13%	70,49%	70,58%
Ours without proximity graph	Lab + Pseudo	Lab + Pseudo	68,06%	65,14%	74,40%
Baseline with CR	Lab	Lab	68,08%	77,77%	61,94%
Baseline	Lab	None	64,41%	61,80%	69,70%

A.5 Results

A.5.1 Ablation study

To evaluate our contributions, we compare POPCORN with other versions of our strategy when isolating key elements. As shown in Table A.1, our full method achieves the highest Dice and the second best result in terms of precision. First, when comparing POPCORN without consistency regularization (corresponds to $\alpha = 0$ in (A.5)) and our full method, we notice a decrease in both precision and sensitivity. This suggests that without CR, the latent space is less meaningful for our selection process of unlabeled data. Second, to underline the impact of the proximity graph, we consider another progressive PL strategy where pseudo labels are randomly selected. Although the strategy without proximity graph is slightly more sensitive, we observe an important drop in both Dice and precision compared to our full method. This demonstrates that the proposed progressive selection based on image proximity in latent space is more robust to confirmation bias than random selection. Next, when running only half the selection steps ($M = 1400$), our method obtained the second best Dice score. This shows that POPCORN with nearly half unlabeled data can achieve better performance than the other variations and methods with full dataset (see also A.5.2). Finally, when combining the proposed CR (on labeled data only) with the baseline (supervised learning), the precision is considerably improved. This shows the importance of our CR on segmentation accuracy, beyond data selection. Overall, the statistical analysis shows that our full method has a significantly higher Dice than the baseline, the version without CR, baseline with CR, and Ours without proximity graph.

A.5.2 Comparison with state-of-the-art approaches

Table A.2 shows the results of POPCORN compared to the reference methods presented in section A.4.2. First, all the SSL strategies obtain a significantly better Dice scores compared to the baseline. Second, POPCORN obtains the highest Dice followed by Uncertainty guided Pseudo-labeling [Sedai et al., 2019]. Next, the multi-task attention-based SSL [Chen et al., 2019] and the semi-supervised transformation consistency [Bortsova et al., 2019] respectively obtain the highest precision and sensitivity rates. Finally, POPCORN obtains the best balance between precision and sensitivity, as opposed to the other strategies which are more prone to FP [Bortsova et al., 2019, Sedai et al., 2019] and FN [Chen et al., 2019]. Overall, POPCORN has a significantly higher Dice compared to the other methods according to our Wilcoxon test.

Table A.2: The table represents results of POPCORN (our method) compared to other state-of-the-art strategies on the testing dataset (see A.5.2 for complementary details).

Strategy	Dice	Precision	Sensitivity
POPCORN	73,09%	73,33%	74,29%
Multi-task Attention-based SSL [Chen et al., 2019]	67,23%	75,72%	61,99%
Uncertainty guided Pseudo-labeling [Sedai et al., 2019]	68,31%	67,93%	71,95%
Semi-supervised transformation consistency [Bortsova et al., 2019]	66,75%	61,52%	78,79%
Baseline (labeled data only)	64,41%	61,80%	69,70%

Fig.A.2 shows image segmentations produced by POPCORN and the compared strategies. A, B, and C are images from the testing dataset, specifically chosen to showcase acquisition and lesion diversity. For A, we observe that POPCORN segmentation is the most accurate. On the contrary, [Chen et al., 2019, Sedai et al., 2019] are the least sensitive with high volumes of false negative. Similarly, the segmentations obtained with the baseline and [Bortsova et al., 2019] do not cover all lesions. On image B, the segmentation provided by [Bortsova et al., 2019] contains several false positive lesions, compared to the other strategies. Both the baseline and [Sedai et al., 2019] only include one or two false detections. POPCORN proposes an accurate segmentation. Last, the method [Chen et al., 2019] misses a small lesion. For C, we notice that [Bortsova et al., 2019, Sedai et al., 2019] and the baseline detect many false positive lesions. POPCORN and [Chen et al., 2019] produce fewer false detection on this challenging sample. To conclude, our strategy segments accurately most lesions while minimizing false detection. Compared to the other strategies, POPCORN maintains the best balance between the sensitivity and the precision of lesion segmentation.

