DEEP LEARNING METHODS FOR EXTRACTION OF NEUROIMAGE MARKERS IN THE PROGNOSIS OF BRAIN PATHOLOGIES

Albert Clèrigues Garcia

Universitat de Girona

DOCTORAL THESIS

Deep learning methods for extraction of neuroimage markers in the prognosis of brain pathologies

Albert Clèrigues Garcia

2022

Doctoral Thesis

Deep learning methods for extraction of neuroimage markers in the prognosis of brain pathologies

Albert Clèrigues Garcia

2022

Doctoral Program in Technology

Supervised by:
Dr. Xavier Lladó
Dr. Arnau Oliver
Dr. Sergi Valverde

Presented to obtain the degree of Doctor of Philosophy
at the University of Girona

“*Prediction is very difficult, especially about the future.*”

– Niels Bohr

Acknowledgments

This PhD thesis is the culmination of four years of intense and sustained research work. It has only been possible thanks to the previous work of institutions, staff and colleagues which have allowed for an environment in which students like me are able to thrive. In a true sense, this thesis is standing in the shoulders of giants. First of all, I would like to thank my supervisors Xavier Lladó, Arnau Oliver and Sergi Valverde for their trust and support in the pursuit of this PhD. The attention, time and dedication they provide to their students is the key of all their continued and increasing success. My sincere admiration goes as well to all the previous colleagues, teachers and administrative staff at the ViCOROB institute for their work in the VIBOT and MAIA masters, their effort and commitment have provided life-changing opportunities for many students. A special mention goes also to the constellation of doctors and researchers from collaborating hospitals, which allow the opportunity for our research to have an impact on the real world. Finally, I am most grateful to all the patients and families who, in such difficult times, grant their consent to share their imaging data with the research community.

Even before all this, I could not have reached this point without the unconditional support of my loved ones, who have painstakingly supported me through many years of school, five years of university, two years abroad in an international master and a whole PhD thesis. To them I am immensely grateful and have no other option but to recognize their fair share in my achievements and success. This thesis is especially dedicated to the loving memory of my father, Joan Salvador Clèrigues, who left us before I could finish this PhD, but never doubted for a second that I would make it.

Last but not least, this thesis is dedicated to all the friends I made along the way, and with whom I shared more successes than failures. It has been a real treat to meet the international colleagues of my VIBOT promotion which, along with all the great people I met from the accompanying MAIA and MSCV master programmes, have been the best companions in this incredible and life-changing experience. I will always remember all the great moments we shared in Le Creusot and, especially, in the Acacias dormitory.

Publications

The presented thesis is a compendium of the following research articles:

- **A. Clèrigues**, S. Valverde, J.Bernal, J. Freixenet, A. Oliver, X. Lladó. "Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks". Computers in Biology and Medicine. 115, art 103487, 2019. [JCR MCB IF 3.434, Q1(8/59)]

- **A. Clèrigues**, S. Valverde, J.Bernal, J. Freixenet, A. Oliver, X. Lladó. "Acute and sub-acute stroke lesion segmentation from multimodal MRI". Computer Methods and Programs in Biomedicine, art 105521, 2020. [JCR CSTM IF 3.424, Q1(15/104)]

- **A. Clèrigues**, S. Valverde, J. Salvi, A. Oliver, X. Lladó. "Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry". Computerized Medical Imaging and Graphics, art 102157, 2023. [JCR RNMMI IF 4.790, Q1(27/314)]

- **A. Clèrigues**, S. Valverde, A. Oliver, X. Lladó. "Improving segmentation-based brain atrophy quantification with unsupervised deep learning using tissue similarity regularization". Submitted to Medical Image Analysis, 2022. [JCR CSAI IF 8.880, Q1(5/133)]

The rest of publications, book chapters, challenges and conferences accomplished during this PhD thesis are the following;

Journals

- V. Abramova, **A. Clèrigues**, A. Quiles, D. García Figueredo, Y. Silva, S. Pedraza, A. Oliver, X. Lladó. "Hemorrhagic stroke lesion segmentation using a 3D U-Net with squeeze-and-excitation blocks". Computerized Medical Imaging and Graphics, vol 90, pp. 101908, 2021. [JCR RNMMI IF 4.790, Q1(27/314)]

- S. Valverde, L. Valencia, Ll. Coll, **A. Clèrigues**, A. Oliver, J. C. Vilanova, Ll. Ramió-Torrentà, À. Rovira, X. Lladó. "Assessing the accuracy and reproducibility of PARIETAL: a deep learning brain extraction algorithm". Journal of Magnetic Resonance Imaging, 2022. [JCR N IF 4.813, Q1(26/134)]

- L. València, **A. Clèrigues**, S. Valverde, A. Oliver, À. Rovira and X. Lladó "Evaluating the use of synthetic T1-w images in new T2 lesion detection in multiple sclerosis". Frontiers in Neuroscience, Brain Imaging Methods, 2022. [JCR N IF 5.152, Q2(87/274)]

- Y. Silva , M. Terceño, D. Puigoriol, M. Buxo, L. Pardo, M. Reina, A. Quiles, **A. Clèrigues**, V. Abramova, X. Lladó , C. Biarnés, S. Pedraza, J. Serena. "Prediction of intracerebral hemorrhage growth by CT-Perfusion: a prospective study". Submitted to Stroke, 2022. [JCR CN IF:7.914 Q1(16/192)]

Conferences and abstracts in medical journals

- **A. Clèrigues**, S. Valverde, J. Bernal, D. Pareto, J.C. Vilanova, Ll. Ramió-Torrentà, A. Rovira, A. Oliver, X. Lladó. "A quantitative analysis of deep learning methods for multiple sclerosis white matter lesion segmentation". Multiple Sclerosis, Berlin. October 2018. [JCR CN IF:5.649 Q1(23/199)]

- L. Valencia, S. Valverde, **A. Clèrigues**, A. Oliver, .F. Corral, A. Rovira, X. Lladó. "Assessing the usability of synthetic images to improve the detection of new T2 lesions". Multiple Sclerosis, Vienna, Austria. October, 2021. [JCR CN IF:6.312 Q1(28/208)]

- S. Valverde, R. Bramon, **A. Clèrigues**, L. Valencia, A. Oliver, N. Nerseyan, M. Puig, R. Robles, G. Álvarez, D. Lourido, L. Ramió-Torrentà, A. Rovira, X. Lladó. "Assessing the performance of an automated decision-support system for detecting active T2 lesions on non-standardized MRI systems and field strengths". Multiple Sclerosis. Amsterdam, 26-28 Oct 2022. [JCR CN IF:6.312 Q1(28/208)]

Book chapter

- J. Bernal, K. Kushibar, **A. Clèrigues**, A. Oliver, X. Lladó. “Deep Learning in Biology and Medicine”. First three authors contributed equally to the chapter titled “Deep learning in Medical Imaging”, Chapter 1. Eds: D. Bacciu, P. J. G. Lisboa, A. Vellido, World Scientific. 2022.

International workshops

- **A. Clérigues**, S. Valverde, J. Bernal, K. Kushibar, M. Cabezas, A. Oliver, and X. Lladó. "Ensemble of convolution neural networks for acute stroke anatomy differentiation". Ischemic Stroke Lesion Segmentation (ISLES) in Medical Imaging. MICCAI Workshop, 2018.

- M. Cabezas, S. Valverde, S. González-Villà, **A. Clèrigues**, M. Salem, K. Kushibar, J. Bernal, A. Oliver, J. Salvi and X. Lladó. "Survival prediction using ensemble tumor segmentation and transfer learning ". Multimodal Brain Tumor Segmentation Challenge 2018 (BRATS) in Medical Imaging. MICCAI Workshop, 2018.

- J. Bernal, M. Salem, K. Kushibar, **A. Clérigues**, S. Valverde, M. Cabezas, S. Gonzàlez-Villà, J. Salvi, A. Oliver, and X. Lladó. "MR Brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors". MR Brain tissue segmentation Challenge in Medical Imaging. MICCAI Workshop, 2018.

- M. Cabezas, S. Valverde, **A. Clèrigues**, M. Salem, K. Kushibar, J. Bernal, A. Oliver, J. Salvi and X. Lladó. "Brain tumour segmentation and prediction via CNNs". Multimodal Brain Tumor Segmentation Challenge 2019 (BRATS) in Medical Imaging. MICCAI Workshop, China, 2019.

- **A. Clèrigues**, V. Abramova, L. Valencia, Uma Lal-Trehan, M. Guirao, J. Salvi, A. Oliver, X. Lladó. "Stroke lesion segmentation from multimodal MRI using a residual U-Net and bilateral modality augmentation". Ischemic Stroke Lesion Segmentation (ISLES) in Medical Imaging. MICCAI Workshop, September 18-22, 2022.