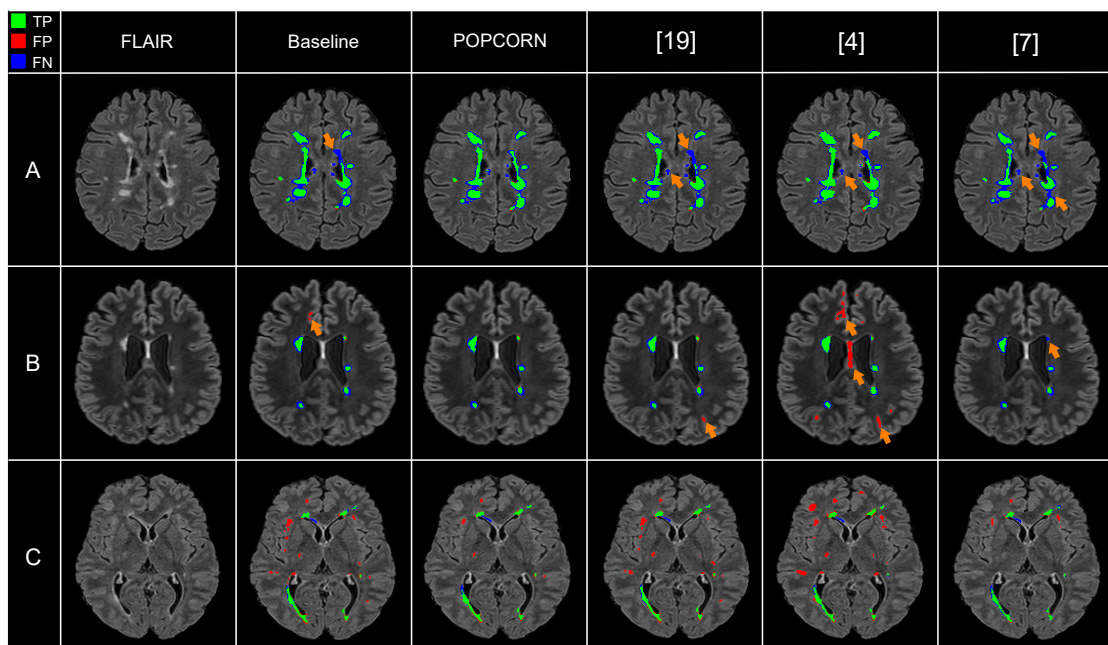


Figure A.2: Comparison of POPCORN, Uncertainty guided Pseudo-labeling [Sedai et al., 2019], multi-task attention-based SSL [Chen et al., 2019], and the semi-supervised transformation consistency [Bortsova et al., 2019] lesion segmentations. Orange arrows indicate key segmentation differences.

A.6 Conclusion

We propose a novel strategy for SSL segmentation. Our method combines consistency regularization and pseudo-labeling. POPCORN progressively selects unlabeled samples with an increasing difficulty using a proximity graph. Overall, we have shown the improvement of using POPCORN compared to other state-of-the-art strategies, as well as the impact of each of our contributions.

Acronyms

CIS Clinically Isolated Syndrome. 20

CNN Convolutional Neural Networks. 31

CNS Central Nervous System. 9

CORR Pearson’s correlation coefficient. 44

DA Data Augmentation. 69

DIR Double Inversion Recovery. 16

DIS Dissemination In Space. 21

DIT Dissemination In Time. 21

DLB DeepLesionBrain. 35

DMT Disease Modifying Treatment. 84

DTI Diffusion Tensor Imaging. 16

FLAIR Fluid-Attenuated Inversion Recovery. 16

GN Group Normalization. 39

GPU Graphical Processor Unit. 26

HSL Hierarchical Specialization Learning. 35

IQDA Image Quality Data Augmentation. 35

LesF1 Lesion Detection F1. 73

LFPR Lesion False Positive Rate. 44

LTPR Lesion True Positive Rate. 44

ML Machine Learning. 24

MNI space Montreal Neurological Institute template space. 39

MRI Magnetic Resonance Imaging. 9

MS Multiple Sclerosis. 9

MTR Magnetization Transfer Ratio. 16

OFSEP "Observatoire français de la sclérose en Plaques". 85

PD Proton Density-weighted. 16

PPMS Primary Progressive MS. 20

PPV Positive Predictive Value. 44

ReLU Rectified Linear Units. 39

RRMS Relapsing-Remitting MS. 20

SD Standard Deviation. 69

SPMS Secondary Progressive MS. 20

T1 T1-weighted. 16

T2 T2-weighted. 16

TPR True Positive Rate. 44

WMH White Matter Hyperintensities. 16

References

- [Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.
- [Akkus et al., 2017] Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of digital imaging*, 30(4):449–459.
- [Al-Midfai et al., 2022] Al-Midfai, Y., Kujundzic, W., Uppal, S., Oakes, D., and Giezy, S. (2022). Acute multiple sclerosis exacerbation after vaccination with the Johnson & Johnson COVID-19 vaccine: Novel presentation and first documented case report. *Cureus*, 14(4).
- [Al-Zubi et al., 2002] Al-Zubi, S., Toennies, K., Bodammer, N., and Hinrichs, H. (2002). Fusing markov random fields with anatomical knowledge and shape based analysis to segment multiple sclerosis white matter lesions in magnetic resonance images of the brain. In *Bildverarbeitung für die Medizin 2002*, pages 185–188. Springer.
- [Alves et al., 2022] Alves, P., Green, E., Leavy, M., Friedler, H., Curhan, G., Marci, C., and Boussios, C. (2022). Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis. *Multiple Sclerosis Journal—Experimental, Translational and Clinical*, 8(2):20552173221108635.
- [Amadasun and King, 1989] Amadasun, M. and King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, 19(5):1264–1274.
- [Andravizou et al., 2019] Andravizou, A., Dardiotis, E., Artemiadis, A., Sokratous, M., Siokas, V., Tsouris, Z., Aloizou, A.-M., Nikolaidis, I., Bakirtzis, C., Tsivgoulis, G., et al. (2019). Brain atrophy in multiple sclerosis: mechanisms, clinical relevance and treatment options. *Autoimmunity Highlights*, 10(1):1–25.

-
- [Arazo et al., 2020] Arazo, E., Ortego, D., Albert, P., O’Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- [Ashburner, 2002] Ashburner, J. (2002). Another MRI bias correction approach. In *8th International Conference on Functional Mapping of the Human Brain*, volume 16.
- [Aslani et al., 2019] Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M. A., and Sona, D. (2019). Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196:1–15.
- [Aslani et al., 2020] Aslani, S., Murino, V., Dayan, M., Tam, R., Sona, D., and Hamarneh, G. (2020). Scanner invariant multiple sclerosis lesion segmentation from MRI. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 781–785. IEEE.
- [Avants et al., 2008] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41.
- [Avants et al., 2011] Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044.
- [Avants et al., 2014] Avants, B. B., Tustison, N. J., Stauffer, M., Song, G., Wu, B., and Gee, J. C. (2014). The insight toolkit image registration framework. *Frontiers in neuroinformatics*, 8:44.
- [Bakshi et al., 2001] Bakshi, R., Benedict, R. H., Bermel, R. A., and Jacobs, L. (2001). Regional brain atrophy is associated with physical disability in multiple sclerosis: semiquantitative magnetic resonance imaging and relationship to clinical findings. *Journal of Neuroimaging*, 11(2):129–136.
- [Barker-Collo, 2006] Barker-Collo, S. L. (2006). Quality of life in multiple sclerosis: does information-processing speed have an independent effect? *Archives of clinical neuropsychology*, 21(2):167–174.
- [Barkhof et al., 1997] Barkhof, F., Filippi, M., Miller, D. H., Scheltens, P., Campi, A., Polman, C. H., Comi, G., Ader, H. J., Losseff, N., and Valk, J. (1997). Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain: a journal of neurology*, 120(11):2059–2069.