Acronyms

MS	Multiple sclerosis
CIS	Clinically isolated syndrome
RRMS	Relapsing-remitting MS
PPMS	Primary progressive MS
SPMS	Secondary progressive MS
AD	Alzheimer's disease
CT	Computed tomography
MRI	Magnetic resonance imaging
T1w	T1-weighted
T2w	T2-weighted
PD	Proton density
FLAIR	Fluid-attenuated inversion-recovery
DWI	Diffusion weighted imaging
PWI	Perfusion weighted imaging
CBF	Cerebral blood flow
CBV	Cerebral blood volume
MTT	Mean transit time
Tmax	Time to maximum
CNN	Convolutional neural network
NCCT	Non-contrast CT
CTP	CT perfusion
CTA	CT angiography
WMH	White matter hyperintensity
FCM	Fuzzy C-means
RDF	Random decision forests
WM	White matter
GM	Gray matter
CSF	Cerebrospinal fluid
PVE	Partial volume effect
BSI	Boundary shift integral
GDL	Generalized Dice loss
DSC	Dice similarity coefficient
HD	Hausdorff distance
PPV	Positive predictive value
COD	Coefficient of determination
CRF	Conditional random field
ReLU	Rectified linear unit
PReLU	Parametric ReLU
NAWM	Normally appearing WM
GAN	Generative adversarial networks

BN	Batch normalization
ROI	Region of interest
HC	Healthy control
PBVC	Percentage of brain volume change
PGMVC	Percentage of gray matter volume change
PWMVC	Percentage of white matter volume change
BPF	Brain parenchymal fraction
GMF	Gray matter fraction
WMF	White matter fraction
ICV	Intracranial volume

Contents

List of Figures

Abstract

This PhD thesis focuses on improving the extraction of neuroimage markers for the prognosis and outcome prediction of neurological pathologies such as ischemic stroke, Alzheimer's disease (AD) and multiple sclerosis (MS). Our work has been developed on two of the most relevant neuroimage markers for diagnosis and prediction, brain lesion segmentation and longitudinal atrophy quantification. Brain lesion segmentation can be directly used in MS and ischemic stroke as a prognostic marker and can also be useful for other downstream segmentation tasks. In MS, disease activity produces very characteristic lesions which can help with diagnosis and prognosis of the pathology. In ischemic stroke, lesion segmentation can inform the treatment decision workflow by quantifying the amount of tissue that could be salvaged against the risks of surgical intervention. We also tackle in this PhD thesis the task of brain tissue segmentation for longitudinal atrophy quantification, a validated prognostic image marker in MS and AD. Measurements of longitudinal atrophy can be used to assess the rate of disease progression and might even help to predict AD onset years in advance. In MS patients, an accelerated rate of brain atrophy is also observed as a result of disease activity and is used as a prognostic marker and to evaluate the response of disease-modifying treatments.

The work in this thesis has been developed in several stages. In stage one, we approach the task of brain lesion segmentation and propose two patch-based deep learning methods for ischemic stroke, a 2D approach for computed tomography (CT) images and a 3D one for magnetic resonance imaging (MRI). Within both of these approaches, we have proposed training patch sampling techniques along with class balancing loss functions to mitigate the imbalance between healthy and lesion classes. We have also explored the use of several post-processing techniques to rectify the classification confidence of the model and filter lesions based on its morphology. Additionally, we have proposed a novel technique to exploit features based on the bilateral symmetry between brain hemispheres. The proposed approaches have shown state-of-the-art performance on two well-known publicly available datasets from the 2015 and 2018 editions of the Ischemic Stroke Lesion Segmentation (ISLES) challenge.

In the subsequent stages of this thesis, we focused on brain tissue segmentation for cross-sectional and longitudinal volumetric analysis. Although deep learning techniques have been at the forefront of many recent breakthroughs, current state-of-the-art methods for brain tissue segmentation have still not found a way to benefit from them. The main issue preventing their application is that the typically employed supervised deep learning methods would require accurate manual measurements of brain volumetry, which are virtually impossible to perform by human raters. Thus, we propose an unsupervised patch-based deep learning framework designed for accurate brain tissue segmentation which does

not rely on manual annotations for training. Instead, we learn from the outputs of a reference classical segmentation method and use data-driven techniques to improve upon their results and compensate its shortcomings. This unsupervised brain tissue segmentation framework is used as the basis for the work performed in the next stages.

Although the effect of WM lesions typically observed in MS patient images has been extensively studied in classical brain tissue segmentation methods, it has still not been evaluated within the more recent deep learning based approaches. In this regard, we begin by studying and evaluating the error that is introduced by WM lesions in our deep learning based tissue segmentation framework. Then, we propose an approach to reduce the error that these lesions introduce in the measured tissue volumes. Typically, the gold standard technique to mitigate WM lesion effect is to perform a lesion filling or inpainting in a previous separate step to prevent the abnormal intensities from interfering with the tissue segmentation. Instead, we propose a data-driven technique that performs the inpainting and segmentation tasks in an end-to-end fashion within our deep learning framework. By jointly optimizing both tasks, we are able to obtain an inpainting model that is also trained to aid in the segmentation task and minimizes the WM lesion influence to almost negligible levels.

Finally, based on our previously developed unsupervised brain tissue segmentation framework, we propose a method for longitudinal atrophy quantification. Within our approach, the network learns from a reference tissue segmentation method while utilizing data priors to regularize the training and avoid learning its errors and biases. More specifically, we propose a tissue similarity regularization during training which penalizes volume differences between pairs of scans from the same patient made within a short time interval. The experimental results show our method has greatly reduced short interval error and improved sensitivity to differences between healthy and AD patients compared to the reference method used for training.

In this PhD thesis, we have worked with diverse neuroimage markers and imaging modalities, which has provided valuable insights on the issues and challenges for their use in prognostic and predictive tasks.

Resum

Aquesta tesi doctoral se centra en millorar l'extracció de marcadors de neuroimatge per al pronòstic i predicció de l'estat del pacient en patologies neurològiques com l'ictus isquèmic, la malaltia d'Alzheimer o l'esclerosi múltiple (EM). El nostre treball ha estat desenvolupat en dos dels marcadors de neuroimatge més rellevants per al diagnòstic i la predicció; la segmentació de lesions cerebrals i la quantificació longitudinal d'atròfia. La segmentació de lesions pot ser utilitzada directament en ictus i en EM com a marcador del pronòstic i també pot ser útil en posteriors tasques de segmentació. A l'EM, l'activitat de la malaltia produeix lesions molt característiques que poden informar el diagnòstic i el pronòstic de la patologia. A l'ictus isquèmic, la segmentació de la lesió pot assistir en la decisió del tractament mitjançant la quantificació del teixit que podria salvar-se front als riscos de la intervenció quirúrgica. També abordem en aquesta tesi doctoral la segmentació del teixit cerebral per a la quantificació de l'atròfia longitudinal, un marcador d'imatge pronòstic validat a la EM i la malaltia d'Alzheimer. Les mesures de l'atròfia longitudinal es poden fer servir per avaluar la velocitat de progressió de la malaltia i fins i tot podrien ajudar a predir l'inici de la malaltia d'Alzheimer anys abans de mostrar els primers símptomes. En pacients amb EM, també s'observa una taxa accelerada d'atròfia cerebral com a resultat de l'activitat de la malaltia i la seva mesura pot servir com a marcador pronòstic i també per avaluar la resposta als tractaments modificadors de la malaltia.

El treball d'aquesta tesi s'ha desenvolupat en diverses fases. En la primera fase, abordem la segmentació de lesions cerebrals i proposem dos mètodes de deep learning basats en blocs per ictus isquèmic, un mètode 2D per a imatges de tomografia computaritzada i un 3D per a imatges de ressonància magnètica. En tots dos mètodes, hem proposat una tècnica de mostreig de blocs d'entrenament juntament amb funcions de pèrdua dissenyades per balancejar la contribució de la classe sana i lesionada i mitigar el seu desequilibri. També hem explorat l'ús de diverses tècniques de postprocessament utilitzant l'estimació de l'incertesa de classificació i el filtratge morfològic de les lesions. A més, hem proposat una tècnica per explotar característiques basades en la simetria bilateral entre els hemisferis cerebrals. Els mètodes proposats han demostrat estar a l'estat de l'art en termes de rendiment en dos coneguts conjunts de dades accessibles públicament en les edicions de 2015 i 2018 del repte internacional Ischemic Stroke Lesion Segmentation (ISLES).

Posteriorment, a les següents etapes de la tesi ens hem centrat en la segmentació de teixit cerebral per a l'anàlisi volumètric transversal i longitudinal. Tot i que les tècniques de deep learning han estat al capdavant dels avenços recents, els mètodes actuals per a la segmentació del teixit cerebral encara no han trobat la manera de beneficiar-se'n. El principal problema que n'impedeix

l'aplicació és que els mètodes típics de deep learning supervisats requereixen mesures manuals precises de la volumetria cerebral, que són virtualment impossibles de realitzar per humans. Per tant, en aquesta tesi proposem un sistema de deep learning no supervisat basat en blocs dissenyat per a una segmentació precisa del teixit cerebral que no requereix anotacions manuals per a l'entrenament. En el seu lloc, aprenem dels resultats d'un mètode de segmentació clàssic com a referència i utilitzem tècniques basades en coneixement a priori de les dades per millorar-ne els resultats i compensar les seves deficiències. Aquest sistema de segmentació de teixit cerebral no supervisat s'utilitza com a base per al treball dut a terme a les pròximes etapes.

Tot i que l'efecte de les lesions de matèria blanca (MB) típicament observades a les imatges de pacients amb EM s'ha estudiat àmpliament en els mètodes clàssics de segmentació del teixit cerebral, encara no s'ha avaluat dins dels mètodes més recents basats en deep learning. En aquest sentit, hem estudiat i avaluat l'error que introdueixen les lesions de MB al nostre sistema de segmentació de teixits basat en deep learning no supervisat. Posteriorment, proposem un mètode per reduir l'error que aquestes lesions introdueixen a les mesures de volums de teixit. En general, la tècnica estàndard per mitigar l'efecte de la lesió de MB és fer un repintat o inpainting de la lesió en un pas previ per evitar que aquestes intensitats anormals interfereixin amb la segmentació del teixit. En el seu lloc, nosaltres proposem una tècnica basada en dades que realitza les tasques de repintat i segmentació íntegrament dins del nostre sistema de deep learning. Optimitzant conjuntament les dues tasques obtenim un model de repintat optimitzat per a la tasca de segmentació i que minimitza la influència de la lesió a nivells gairebé insignificants.

Finalment, partint del sistema de segmentació de teixit cerebral no supervisat desenvolupat prèviament, proposem un mètode per a la quantificació de l'atròfia longitudinal. En el nostre mètode, la xarxa aprèn d'un mètode de segmentació de teixit com a referència mentre utilitza coneixement previ sobre les dades per regularitzar l'entrenament i evitar aprendre així dels seus errors i biaixos. Més específicament, proposem una regularització basada en la similitud de teixits durant l'entrenament que penalitza les diferències de volum entre parells d'imatges del mateix pacient fetes en un curt interval de temps. Els resultats experimentals mostren que el nostre mètode redueix en gran mesura l'error a curt termini i millora la sensibilitat a les diferències entre pacients sans i amb malaltia d'Alzheimer en comparació amb el mètode de referència utilitzat per a entrenar.

En resum, en aquesta tesi doctoral hem desenvolupat mètodes per a l'extracció de diversos marcadors neurològics procedents de diferents modalitats d'imatge, obtenint informació valuosa per al seu ús en tasques de pronòstic i predicció en malalties neurològiques.

Resumen

Esta tesis doctoral se centra en mejorar la extracción de marcadores de neuroimagen para el pronóstico y la predicción del estado del paciente en patologías neurológicas como el ictus isquémico, enfermedad de Alzheimer (EA) o la esclerosis múltiple (EM). Nuestra investigación se ha desarrollado en dos de los marcadores de neuroimagen más relevantes para diagnóstico y predicción; la segmentación de lesiones cerebrales y la cuantificación longitudinal de atrofia. La segmentación de lesiones puede ser utilizada directamente en ictus y en MS como un marcador del pronóstico y también puede ser útil en posteriores tareas de segmentación. En la EM, la actividad de la enfermedad produce lesiones muy características que pueden informar el diagnóstico y pronóstico de la patología. En el ictus isquémico, la segmentación de la lesión puede asistir en la decisión del tratamiento mediante la cuantificación del tejido que podría salvarse frente a los riesgos de la intervención quirúrgica. También abordamos en esta tesis doctoral la tarea de la segmentación del tejido cerebral para la cuantificación de atrofia longitudinal, un marcador de imagen pronóstico validado en la EM y la EA. Las mediciones de atrofia longitudinal se pueden usar para evaluar la velocidad de progresión de la EA e incluso podrían ayudar a predecir el inicio de la enfermedad años antes de mostrar los primeros síntomas. En pacientes con EM, también se observa una tasa acelerada de atrofia cerebral como resultado de la actividad de la enfermedad y su medición puede usarse como marcador pronóstico y para evaluar la respuesta a los tratamientos modificadores de la enfermedad.

El trabajo de esta tesis se ha desarrollado en varias fases. En la primera fase abordamos la segmentación de lesiones cerebrales y proponemos dos métodos de deep learning basados en parches para ictus isquémico, un método 2D para imágenes de tomografía computarizada y uno 3D para imágenes de resonancia magnética. En ambos métodos, hemos propuesto una técnica de muestreo de parches de entrenamiento junto con funciones de pérdida balanceadoras para mitigar el desequilibrio entre la clase sana y la lesionada. También hemos explorado el uso de varias técnicas de post-procesamiento usando la estimación de la incerteza de clasificación y el filtrado morfológico de lesiones. Además, hemos propuesto una técnica para explotar características basadas en la simetría bilateral entre los hemisferios cerebrales. Los métodos propuestos han demostrado un rendimiento al nivel del estado del arte en dos conocidos conjuntos de datos accesibles públicamente en las ediciones de 2015 y 2018 del reto internacional Ischemic Stroke Lesion Segmentation (ISLES).

En las siguientes etapas de esta tesis nos hemos centrado en la segmentación de tejido cerebral para el análisis volumétrico transversal y longitudinal. Aunque las técnicas de deep learning han estado a la vanguardia de los recientes avances, los métodos actuales para la segmentación del tejido cerebral aún no han

encontrado la manera de beneficiarse de ellos. El principal problema que impide su aplicación es que los típicos métodos de aprendizaje profundo supervisados requieren mediciones manuales precisas de la volumetría cerebral, que son virtualmente imposibles de realizar por humanos. Por lo tanto, en esta tesis proponemos un sistema de deep learning no supervisado basado en parches diseñado para una segmentación precisa del tejido cerebral que no requiere anotaciones manuales para el entrenamiento. En su lugar, aprendemos de los resultados de un método de segmentación clásico como referencia y utilizamos técnicas basadas en los datos para mejorar sus resultados y compensar sus deficiencias. Este sistema de segmentación de tejido cerebral no supervisado se utiliza como base para el trabajo realizado en las próximas etapas.

Aunque el efecto de las lesiones de materia blanca (MB) típicamente observadas en las imágenes de pacientes con EM se ha estudiado ampliamente en los métodos clásicos de segmentación del tejido cerebral, aún no se ha evaluado dentro de los métodos más recientes basados en deep learning. En este sentido, comenzamos estudiando y evaluando el error que introducen las lesiones de MB en nuestro sistema de segmentación de tejidos basado en deep learning no supervisado. Posteriormente, proponemos un método para reducir el error que estas lesiones introducen en la medición de los volúmenes de tejidos. Por lo general, la técnica estándar para mitigar el efecto de la lesión de MB es realizar un repintado o inpainting de la lesión en un paso previo para evitar que estas intensidades anormales interfieran con la segmentación del tejido. En su lugar, proponemos una técnica basada en datos que realiza las tareas de repintado y segmentación íntegramente dentro de nuestro sistema de deep learning. Al optimizar conjuntamente ambas tareas, obtenemos un modelo de repintado optimizado para la tarea de segmentación y que minimiza la influencia de la lesión a niveles casi insignificantes.

Por último, partiendo del sistema de segmentación de tejido cerebral no supervisado desarrollado previamente, proponemos un método para la cuantificación de la atrofia longitudinal. En nuestro método, la red aprende de un método de segmentación de tejido como referencia mientras utiliza conocimiento previo sobre los datos para regularizar el entrenamiento y evitar aprender también sus errores y sesgos. Más específicamente, proponemos una regularización basada en la similitud de tejidos durante el entrenamiento que penaliza las diferencias de volumen entre pares de imágenes del mismo paciente realizadas en un intervalo corto de tiempo. Los resultados experimentales muestran que nuestro método reduce en gran medida el error de corto plazo y mejora la sensibilidad a las diferencias entre pacientes sanos y con EA en comparación con el método que se ha usado como referencia para el entrenamiento.

En resumen, en esta tesis doctoral hemos desarrollado métodos para la extracción de diversos marcadores neurológicos en diferentes modalidades de imagen, lo que ha proporcionado información valiosa para su uso en tareas de pronóstico y predicción de enfermedades neurológicas.

Chapter 1

Introduction

1.1 Research Context

1.1.1 Ischemic stroke

Stroke is a medical condition by which an abnormal blood flow in the brain causes the death of cerebral tissue. Stroke is the third most common cause of morbidity worldwide, after myocardial infarction and cancer, and is the leading cause of acquired disability [1]. Depending on the type of abnormality, strokes can be divided in ischemic (80%) due to insufficient blood supply, and hemorrhagic (20%), typically due to a vessel rupture inside the brain generating toxic hematoma and swelling.

In an ischemic stroke, a blockage in an artery reduces the blood flow to its irrigated region, which starves the area of nutrients and puts it at risk of irreversible damage. The infarcted tissue during an episode is divided into three regions depending on the potential for its recovery, also referred as salvageablity: core, penumbra and benign oligemia. The core is formed by irreversibly damaged tissue, and is characterized by a fatally low blood supply. The penumbra represents tissue at risk but with enough blood supply that can be eventually salvaged depending on factors such as revascularization, collateral blood supply, tissue resistance, etc. The benign oligemia is an area whose vascularity has been altered by the stroke but is not at risk of permanent damage. In the affected area of the brain, the stroke lesion undergoes dynamic temporal evolution as depicted in Figure 1.1. Progressively, the lesion core grows as the affected tissue from the penumbra exhausts its nutrients and undergoes infarction. The stages of evolution are typically subdivided according to the time passed since stroke onset into *acute* during the first 24 hours, *sub-acute* from day one up to the second week, and *chronic* from the second week onward.