- [Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.
- [Benoit-Cattin, 2006] Benoit-Cattin, H. (2006). *Texture analysis for magnetic resonance imaging*. Texture Analysis Magn Resona.
- [Bermel and Bakshi, 2006] Bermel, R. A. and Bakshi, R. (2006). The measurement and clinical relevance of brain atrophy in multiple sclerosis. *The Lancet Neurology*, 5(2):158–170.
- [Bodis-Wollner and Brannan, 1997] Bodis-Wollner, I. and Brannan, J. R. (1997). Hidden visual loss in optic neuropathy is revealed using gabor patch contrast perimetry. *Clinical Neuroscience (New York, NY)*, 4(5):284–291.
- [Boesen et al., 2004] Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Lüders, E., and Rottenberg, D. (2004). Quantitative comparison of four brain extraction algorithms. *NeuroImage*, 22(3):1255–1261.
- [Bortsova et al., 2019] Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., and de Bruijne, M. (2019). Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 810–818. Springer.
- [Bosc et al., 2003] Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D., and Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage*, 20(2):643–656.
- [Bottou, 2010] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer.
- [Boucekine et al., 2013] Boucekine, M., Loundou, A., Baumstarck, K., Minaya-Flores, P., Pelletier, J., Ghattas, B., and Auquier, P. (2013). Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. *BMC medical research methodology*, 13(1):20.
- [Branco et al., 2017] Branco, P., Torgo, L., and Ribeiro, R. P. (2017). SMOGN: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pages 36–50. PMLR.

-
- [Brisset et al., 2020] Brisset, J.-C., Kremer, S., Hannoun, S., Bonneville, F., Durand-Dubief, F., Tourdias, T., Barillot, C., Guttman, C., Vukusic, S., Dousset, V., et al. (2020). New OFSEP recommendations for MRI assessment of multiple sclerosis patients: special consideration for gadolinium deposition and frequent acquisitions. *Journal of Neuroradiology*, 47(4):250–258.
- [Bron et al., 2021] Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., Aalten, P., De Deyn, P. P., Biessels, G. J., Claassen, J. A., et al. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer’s disease. *NeuroImage: Clinical*, 31:102712.
- [Brosch et al., 2016] Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., and Tam, R. (2016). Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239.
- [Byrd and Lipton, 2019] Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR.
- [Cabezas et al., 2021] Cabezas, M., Luo, Y., Kyle, K., Ly, L., Wang, C., and Barnett, M. (2021). Estimating lesion activity through feature similarity: A dual path Unet approach for the MSSEG2 MICCAI challenge. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 107.
- [Cao et al., 2020] Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., and Cheng, L. (2020). Uncertainty aware temporal-ensembling model for semi-supervised abuse mass segmentation. *IEEE Transactions on Medical Imaging*, 40(1):431–443.
- [Carass et al., 2017] Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C. H., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102.
- [Carass et al., 2007] Carass, A., Wheeler, M. B., Cuzzocreo, J., Bazin, P.-L., Bassett, S. S., and Prince, J. L. (2007). A joint registration and segmentation approach to skull stripping. In *2007 4th IEEE international symposium on biomedical imaging: from nano to macro*, pages 656–659. IEEE.
- [Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

- [Chen et al., 2019] Chen, S., Bortsova, G., Juárez, A. G.-U., van Tulder, G., and de Bruijne, M. (2019). Multi-task attention-based semi-supervised learning for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–465. Springer.
- [Cheng et al., 2018] Cheng, M., Galimzianova, A., Lesjak, Ž., Špiclin, Ž., Lock, C. B., and Rubin, D. L. (2018). A multi-scale multiple sclerosis lesion change detection in a multi-sequence MRI. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 353–360. Springer.
- [Chiaravalloti and DeLuca, 2008] Chiaravalloti, N. D. and DeLuca, J. (2008). Cognitive impairment in multiple sclerosis. *The Lancet Neurology*, 7(12):1139–1151.
- [Chollet et al., 2015] Chollet, F. et al. (2015). Keras. <https://keras.io>.
- [Ciccarelli et al., 2001] Ciccarelli, O., Werring, D., Wheeler-Kingshott, C., Barker, G., Parker, G., Thompson, A., and Miller, D. (2001). Investigation of MS normal-appearing brain using diffusion tensor MRI with clinical correlations. *Neurology*, 56(7):926–933.
- [Çiçek et al., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer.
- [Commowick et al., 2016] Commowick, O., Cervenansky, F., and Ameli, R. (2016). MSSEGchallenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure.
- [Commowick et al., 2021] Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 1–118.
- [Commowick et al., 2018] Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Ameli, R., Ferré, J.-C., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1):1–17.
- [Commowick et al., 2012] Commowick, O., Wiest-Daesslé, N., and Prima, S. (2012). Block-matching strategies for rigid registration of multimodal medical images. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 700–703. IEEE.