Once the symptoms of stroke have been identified, a shorter time to treatment is strongly correlated with a positive outcome [3]. Acute ischemic stroke therapies try to stop a stroke while it is happening by quickly dissolving the blood clot with medication or physical intervention. If the patient arrives at the hospital within the first 3 to 4.5 hours after the stroke, a tissue plasminogen activator is administered intravenously. It is a type of anticoagulant that works by dissolving the clot and improving flow to the part of the brain being deprived of blood. When promptly administered, it can save lives and reduce the long-term effects of stroke. Physical removal of a large blood clot, called an endovascular procedure or a mechanical thrombectomy, is another strongly recommended treatment

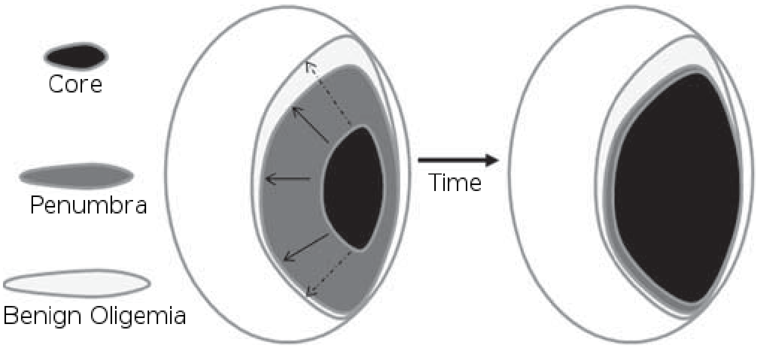


FIGURE 1.1: Temporal evolution of an ischemic stroke lesion [2].

option with successful reperfusion in up to 80% of receivers [4]. However, this surgery is not free of risks. A study found the overall complication rate was about 15.3% including symptomatic intracranial hemorrhage in 8.1% and 6.6% deaths [5]. In the acute setting, brain imaging plays a key role in assessing if the volume of penumbra, i.e. tissue that could be salvaged, is large enough to outweigh the risks of a surgical intervention.

1.1.2 Multiple Sclerosis

Multiple sclerosis (MS) is an immune-mediated inflammatory disease in the central nervous system affecting more than 2.8M people worldwide [6]. With a prevalence of 36 per 100.000 inhabitants, MS is the most common neurological cause of disability for young adults, with an average age of diagnosis ranging from 30 to 33 years. MS is a demyelinating disease, in which the insulating covers of nerve cells in the brain and spinal cord are damaged by the immune system. This damage interferes with the signal transmission between parts of the nervous system and leads to a range of physical, mental and psychiatric symptoms. MS is also characterized by the formation of focal lesions in the central nervous system and inflammation. Among the most common symptoms, MS patients can both temporarily and progressively experience partial paralysis, fatigue, cognitive dysfunction, muscle weakness and trouble with sensation or coordination.

Clinically isolated syndrome (CIS) is a first episode of neurologic symptoms, lasting for at least 24 hours, caused by inflammation and demyelination in the central nervous system. This kind of episodes are characteristic of MS but do not constitute a definitive criteria for diagnosis because people who experience a CIS may or may not go on to develop MS. When CIS is accompanied by the characteristic small lesions typical of MS, such as the ones depicted in Figure 1.2, the person has a high likelihood of a second episode of neurologic symptoms and definitive diagnosis. MS is typically subdivided in different types according to the appearance of new symptoms, which can either occur in isolated attacks (relapsing forms) or progressively build up over time (progressive forms). Relapsing-remitting MS (RRMS) is the most common disease course covering over 85% of the cases. It is characterized by clearly defined attacks of new relapses followed by periods of partial or complete recovery (remissions). Approximately 12% of MS patients are initially diagnosed with

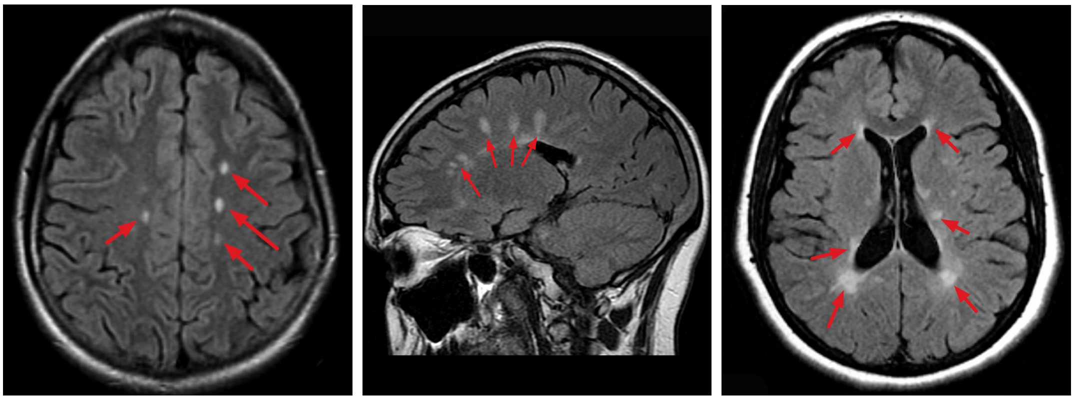


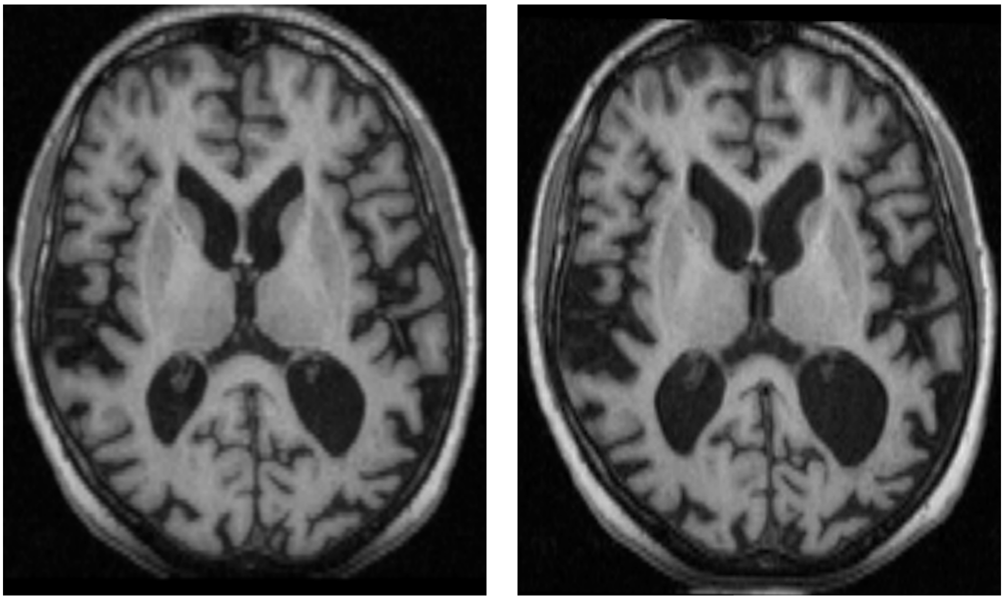
Figure 1.2: Examples of small hyperintense lesions characteristic of MS.

primary progressive MS (PPMS), which is characterized by worsening neurologic function (accumulation of disability) from the onset of symptoms, without early relapses or remissions. Some patients with RRMS diagnosis might eventually transition into a form of secondary progressive MS (SPMS). In this stage, there is a progressive worsening of neurologic function (accumulation of disability) over time, occasional relapses may occur as well as periods of stability.

In the clinical setting, brain imaging has become an essential tool for management of multiple sclerosis patients. After the initial diagnosis, disease monitoring is regularly performed to MS patients with brain imaging for assessment of new disease activity, such as the appearance of new lesions or increased brain atrophy [7]. Moreover, it is also used to assess the effectiveness of disease-modifying treatments both in routine clinical practice as well as in experimental trials [8].

1.1.3 Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative pathology characterized by the loss of neurons and synapses in the cerebral cortex and certain subcortical regions. As of 2020, there were approximately 50 million people worldwide with Alzheimer's disease [9]. It is most commonly diagnosed in people over 65 years of age but up to 10% of the cases are early-onset affecting those in their 30s to mid-60s. The main early symptom is difficulty in remembering recent events, with more symptoms appearing progressively as the disease advances such as language problems, disorientation, mood swings, behavioral issues, loss of bodily functions and ultimately death. The disease process if mainly associated to amyloid plaques, neurofibrillary tangles, loss of brain connectivity and generalized brain atrophy (see Figure 1.3). Although some environmental and genetic risk factors have been identified in correlation studies, the cause of Alzheimer's disease is poorly understood. Since initial symptoms are often mistaken for normal aging, a typical diagnosis requires cognitive testing, blood tests and brain imaging. A growing number of studies report that brain imaging can play an important role in early detection [10] and predicting the evolution of patients [11].



(A) Baseline (B) 2 year follow-up

Figure 1.3: Longitudinal scans of an Alzheimer's disease patient. The reduction in gray and white matter volume due to the generalized atrophy process is compensated by an increase in cerebrospinal fluid (CSF), producing an apparent enlargement of the ventricles and sulci in the cortex.

1.2 Research Background

Within the University of Girona, the Computer Vision and Robotics (ViCOROB) research group was established in 1996. Stemming from its computer vision expertise, the group began working in the field of medical image analysis, initially dealing with segmentation and registration in X-ray breast imaging. In 2009, the group started collaborating with several medical institutions and physicians in MS to develop new tools for brain image analysis with a focus on transference to the clinical practice. In the last years, the group has widened and generalized its brain imaging research framework which currently focuses on preprocessing, registration, segmentation of brain lesions, new lesions detection and longitudinal brain volumetry for atrophy quantification in different brain diseases.

The research performed in this PhD is framed within the following projects:

1. [2016 – 2019] wASSABI: "Automatic brain Structures Segmentation As potential imaging BIomarkers". Awarded in 2016 by the Ministerio de ciencia y tecnología. Ref: TIN2015-73563-JIN.

2. [2018 – 2021] EVOLUTION: "Predictive models for multiple sclerosis using brain magnetic resonance imaging biomarkers". Awarded in 2017 by Ministerio de ciencia y tecnología. RETOS 2017. Ref: DPI2017-86696-R.

3. [2021 – 2024] Modelling: "Models for Multiple Sclerosis using Deep Learning on Radiological, Clinical and Laboratory Data". Awarded in 2020 by Ministerio de ciencia y tecnología. RETOS 2020. Ref: PID2020-114769RB-I00.

Along with these projects, there has been a strong relationship with medical expert teams. Between 2021 and 2022, we collaborated with Dra. Yolanda Silva from Hospital Dr. Josep Trueta in a hemorrhagic stroke project titled "Medical image processing with artificial intelligence techniques for the segmentation and volume quantification of cerebral hematoma and edema in stroke patients". In the field of MS, collaborations were also carried out with Dr. Lluís Ramió and Dr. Joan Carles Vilanova from Hospital Dr. Josep Trueta and Dr. Àlex Rovira from Hospital Vall d'Hebron.

1.3 Objectives

As part of the wASSABI, EVOLUTION and Modelling research frameworks, the goal of the thesis is described as:

To develop novel deep learning based methods for improving the quality and robustness of cross-sectional and longitudinal neuroimage markers for the prognosis and outcome prediction of brain pathologies.

The sub-goals for this thesis revolve around the most relevant brain imaging markers for prognosis and outcome prediction of the considered pathologies. Dealing with several image based markers will allow to gain knowledge and insights about the challenges and particularities of each marker and how they can be improved and combined for its future use in predictive models. We propose and detail the following sub-goals:

- **to propose deep learning based approaches for improving brain lesion segmentation in ischemic stroke.** In recent years, the proliferation of medical imaging challenges has sparked an increase in the number of high-quality publicly accessible brain imaging datasets. At the same time, there is a surge of deep learning works proposing and evaluating a huge number of novel techniques in a wide variety of medical imaging related tasks. The field of brain lesion segmentation covers a huge array of neurological pathologies and imaging modalities, which has ultimately ended up in a highly fragmented literature with many preprocessing and data-driven techniques validated only for certain imaging modalities and pathologies. While many of these techniques and innovations are very specifically tuned to the particular physiopathology of the imaged disease, others are general enough to also benefit other unrelated tasks. As part of our goal, we will review the literature for recently proposed brain lesion segmentation methods to identify and apply techniques with the potential to improve our ischemic stroke lesion segmentation framework. We also aim at developing and validating novel techniques and approaches to improve results by leveraging the physiopathology of stroke and the particularities of typically employed imaging modalities.

- **to qualitatively and quantitatively evaluate the effect of WM lesions within deep learning based brain tissue segmentation**

methods and to develop novel data-driven techniques to reduce its influence. Brain lesions affecting the white matter of the brain can appear due to a wide array of causes and are a common occurrence in demyelinating inflammatory disorders, such as MS, and degenerative disorders, such as AD. The presence of WM lesions interferes with longitudinal brain atrophy quantification methods and distort brain volume measurements. Although the effect of WM lesions on tissue segmentation methods has been extensively documented in the literature, it has still not been evaluated within the more recent deep learning based approaches. Hence, our goal is to evaluate the effect of these lesions within deep learning based brain tissue segmentation methods and propose novel techniques to reduce their effect. Our hypothesis is that WM lesions will have a different effect on deep learning based methods and that their influence can be reduced more effectively by leveraging of data-driven techniques.

- **to develop novel unsupervised deep learning methods for brain atrophy quantification.** Although deep learning techniques have been at the forefront of many recent breakthroughs in medical image processing, current state-of-the-art techniques for brain atrophy quantification do not make use of these data-driven techniques and are yet to benefit from their advancements. The main issue is that most of the cutting edge techniques are based on supervised deep learning, where a ground truth annotation made by an expert physicians is used to train the model. The challenge for brain atrophy quantification methods is that sufficiently accurate manual ground truth annotations for effective training are virtually impossible to achieve by human raters. Thus, our goal is to develop novel unsupervised deep learning techniques that can improve upon the state-of-the-art methods for brain atrophy quantification without the use of manual annotations.

Throughout the completion of the considered objectives, we made most of the source code publicly available to the medical imaging community to improve the reproducibility of the research carried out in this thesis.

1.4 Document Structure

This thesis is done as the compendium of three Q1 JCR journal publications and one submission to a Q1 JCR journal, covering chapters 3 to 6. The articles in this compendium draw the clear thematic unity of presenting deep learning methods for extraction of brain image markers relevant to the prognosis and outcome prediction of neurological pathologies. Predictive models typically rely on the use of several image-based markers, along with a range of clinical and patient data, to perform prognostic and functional outcome prediction tasks. The set of articles in this compendium encompass two of the most clinically relevant imaging markers for prediction tasks, brain lesion segmentation and brain tissue segmentation for cross-sectional and longitudinal volumetry. The following parts of the thesis are structured as:

- **Chapter 2. Thesis Context.** This chapter presents the theoretical and technical background in which this thesis is framed. It has been divided in four main sections consisting on the basis for volumetric brain imaging, a brief introduction to deep learning and the research background for the tasks of stroke lesion segmentation and brain atrophy quantification.

- **Chapter 3. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks.** In this chapter, we present a 2D patch-based deep learning methodology for ischemic stroke lesion segmentation in non-contrast CT and CT perfusion imaging.

- **Chapter 4. Acute and sub-acute stroke lesion segmentation from multimodal MRI.** We present a 3D patch-based deep learning approach for acute and sub-acute ischemic stroke lesion segmentation from multimodal MRI.

- **Chapter 5. Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry.** We present a novel technique for training a deep learning based brain tissue segmentation method with built-in white matter lesion effect reduction onto the measured tissue volumes.

- **Chapter 6. Improving segmentation-based brain atrophy quantification with unsupervised deep learning using tissue similarity regularization.** In this chapter, we present a novel unsupervised deep learning pipeline for segmentation-based longitudinal atrophy quantification which uses a tissue similarity regularization to achieve state-of-the-art results.

- **Chapter 7. Main results and discussion.** We summarize and discuss the major results of the work realized within this thesis.

- **Chapter 8. Conclusions.** This chapter presents the main contributions and discusses future research that might derive from the work realized in this thesis.

Chapter 2

Thesis Context

2.1 Volumetric brain imaging

During the last few decades, advances in physics, electronics and computer science have sparked the proliferation and availability of in vivo non-invasive volumetric imaging of the human anatomy. In its digital form, the imaged space is discretized into a grid of small cubes and represented as a three dimensional structured array of numbers. In contrast to conventional imaging techniques, such as photography, that typically produce two-dimensional projections from a single viewpoint, volumetric imaging captures a one to one absolute magnitude representation of the imaged space without projection of any kind. Each of the volumetric picture elements, commonly refered to as a voxel, represents a cube centered on a specific point in space and of known absolute volume, characterized by the *voxel size* in each dimension which is measured in mm. Within this 3D grid, each voxel contains the average signal intensity, which depends on the imaging modality, within the volume it represents. Depending on the scanning principle, 3D images may be acquired as a series of stacked two-dimensional (2D) images each having a specific slice thickness and acquired at regular spatial intervals according to the slice spacing.

In medical imaging, the three dimensional image array is spatially referenced by accompanying metadata to standard human anatomical axes as well as to standardized world coordinates, expressed in mm, as illustrated in Figure 2.1. The relationship between voxel and world coordinates is typically encoded in an

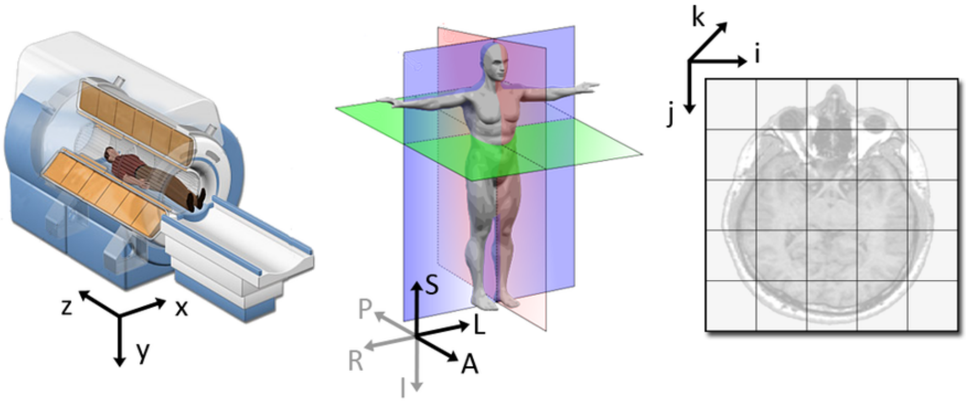
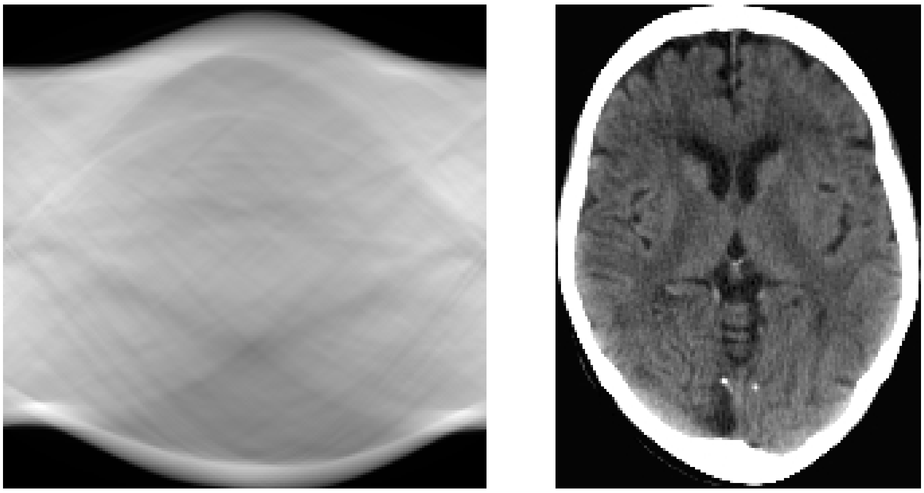


FIGURE 2.1: Relation between world (left), anatomical (center) and image (right) coordinate systems. (Source: `https://www.slicer.org/wiki/Coordinate_systems`)



(A) Captured sinogram

(B) Reconstructed slice

FIGURE 2.2: Sinogram and tomographical reconstruction of a CT brain scan slice.