-
- [Confavreux et al., 2003] Confavreux, C., Vukusic, S., and Adeleine, P. (2003). Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. *Brain*, 126(4):770–782.
- [Coupé et al., 2017] Coupé, P., Catheline, G., Lanuza, E., Manjón, J. V., and Initiative, A. D. N. (2017). Towards a unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis. *Human brain mapping*, 38(11):5501–5518.
- [Coupé et al., 2020] Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B. D., Ta, V.-T., Lepetit, V., and Manjon, J. V. (2020). AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation. *NeuroImage*, page 117026.
- [Coupé et al., 2018] Coupé, P., Tourdias, T., Linck, P., Romero, J. E., and Manjón, J. V. (2018). Lesionbrain: an online tool for white matter lesion segmentation. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 95–103. Springer.
- [Coupé et al., 2008] Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., and Barillot, C. (2008). An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. *IEEE transactions on medical imaging*, 27(4):425–441.
- [Dadar et al., 2021] Dadar, M., Potvin, O., Camicioli, R., Duchesne, S., and Initiative, A. D. N. (2021). Beware of white matter hyperintensities causing systematic errors in freesurfer gray matter segmentations! *Human brain mapping*, 42(9):2734–2745.
- [Dalbis et al., 2021] Dalbis, T., Fritz, T., Grilo, J., Hitziger, S., and Ling, W. X. (2021). Triplanar U-net with orientation aggregation for new lesions segmentation. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 57.
- [Dalton et al., 2004] Dalton, C. M., Chard, D. T., Davies, G. R., Miszkiet, K. A., Altmann, D. R., Fernando, K., Plant, G. T., Thompson, A. J., and Miller, D. H. (2004). Early development of multiple sclerosis is associated with progressive grey matter atrophy in patients presenting with clinically isolated syndromes. *Brain*, 127(5):1101–1107.
- [Danelakis et al., 2018] Danelakis, A., Theoharis, T., and Verganelakis, D. A. (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 70:83–100.

- [Desolneux et al., 2003] Desolneux, A., Moisan, L., and More, J.-M. (2003). A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513.
- [Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [Elliott et al., 2013] Elliott, C., Arnold, D. L., Collins, D. L., and Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE transactions on medical imaging*, 32(8):1490–1503.
- [Eloyan et al., 2014] Eloyan, A., Shou, H., Shinohara, R. T., Sweeney, E. M., Nebel, M. B., Cuzzocreo, J. L., Calabresi, P. A., Reich, D. S., Lindquist, M. A., and Crainiceanu, C. M. (2014). Health effects of lesion localization in multiple sclerosis: spatial registration and confounding adjustment. *PloS one*, 9(9):e107263.
- [Eshaghi et al., 2021] Eshaghi, A., Young, A. L., Wijeratne, P. A., Prados, F., Arnold, D. L., Narayanan, S., Guttmann, C. R., Barkhof, F., Alexander, D. C., Thompson, A. J., et al. (2021). Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nature communications*, 12(1):1–12.
- [Feng et al., 2019] Feng, Y., Pan, H., Meyer, C., and Feng, X. (2019). A self-adaptive network for multiple sclerosis lesion segmentation from multi-contrast MRI with various imaging sequences. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 472–475. IEEE.
- [Filippi et al., 2019a] Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., De Stefano, N., Geurts, J. J., Paul, F., Reich, D. S., Toosy, A. T., et al. (2019a). Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain*, 142(7):1858–1875.
- [Filippi et al., 2019b] Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., De Stefano, N., Geurts, J. J. G., Paul, F., Reich, D. S., Toosy, A. T., Traboulsee, A., Wattjes, M. P., Yousry, T. A., Gass, A., Lubetzki, C., Weinshenker, B. G., and Rocca, M. A. (2019b). Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines.
- [Fischl, 2012] Fischl, B. (2012). Freesurfer. *Neuroimage*, 62(2):774–781.

-
- [Fisniku et al., 2008] Fisniku, L., Brex, P., Altmann, D., Miszkiet, K., Benton, C., Lanyon, R., Thompson, A., and Miller, D. (2008). Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain*, 131(3):808–817.
- [French et al., 2020] French, G., Aila, T., Laine, S., Mackiewicz, M., and Finlayson, G. (2020). Semi-supervised semantic segmentation needs strong, high-dimensional perturbations.
- [Galloway, 1975] Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179.
- [Ganiler et al., 2014] Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., Ramió-Torrentà, L., Rovira, À., and Lladó, X. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology*, 56(5):363–374.
- [García-Lorenzo, 2010] García-Lorenzo, D. (2010). *Robust segmentation of focal lesions on multi-sequence MRI in multiple sclerosis*. PhD thesis, Université Rennes 1.
- [García-Lorenzo et al., 2013] García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical image analysis*, 17(1):1–18.
- [Gedamu, 2011] Gedamu, E. (2011). Guidelines for developing automated quality control procedures for brain magnetic resonance images acquired in multi-centre clinical trials. *Applications and Experiences of Quality Control*, ed Ognyan Ivanov (Rijeka: InTech), pages 135–158.
- [Ghafoorian et al., 2017] Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I. W., Sanchez, C. I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., and Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):1–12.
- [Goldberg-Zimring et al., 1998] Goldberg-Zimring, D., Achiron, A., Miron, S., Faibel, M., and Azhari, H. (1998). Automated detection and characterization of multiple sclerosis lesions in brain MR images. *Magnetic resonance imaging*, 16(3):311–318.
- [Goldenberg, 2012] Goldenberg, M. M. (2012). Multiple sclerosis review. *Pharmacy and Therapeutics*, 37(3):175.

- [González-Villà et al., 2016] González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., and Lladó, X. (2016). A review on brain structures segmentation in magnetic resonance imaging. *Artificial intelligence in medicine*, 73:45–69.
- [Griffanti et al., 2016] Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M., et al. (2016). Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage*, 141:191–205.
- [Hao et al., 2019] Hao, J., Kosaraju, S. C., Tsaku, N. Z., Song, D. H., and Kang, M. (2019). Page-net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In *Pacific Symposium on Biocomputing 2020*, pages 355–366. World Scientific.
- [Haralick et al., 1973] Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.
- [Hashemi et al., 2018] Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Prabhu, S. P., Warfield, S. K., and Gholipour, A. (2018). Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735.
- [Havaei et al., 2017] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31.
- [Herbet et al., 2016] Herbert, G., Maheu, M., Costi, E., Lafargue, G., and Duffau, H. (2016). Mapping neuroplastic potential in brain-damaged patients. *Brain*, 139(3):829–844.
- [Houtchens et al., 2007] Houtchens, M., Benedict, R., Killiany, R., Sharma, J., Jaisani, Z., Singh, B., Weinstock-Guttman, B., Guttmann, C., and Bakshi, R. (2007). Thalamic atrophy and cognition in multiple sclerosis. *Neurology*, 69(12):1213–1223.
- [Hu et al., 2018] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

-
- [Huang et al., 2017] Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.
- [Huo et al., 2019] Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S. M., Cutting, L. E., and Landman, B. A. (2019). 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119.
- [Hutchinson, 2015] Hutchinson, M. (2015). Neurodegeneration in multiple sclerosis is a process separate from inflammation: No. *Multiple Sclerosis Journal*, 21(13):1628–1631.
- [Iheme and Unay, 2005] Iheme, L. O. and Unay, D. (2005). Automatic white matter hyperintensity segmentation using FLAIR MRI: The MS lesion segmentation challenge. *NeuroImage*, 28(3):607–17.
- [Ion-Mărgineanu et al., 2017a] Ion-Mărgineanu, A., Kocevar, G., Stamile, C., Sima, D. M., Durand-Dubief, F., Huffel, S. V., and Sappey-Marini er, D. (2017a). A comparison of machine learning approaches for classifying multiple sclerosis courses using MRSI and brain segmentations. In *International Conference on Artificial Neural Networks*, pages 643–651. Springer.
- [Ion-Mărgineanu et al., 2017b] Ion-Mărgineanu, A., Kocevar, G., Stamile, C., Sima, D. M., Durand-Dubief, F., Van Huffel, S., and Sappey-Marini er, D. (2017b). Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. *Frontiers in neuroscience*, 11:398.
- [Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.
- [Isensee et al., 2019] Isensee, F., Petersen, J., Kohl, S. A., J ager, P. F., and Maier-Hein, K. H. (2019). nnU-Net: Breaking the spell on successful medical image segmentation. *arXiv preprint arXiv:1904.08128*, 1:1–8.
- [Jacobsen et al., 2014] Jacobsen, C., Hagemeyer, J., Myhr, K.-M., Nyland, H., Lode, K., Bergsland, N., Ramasamy, D. P., Dalaker, T. O., Larsen, J. P., Farbu, E., et al. (2014). Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. *Journal of Neurology, Neurosurgery & Psychiatry*, 85(10):1109–1115.
-

- [Johnson and Khoshgoftaar, 2019] Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.
- [Johnston et al., 1996] Johnston, B., Atkins, M. S., Mackiewicz, B., and Anderson, M. (1996). Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. *IEEE transactions on medical imaging*, 15(2):154–169.
- [Kamnitsas et al., 2015] Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., and Glocker, B. (2015). Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. *Ischemic stroke lesion segmentation*, 13:46.
- [Kamnitsas et al., 2016] Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., Rueckert, D., and Glocker, B. (2016). Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer.
- [Kamnitsas et al., 2017] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78.
- [Kaur et al., 2020] Kaur, A., Kaur, L., and Singh, A. (2020). State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions. *Archives of Computational Methods in Engineering*, pages 1–27.
- [Kervadec et al., 2019] Kervadec, H., Dolz, J., Granger, É., and Ayed, I. B. (2019). Curriculum semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–576. Springer.
- [Keskar et al., 2016] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.
- [Khan et al., 2021] Khan, A. R., Khan, S., Harouni, M., Abbasi, R., Iqbal, S., and Mehmood, Z. (2021). Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification. *Microscopy Research and Technique*, 84(7):1389–1399.
- [Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.

-
- [Kocevar et al., 2016] Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., and Sappey-Marini r, D. (2016). Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses. *Frontiers in neuroscience*, 10:478.
- [Koudriavtseva and Mainero, 2016] Koudriavtseva, T. and Mainero, C. (2016). Neuroinflammation, neurodegeneration and regeneration in multiple sclerosis: intercorrelated manifestations of the immune response. *Neural regeneration research*, 11(11):1727.
- [Kr ger et al., 2020] Kr ger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., Schlaefer, A., and Schippling, S. (2020). Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage: Clinical*, 28:102445.
- [Kurtzke, 1983] Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1444.
- [Laine and Fan, 1993] Laine, A. and Fan, J. (1993). Texture classification by wavelet packet signatures. *IEEE Transactions on pattern analysis and machine intelligence*, 15(11):1186–1191.
- [Lee et al., 2016] Lee, H., Park, M., and Kim, J. (2016). Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *2016 IEEE international conference on image processing (ICIP)*, pages 3713–3717. IEEE.
- [Leray et al., 2010] Leray, E., Yaouanq, J., Le Page, E., Coustans, M., Laplaud, D., Oger, J., and Edan, G. (2010). Evidence for a two-stage disability progression in multiple sclerosis. *Brain*, 133(7):1900–1913.
- [Li et al., 2018] Li, H., Jialin Pan, S., Wang, S., and Kot, A. C. (2018). Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.
- [Li et al., 2020] Li, S., Zhang, C., and He, X. (2020). Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer.
- [Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