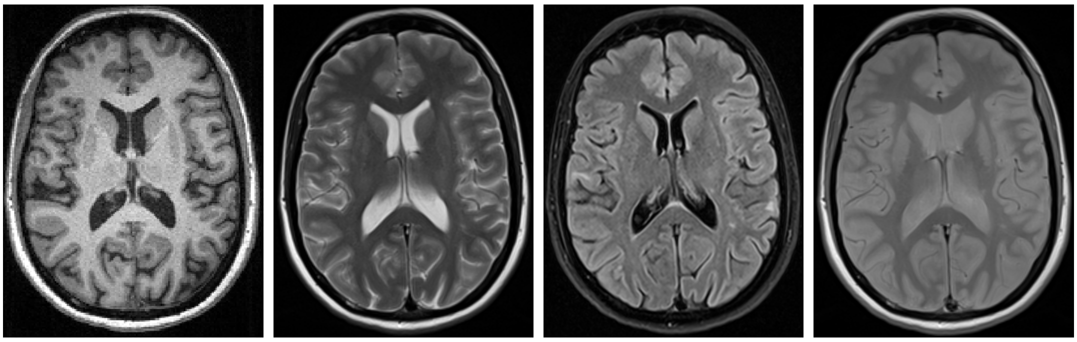
affine transformation matrix to a reference origin and orientation that depends on the convention adopted by the file format. Additional metadata is usually included to also encode the spatial properties of the voxel dimensions, such as voxel size or spacing. In brain imaging, typical voxel sizes are in the order of $1 \times 1 \times 1$mm, with higher resolution images in the order of $0.5 \times 0.5 \times 0.5$mm.

2.1.1 Computed Tomography

A computed tomography (CT) scan involves a motorized moving array of X-ray emitting diodes, and its corresponding detectors, which are used to take multiple 1D measurements from many different angles and positions and use computational methods to reconstruct the underlying 3D anatomy. The most common type of CT scan is refered to as spiral CT, or helical CT, in which an entire X-ray tube is spun around the central axis of the area being scanned. Due to the physical constraints of the acquisition, CT scans are typically acquired as a series of stacked 2D slices. For each 2D slice, multiple x-ray attenuation measurements are taken from many different angles around the central axis during each rotation of the scanner. All of the measurements taken are then merged onto a sinogram, which relates the axial displacement of the diodes and the radial angles at which the measurements where taken. The sinogram is then computationally processed using tomographic reconstruction algorithms which reconstruct cross-sectional tomographic images of the internal anatomy. An example of a sinogram and the final reconstructed image can be found in Figure 2.2.

2.1.2 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a non-invasive medical imaging technique that produces high quality representations of the internal anatomy and physiological process of the body. Its imaging principle is based on the magnetic resonance of the single proton that forms the hydrogen nuclei and which is



(A) T1-weighted (B) T2-weighted (C) FLAIR (D) PD

Figure 2.3: Common MRI sequences used for brain imaging.

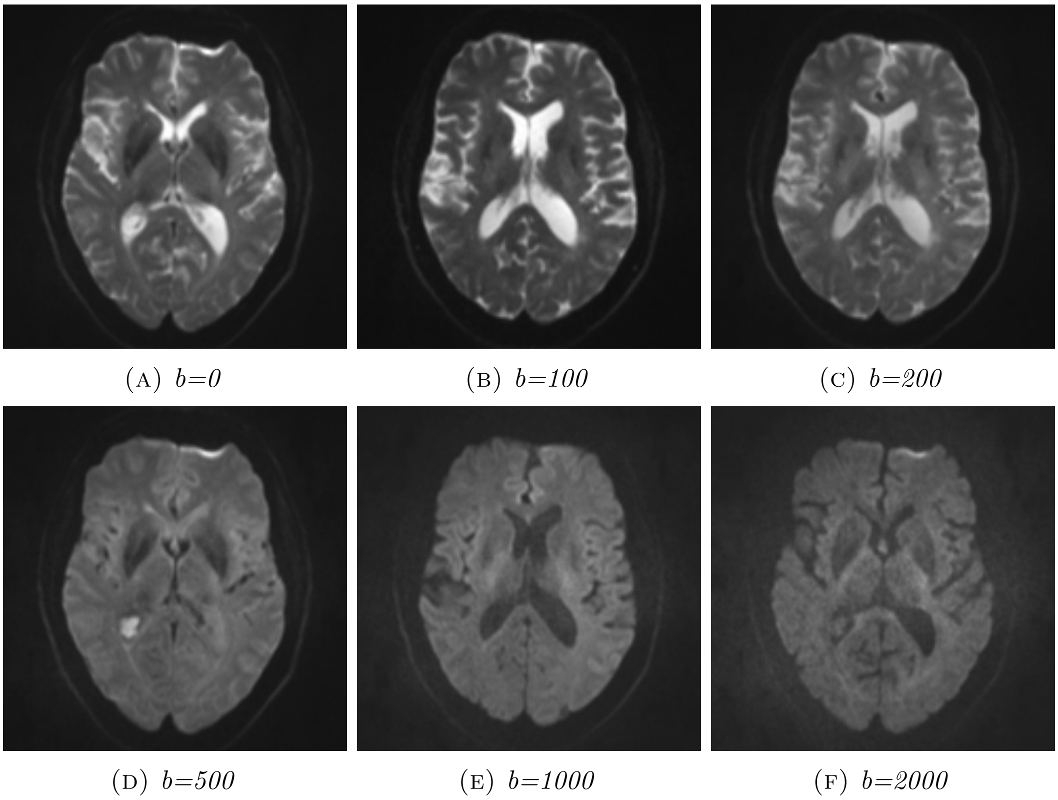
present in almost all living tissue. To perform an MRI scan, a strong oscillating magnetic field made with an emitter coil array is used to excite the hydrogen atoms and align their rotation axis to that of the applied field. This creates a phase coherence in the precession of all the proton spins. After the initial alignment, information about the underlying anatomy can be inferred from the way in which the hydrogen atoms return to its natural spin and which induces an electric current through the receiver coil. The key factor is that the behavior of protons is affected by fields from other atoms to which they are bonded, which makes it possible to separate the responses from hydrogen bonded within specific compounds in different tissues. Spatial encoding of the MRI signal is accomplished through the use of magnetic field gradients, which are compounded onto the base field, having an intensity offset that depends linearly on the spatial position. Brain MRI requires a very strong magnetic field, typically in the order of 1.5 to 3.0 Teslas, and which is uniform across the imaged volume within very small tolerances. In particular circumstances, extraneous contrast agents highly sensitive to magnetization may be given to highlight specific regions or processes.

After the initial magnetic resonance excitation, the measured magnetization signal of hydrogen nuclei is attenuated due to two simultaneous relaxation processes. The magnetization vector slowly relaxes towards its equilibrium orientation, parallel to the static magnetic field, at a tissue dependent time constant called the spin-lattice relaxation time (T1). Concurrently, the loss of coherence in the phases of spin precession attenuates the measured signal with a tissue-dependent time constant called the transverse relaxation time (T2). Since each tissue has particular responses of T1 and T2 times, specific highlighting of tissues or different contrasts between them can be obtained by tuning the excitation and measurement parameters. In T1 weighted (T1w) images, such as the one in Figure 2.3a, fat tissue quickly realigns its longitudinal magnetization and it therefore appears bright. In contrast, water has much slower longitudinal magnetization realignment and therefore has less transverse signal intensity, appearing dark. T1w images are typically used for assessing the cerebral cortex, identifying fatty tissue and observing the brain morphology in general. T2 weighted images (T2w) show different intensities and contrasts between tissues (see Figure 2.3b), and are useful in brain MRI for detecting edema, inflammation and revealing white matter lesions. In T2w images, external factors such as magnetic field inhomogeneity can alter the T2 relaxation time. This additional effect is initially captured on T2* weighted imaging, which is what the coil receiver actually detects during T2w imaging. To mitigate these extraneous influences, a refocusing pulse on spin-echo sequences is emitted to obtain the desired T2w image.