- [Louapre and Lubetzki, 2015] Louapre, C. and Lubetzki, C. (2015). Neurodegeneration in multiple sclerosis is a process separate from inflammation: Yes. *Multiple Sclerosis Journal*, 21(13):1626–1628.
- [Lublin et al., 2014] Lublin, F. D., Reingold, S. C., Cohen, J. A., Cutter, G. R., Sørensen, P. S., Thompson, A. J., Wolinsky, J. S., Balcer, L. J., Banwell, B., Barkhof, F., et al. (2014). Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, 83(3):278–286.
- [Lucchinetti et al., 2011] Lucchinetti, C. F., Popescu, B. F., Bunyan, R. F., Moll, N. M., Roemer, S. F., Lassmann, H., Brück, W., Parisi, J. E., Scheithauer, B. W., Giannini, C., et al. (2011). Inflammatory cortical demyelination in early multiple sclerosis. *New England Journal of Medicine*, 365(23):2188–2197.
- [Luo et al., 2020] Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., and Zhang, S. (2020). Semi-supervised medical image segmentation through dual-task consistency. *arXiv preprint arXiv:2009.04448*.
- [Manjón, 2017] Manjón, J. V. (2017). MRI preprocessing. In *Imaging Biomarkers*, pages 53–63. Springer.
- [Manjón and Coupé, 2016] Manjón, J. V. and Coupé, P. (2016). volBrain: an online MRI brain volumetry system. *Frontiers in neuroinformatics*, 10:30.
- [Manjón and Coupé, 2018] Manjón, J. V. and Coupé, P. (2018). MRI denoising using deep learning. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 12–19. Springer.
- [Manjón et al., 2010] Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L., and Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, 31(1):192–203.
- [Manjón et al., 2014] Manjón, J. V., Eskildsen, S. F., Coupé, P., Romero, J. E., Collins, D. L., and Robles, M. (2014). Nonlocal intracranial cavity extraction. *International journal of biomedical imaging*, 2014.
- [Manjón et al., 2021] Manjón, J. V., Romero, J. E., and Coupe, P. (2021). Deep learning based MRI contrast synthesis using full volume prediction using full volume prediction. *Biomedical Physics & Engineering Express*, 8(1):015013.
- [Manjón et al., 2022] Manjón, J. V., Romero, J. E., and Coupe, P. (2022). A novel deep learning based hippocampus subfield segmentation method. *Scientific Reports*, 12(1):1–9.

- [Manjón et al., 2020] Manjón, J. V., Romero, J. E., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., Tourdias, T., and Coupé, P. (2020). Blind MRI brain lesion inpainting using deep learning. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 41–49. Springer.
- [Mårtensson et al., 2020] Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M. G., et al. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Medical Image Analysis*, page 101714.
- [Masson et al., 2021] Masson, A., Le Bon, B., Kerbrat, A., Edan, G., Galassi, F., and Combes, B. (2021). A nnUnet implementation of new lesions segmentation from serial FLAIR images of MS patients. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 5.
- [Mayerhoefer et al., 2020] Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., and Cook, G. (2020). Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4):488–495.
- [Medical, 2014] Medical, B. (2014). Medical gallery of blausen medical 2014. *WikiJournal of Medicine*, 1(2):1–79.
- [MICCAI, 2021] MICCAI (2021). Longitudinal multiple sclerosis lesion segmentation challenge. <https://portal.fli-iam.irisa.fr/msseg-2/data/>.
- [Minagar et al., 2013] Minagar, A., Barnett, M. H., Benedict, R. H., Pelletier, D., Pirko, I., Sahraian, M. A., Frohman, E., and Zivadinov, R. (2013). The thalamus and multiple sclerosis: modern views on pathologic, imaging, and clinical aspects. *Neurology*, 80(2):210–219.
- [Motiian et al., 2017] Motiian, S., Piccirilli, M., Adjero, D. A., and Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725.
- [Muandet et al., 2013] Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18.
- [Narayana et al., 2020] Narayana, P. A., Coronado, I., Sujit, S. J., Sun, X., Wolinsky, J. S., and Gabr, R. E. (2020). Are multi-contrast magnetic resonance images

- necessary for segmenting multiple sclerosis brains? a large cohort study based on deep learning. *Magnetic resonance imaging*, 65:8–14.
- [Noseworthy et al., 1990] Noseworthy, J., Vandervoort, M., Wong, C., and Ebers, G. (1990). Interrater variability with the expanded disability status scale (EDSS) and functional systems (fs) in a multiple sclerosis clinical trial. *Neurology*, 40(6):971–971.
- [Nyúl et al., 2000] Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150.
- [Omoumi et al., 2021] Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn, C. E., Louvet-de Verchère, F., Dos Santos, D. P., Kober, T., and Richiardi, J. (2021). To buy or not to buy—evaluating commercial AI solutions in radiology (the eclair guidelines). *European Radiology*, pages 1–11.
- [Orbes-Arteaga et al., 2019] Orbes-Arteaga, M., Varsavsky, T., Sudre, C. H., Eaton-Rosen, Z., Haddow, L. J., Sørensen, L., Nielsen, M., Pai, A., Ourselin, S., Modat, M., et al. (2019). Multi-domain adaptation in brain MRI through paired consistency and adversarial learning. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 54–62. Springer.
- [Ouali et al., 2020] Ouali, Y., Hudelot, C., and Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684.
- [Pérez-García et al., 2021] Pérez-García, F., Sparks, R., and Ourselin, S. (2021). Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, page 106236.
- [Perone and Cohen-Adad, 2018] Perone, C. S. and Cohen-Adad, J. (2018). Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 12–19. Springer.
- [Pfefferbaum et al., 2003] Pfefferbaum, A., Adalsteinsson, E., and Sullivan, E. V. (2003). Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 18(4):427–433.

-
- [Pham and Prince, 1999] Pham, D. L. and Prince, J. L. (1999). Adaptive fuzzy segmentation of magnetic resonance images. *IEEE transactions on medical imaging*, 18(9):737–752.
- [Pinto et al., 2020] Pinto, M. F., Oliveira, H., Batista, S., Cruz, L., Pinto, M., Correia, I., Martins, P., and Teixeira, C. (2020). Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific reports*, 10(1):1–13.
- [Planche et al., 2022] Planche, V., Manjon, J. V., Mansencal, B., Lanuza, E., Tourdias, T., Catheline, G., and Coupé, P. (2022). Structural progression of Alzheimer’s disease over decades: the MRI staging scheme. *Brain Communications*, 4(3):fcac109.
- [Pontillo et al., 2021] Pontillo, G., Tommasin, S., Cuocolo, R., Petracca, M., Petasas, N., Ugga, L., Carotenuto, A., Pozzilli, C., Iodice, R., Lanzillo, R., et al. (2021). A combined radiomics and machine learning approach to overcome the clinicoradiologic paradox in multiple sclerosis. *American Journal of Neuroradiology*, 42(11):1927–1933.
- [Pouyanfar et al., 2018] Pouyanfar, S., Tao, Y., Mohan, A., Tian, H., Kaseb, A. S., Gauen, K., Dailey, R., Aghajanzadeh, S., Lu, Y.-H., Chen, S.-C., et al. (2018). Dynamic sampling in convolutional neural networks for imbalanced data classification. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 112–117. IEEE.
- [Ran et al., 2019] Ran, M., Hu, J., Chen, Y., Chen, H., Sun, H., Zhou, J., and Zhang, Y. (2019). Denoising of 3D magnetic resonance images using a residual encoder–decoder wasserstein generative adversarial network. *Medical image analysis*, 55:165–180.
- [Ravnborg et al., 2005] Ravnborg, M., Blinkenberg, M., Sellebjerg, F., Ballegaard, M., Larsen, S. H., and Sørensen, P. S. (2005). Responsiveness of the multiple sclerosis impairment scale in comparison with the expanded disability status scale. *Multiple Sclerosis Journal*, 11(1):81–84.
- [Reinke et al., 2021] Reinke, A., Eisenmann, M., Tizabi, M. D., Sudre, C. H., Rädtsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M. J., Cheplygina, V., et al. (2021). Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*.
- [Roca et al., 2020] Roca, P., Attie, A., Colas, L., Tucholka, A., Rubini, P., Cackowski, S., Ding, J., Budzik, J.-F., Renard, F., Doyle, S., et al. (2020). Artificial

- intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagnostic and Interventional Imaging*, 101(12):795–802.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- [Roy et al., 2018] Roy, S., Butman, J. A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2018). Multiple sclerosis lesion segmentation from brain MRI via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172*.
- [Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- [Salem et al., 2018] Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., Rovira, À., and Lladó, X. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage: Clinical*, 17:607–615.
- [Scalfari et al., 2010] Scalfari, A., Neuhaus, A., Degenhardt, A., Rice, G. P., Muraro, P. A., Daumer, M., and Ebers, G. C. (2010). The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability. *Brain*, 133(7):1914–1929.
- [Schmidt et al., 2019] Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., Bellenberg, B., Zipp, F., Groppa, S., Sämann, P. G., et al. (2019). Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage: Clinical*, 23:101849.
- [Schmidt and Wink, 2017] Schmidt, P. and Wink, L. (2017). LST: A lesion segmentation tool for spm. *Manual/Documentation for version*, 2:15.
- [Sedai et al., 2019] Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W., and Garnavi, R. (2019). Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 282–290. Springer.
- [Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep

- networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.
- [Shaw et al., 2018] Shaw, R., Sudre, C., Ourselin, S., and Cardoso, M. J. (2018). MRI k-space motion artefact augmentation: model robustness and task-specific uncertainty. In *International Conference on Medical Imaging with Deep Learning–Full Paper Track*.
- [Shiee et al., 2014] Shiee, N., Bazin, P.-L., Cuzzocreo, J. L., Ye, C., Kishore, B., Carass, A., Calabresi, P. A., Reich, D. S., Prince, J. L., and Pham, D. L. (2014). Reconstruction of the human cerebral cortex robust to white matter lesions: method and validation. *Human brain mapping*, 35(7):3385–3401.
- [Shiee et al., 2010] Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535.
- [Shoeibi et al., 2021] Shoeibi, A., Khodatars, M., Jafari, M., Moridian, P., Rezaei, M., Alizadehsani, R., Khozeimeh, F., Gorriz, J. M., Heras, J., Panahiazar, M., et al. (2021). Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Computers in Biology and Medicine*, 136:104697.
- [Shorten and Khoshgoftaar, 2019] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.
- [Sled et al., 1998] Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A non-parametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging*, 17(1):87–97.
- [Small, 2014] Small, S. A. (2014). Isolating pathogenic mechanisms embedded within the hippocampal circuit through regional vulnerability. *Neuron*, 84(1):32–39.
- [Snell et al., 2017] Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087.
- [Sudre et al., 2017] Sudre, C. H., Cardoso, M. J., Ourselin, S., Initiative, A. D. N., et al. (2017). Longitudinal segmentation of age-related white matter hyperintensities. *Medical Image Analysis*, 38:50–64.