Another common MRI modality is Proton density (PD) weighting, depicted in Figure 2.3d, where intensity is proportional to the number of hydrogen protons in the area being imaged. PD weighted images are obtained when the contribution of both T1 and T2 contrast is minimized, and its particularly useful for imaging of structures adjacent to the ventricles, since it reduces the confounding signal coming from the cerebrospinal fluid. PD is often aquired along with T2w images since they can be acquired in a single sequence by using a dual spin-echo technique.

Fluid-attenuated inversion-recovery (FLAIR) is another MRI modality, which is a type of heavy T2-weighting modality that selectively nulls the signal coming from cerebrospinal fluid (see Figure 2.3c). In this sequence, a preliminary pulse is used to induce a magnetization in the inverse direction in which the upcoming T2w sequence will be performed. After the initial inverting pulse, tissues relax into their original magnetization at different longitudinal (T1) relaxation rates. The T2w sequence is delayed to the point when the longitudinal magnetization reaches the null point for the fluid, which we wish to suppress. This delay is called the time to inversion (TI) and it can be varied to suppress other kind of tissues depending on their T1 relaxation characteristics. Similar to PD, this kind of sequence improves the contrast of brain structures adjacent to cerebrospinal fluid.

Another form of MRI is that of diffusion weighted imaging (DWI), which is based on the measurement of the random Brownian motion of water within a voxel of tissue. As opposed to the free diffusion of pure water in a container, the diffusion of the water present in tissue is restricted by the cell membrane boundaries. To measure the diffusion characteristics in MRI, the attenuation of the T2* signal is measured proportional to the freedom of movement of water molecules. Thus, in DWI there is no need for the refocusing pulse used in T2w images. To perform DWI, an initial T2* weighted image is taken with no diffusion attenuation which will set the baseline b=*0* image. Afterwards, the diffusion characteristics of tissue are measured in various directions, a minimum of 3 orthogonal directions. Each of the measurements is done by first applying a strong gradient in one direction and then applying a symmetrically opposite one. Depending on the amplitude, duration and intervals of each pulse, a corresponding b value is associated as illustrated in Figure 2.4. At a physical level, water molecules will acquire different precession speeds depending on the magnitude of the gradient at their position. Molecules in the stronger areas of the gradient will acquire a fast precession speed and a slower one in the other end of the gradient. Since water molecules with restricted diffusion do not change location easily, they are exposed to the second gradient with a symmetrically opposite strength and phase. This creates a *rephasing* effect where the faster precessing protons will be slowed down and the slower ones will be sped up. When this happens, at the end of the pulse, all protons which stayed in the


(A) *b=0*

(B) *b=100*

(C) *b=200*

(D) *b=500*

(E) *b=1000*

(F) *b=2000*

FIGURE 2.4: DWI images with increasing *b* value. Case courtesy of A.Prof Frank Gaillard, Radiopaedia.org, rID: 33859. As the *b* value increases denser structures, such as the central ventricles, with more restricted diffusivity attenuate more.

same voxel will spin in unison and emit strong echo, appearing as bright areas in the image. On the other hand, freely moving water molecules will jump to neighboring voxels between the gradient pulses. Hence, the second gradient will not create this rephasing effect where the precession speeds of all molecules in the voxel are uniformly compensated. When the echo is measured from freely moving water, the different precession speeds do not generate a strong echo and appear as darker intensities. In a nutshell, the further an individual water molecule diffuses during the sequence, the more different the rephasing gradient strength will be and thus it will be dephased with respect to the rest of the molecules in the imaged voxel, reducing the amount of signal returned. In brain tissue, water within cerebrospinal fluid can diffuse very easily and, therefore, very little signal remains after rephasing, making the ventricles appear darker. In contrast, water in gray or white matter cannot diffuse as much since cell membranes restrict it, hence the protons spin in unison and produce brighter image intensities.

2.1.3 Perfusion Weighted Imaging

Perfusion Weighted Imaging (PWI) measures the arrival and perfusion characteristics of a contrast agent entering the brain through the vascular system. It can be performed with either magnetic resonance or computed tomography, essentially conveying the same information. The concept of perfusion is simple,

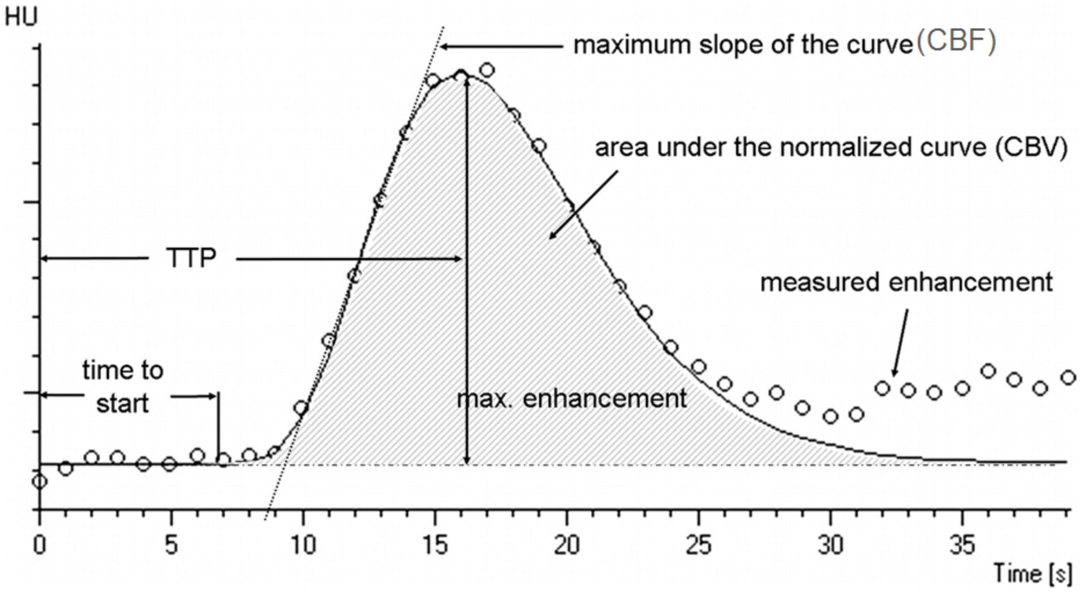


FIGURE 2.5: Typical perfusion curve of contrast agent in CT-PWI [2].

a bolus of contrast is injected and the target tissue is repeatedly scanned while the bolus enters and spreads through the vascular system. From the temporal information on the arrival, absorption and decay of the contrast agent in the tissue, information about its perfusion characteristics can be summarized in the form of derived perfusion parameters.

The typical contrast agent absorption curve as well as the interpretation of the different derived parameters is depicted in Fig. 2.5. Mainly, computed parameters include Cerebral Blood Flow (CBF), Cerebral Blood Volume (CBV), Mean Transit Time (MTT) and Time to maximum of deconvolved tissue residue function (Tmax):

- **CBF** is proportional to the ascending slope of the curve as the contrast agent reaches the area, the faster the flow, the more sudden the arrival of the contrast agent. Areas perfused by collateral vessels have slower and less robust flow than areas perfused by main arteries.

- **CBV** is the overall volume of contrast agent present in the area throughout the imaging process. For the sake of simplicity, CBV can be interpreted as a rough measure of whether or not blood actually reaches the area, even if at a slower pace.

- **MTT** corresponds to the average time, in seconds, that red blood cells spend within a determinate volume of capillary circulation. It represents the relation between volume and flow, i.e. an area can have a slower inflow but overall a greater total volume of blood.

- **Tmax** is the time to peak of the deconvolved tissue residue function. This mathematically derived parameter is not as intuitive to understand but has shown promising results in acute stroke imaging tasks.

Each of these parameters is computed for each voxel from the absorption curve as described. Finally, the extracted parameters from each voxel are

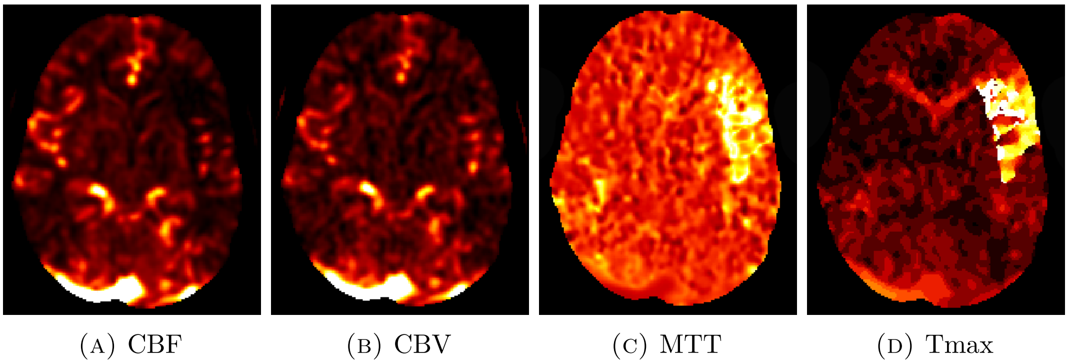

(A) CBF (B) CBV (C) MTT (D) Tmax

Figure 2.6: Derived parameter maps from the perfusion weighted imaging temporal series. An ischemic stroke lesion can be observed on the temporal lobe of the right hemisphere.

merged into several maps, one per each parameter, as shown in Figure 2.6. The modalities derived from perfusion weighted imaging usually show a lack of spatial homogeneity due to the single voxel curve parameter extraction.