- [Sweeney et al., 2013] Sweeney, E. M., Shinohara, R. T., Shiee, N., Mateen, F. J., Chudgar, A. A., Cuzzocreo, J. L., Calabresi, P. A., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2013). OASISis automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI. *NeuroImage: clinical*, 2:402–413.
- [Thibault et al., 2013] Thibault, G., Angulo, J., and Meyer, F. (2013). Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Transactions on Biomedical Engineering*, 61(3):630–637.
- [Thiebaut de Schotten et al., 2020] Thiebaut de Schotten, M., Foulon, C., and Nachev, P. (2020). Brain disconnections link structural connectivity with function and behaviour. *Nature communications*, 11(1):1–8.
- [Thiebaut de Schotten et al., 2014] Thiebaut de Schotten, M., Tomaiuolo, F., Aiello, M., Merola, S., Silvetti, M., Lecce, F., Bartolomeo, P., and Doricchi, F. (2014). Damage to white matter pathways in subacute and chronic spatial neglect: a group study and 2 single-case studies with complete virtual “in vivo” tractography dissection. *Cerebral cortex*, 24(3):691–706.
- [Thompson et al., 2018] Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2):162–173.
- [Torgo et al., 2013] Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. (2013). SMOTE for regression. In *Portuguese conference on artificial intelligence*, pages 378–389. Springer.
- [Torrey and Shavlik, 2010] Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.
- [Tremblay et al., 2018] Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., and Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977.
- [Tripathi et al., 2019] Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., and Chari, V. (2019). Learning to generate synthetic data via compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- [Tustison et al., 2010] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320.
- [Valcarcel et al., 2018] Valcarcel, A. M., Linn, K. A., Vandekar, S. N., Satterthwaite, T. D., Muschelli, J., Calabresi, P. A., Pham, D. L., Martin, M. L., and Shinohara, R. T. (2018). MIMoSA: an automated method for intermodal segmentation analysis of multiple sclerosis brain lesions. *Journal of Neuroimaging*, 28(4):389–398.
- [Valverde et al., 2017] Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., Rovira, À., Oliver, A., and Lladó, X. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168.
- [Valverde et al., 2019] Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., Rovira, À., Salvi, J., Oliver, A., and Lladó, X. (2019). One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21:101638.
- [Vinitski et al., 1999] Vinitski, S., Gonzalez, C. F., Knobler, R., Andrews, D., Iwanaga, T., and Curtis, M. (1999). Fast tissue segmentation based on a 4D feature map in characterization of intracranial lesions. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 9(6):768–776.
- [Vukusic et al., 2020] Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D.-A., De Sèze, J., Cotton, F., Moreau, T., Stankoff, B., et al. (2020). Observatoire français de la sclérose en plaques (OFSEP): A unique multimodal nationwide MS registry in France. *Multiple Sclerosis Journal*, 26(1):118–122.
- [Wang et al., 2018] Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.
- [Weiss et al., 2016] Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.
- [Wolf et al., 2022] Wolf, T. N., Pölsterl, S., and Wachinger, C. (2022). DAFT: A universal module to interweave tabular data and 3D images in CNNs. *NeuroImage*, page 119505.

- [Wu et al., 2017] Wu, L., Zhu, Z., et al. (2017). Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*.
- [Wu and He, 2018] Wu, Y. and He, K. (2018). Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19.
- [Wylezinska et al., 2003] Wylezinska, M., Cifelli, A., Jezard, P., Palace, J., Alecci, M., and Matthews, P. (2003). Thalamic neurodegeneration in relapsing-remitting multiple sclerosis. *Neurology*, 60(12):1949–1954.
- [Xia et al., 2020] Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., and Roth, H. (2020). 3D semi-supervised learning with uncertainty-aware multi-view co-training. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3646–3655.
- [Yang and Gao, 2013] Yang, P. Y. and Gao, W. (2013). Multi-view discriminant transfer learning.
- [Yeh, 2022] Yeh, F.-C. (2022). Population-based tract-to-region connectome of the human brain and its hierarchical topology. *Nature communications*, 13(1):1–13.
- [Yip and Aerts, 2016] Yip, S. S. and Aerts, H. J. (2016). Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150.
- [Yonar et al., 2018] Yonar, D., Ocek, L., Tiftikcioglu, B. I., Zorlu, Y., and Severcan, F. (2018). Relapsing-remitting multiple sclerosis diagnosis from cerebrospinal fluids via Fourier transform infrared spectroscopy coupled with multivariate analysis. *Scientific reports*, 8(1):1–13.
- [Yushkevich et al., 2016] Yushkevich, P. A., Gao, Y., and Gerig, G. (2016). ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3342–3345. IEEE.
- [Zeng et al., 2020] Zeng, C., Gu, L., Liu, Z., and Zhao, S. (2020). Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Frontiers in Neuroinformatics*, 14:610967.
- [Zhang et al., 2019a] Zhang, H., Alberts, E., Pongratz, V., Mühlau, M., Zimmer, C., Wiestler, B., and Eichinger, P. (2019a). Predicting conversion from clinically isolated syndrome to multiple sclerosis—an imaging-based machine learning approach. *NeuroImage: Clinical*, 21:101593.

-
- [Zhang et al., 2021b] Zhang, H., Li, H., and Oguz, I. (2021b). Segmentation of new MS lesions with Tiramisu and 2.5 D stacked slices. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 61.
- [Zhang et al., 2019b] Zhang, H., Valcarcel, A. M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R. T., Hett, K., and Oguz, I. (2019b). Multiple sclerosis lesion segmentation with Tiramisu and 2.5 D stacked slices. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 338–346. Springer.
- [Zhang et al., 2020a] Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. (2020a). How does Mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*.
- [Zhang et al., 2020b] Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B. J., Roth, H., Myronenko, A., Xu, D., et al. (2020b). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*.
- [Zhao et al., 2017] Zhao, Y., Healy, B. C., Rotstein, D., Guttmann, C. R., Bakshi, R., Weiner, H. L., Brodley, C. E., and Chitnis, T. (2017). Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PloS one*, 12(4):e0174866.
- [Zhou et al., 2021] Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., and Liang, J. (2021). Models genesis. *Medical image analysis*, 67:101840.
- [Zijdenbos et al., 1994] Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., and Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE transactions on medical imaging*, 13(4):716–724.
- [Zijdenbos et al., 2002] Zijdenbos, A. P., Forghani, R., and Evans, A. C. (2002). Automatic " pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE transactions on medical imaging*, 21(10):1280–1291.
- [Zivadinov et al., 2008] Zivadinov, R., Reder, A., Filippi, M., Minagar, A., Stüve, O., Lassmann, H., Racke, M., Dwyer, M., Frohman, E., and Khan, O. (2008). Mechanisms of action of disease-modifying agents and brain volume changes in multiple sclerosis. *Neurology*, 71(2):136–144.