2.2 Deep learning for dense volumetric segmentation

Deep learning is a type of machine learning technique in which the parameters of an arbitrarily huge highly-dimensional parametric function are trained to approximate a desired mapping between input and output. Deep learning relies heavily upon the concept of automatic differentiation to make this gradient-based optimization task computationally tractable. Automatic differentiation exploits the fact that every deep learning model, no matter how complicated, is composed of a sequence of elementary arithmetic operations and functions. By representing the deep learning model as a directed graph of operations, the chain rule can be repeatedly applied to compute the derivative at each step of the model with time-constant complexity. The derivative is always performed with respect to a differentiable *loss function*, which provides a measure of error between the current and desired output. In each training iteration, an input is forward-passed through the model and the loss term is computed between the predicted output and the known desired ground truth output. Then, a backward-pass is performed through automatic differentiation, commonly known as backpropagation, which computes the gradients for each parameter of the model with respect to the loss function. Finally, an optimizer will scale down the gradient and update the model parameters towards achieving the desired mapping. Progressively and through the use of many and varied training samples, the model parameters can be trained to approximate highly complex non-linear mappings.

In practice, deep learning models are built with modular differentiable layers implementing simple or composite operations. Although any kind of differentiable mathematical operation or function can be used within deep learning layers, the convolution operator has demonstrated excellent generalization performance and sparked a family of models called convolutional neural networks (CNNs). They

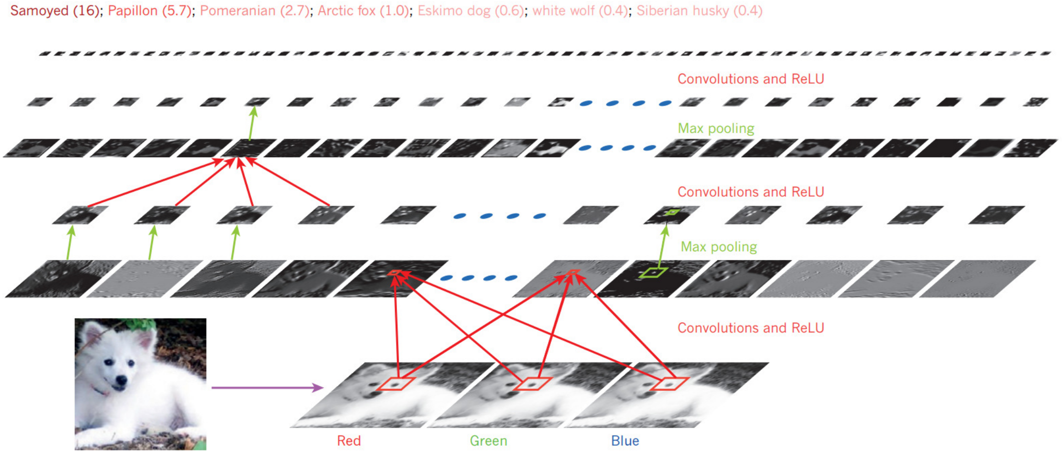

Samoyed (16); Papillon (5.7); Pomeranian (2.7); Arctic fox (1.0); Eskimo dog (0.6); white wolf (0.4); Siberian husky (0.4)
Convolutions and ReLU
Max pooling
Convolutions and ReLU
Max pooling
Convolutions and ReLU
Red
Green
Blue

FIGURE 2.7: Schematic diagram of a typical CNN architecture [12] detecting the species of the dog depicted in the bottom left, a Samoyed dog. Convolutional layers act on all the channels of the input with a small x-y neighborhood (red square), the resulting activations (red arrows) form the feature maps are in the row above. Max pooling (green arrow) halves the spatial dimensions and results in smaller feature maps for subsequent layers. The top most row represents low-dimensional images that could be employed as features for classification.

are a variation of multilayer perceptrons based on shared-weights architecture and translation invariance characteristics. Inspired by biological processes, the connectivity pattern resembles the organization of the animal visual cortex [12]. These models are implemented as a bank of filters or neurons, each characterized by its kernel weights. An activation map is generated by convolving the input image in a sliding-window fashion with the learned kernel weights. When correctly trained, a kernel encodes a particular feature and produces a high response, or activation, if the feature is present. The output of a convolutional layer consists on one activation map per kernel. In CNNs, convolutional layers are stacked in deep configurations along with other types of layers to extract and process increasingly complex features. The first convolutional layer of a CNN receives as input the image intensity values and detects low-level simple features from them, i.e. edges, dots, stripes, corners... The subsequent convolutional layers will receive these activation maps as inputs, and will extract higher-level features based on combinations of them, i.e. texture, lines, squares... For example, a kernel could detect a vertical line as several adjacent vertical edge activations. Then, a square could be detected as the combination of two horizontal lines on top of each other and two vertical lines side by side. This effectively builds a hierarchy of features where many different complex structures can be detected as combinations of a small number of simpler components, as illustrated in Fig. 2.7. Consequently, CNNs typically have a small number of kernels in the first layers and grow progressively with the aim of capturing the increasing number of feature combinations. CNNs were first introduced in 1989 [13] and only received mainstream attention after the excellent results on the ImageNet competition in 2012 [14]. The ImageNet dataset featured millions of images depicting objects from a thousand different classes, where CNNs nearly halved error rates of the previous best methods.

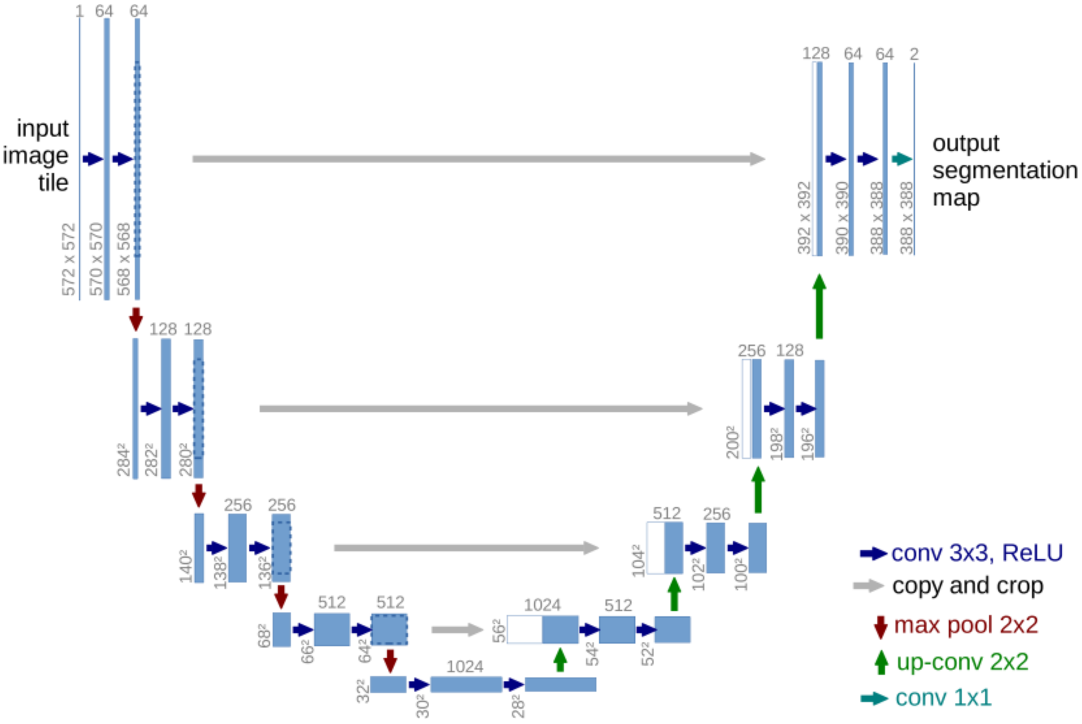


FIGURE 2.8: The 2D U-Net architecture proposed by Ronneberger et al. [16]. Each blue box corresponds to a multi-channel feature map. The number of channels is specified on top of the box. The original maps size is provided at the lower left edge of the box. White boxes represent the feature maps added via skip connections.

As 3D and 4D imaging are becoming routine in the clinical practice, and with physiological and functional imaging increasing, medical imaging data is increasing in size and complexity. Deep learning techniques are gaining popularity in many areas of medical image analysis [15]. CNN segmentation approaches for brain MRI are gaining interest due to their self-learning of features and generalization ability over large amounts of data. The neuroimaging field has greatly benefited from the advent of CNNs as the tasks of tissue and lesion segmentation involve non-linear complex relations. More recently, the U-Net architecture was first applied to medical imaging by Ronneberger et al. and achieved first place at the ISBI 2012 cell tracking challenge [16]. It has since then set the foundation for many subsequent approaches performing dense segmentation. The U-Net architecture, depicted in Fig. 2.8, consists of a contracting path (left side) and an expansive path (right side). The contracting path, also called encoder, follows the typical architecture of a CNN, which consists on a repeating sequence of convolutional layers and downsampling steps. At each downsampling level, the spatial size is halved and the number of feature channels is doubled. At the end of this path, a set of low-dimensional feature maps is obtained, also called the latent space, which is effectively a low-dimensional representation of the input which encodes meaningful global features. The expansive path, also called decoder, is a mirrored version of the encoder which essentially performs the inverse operation. Starting from the latent space, each level will upsample the feature maps to double its spatial dimensions and halve the number of features. Upsampling is usually performed through transpose convolutions which use trained interpolation kernels. In each level of the decoder, the native high-resolution feature maps from the corresponding encoder level are merged by concatenation to the upsampled low-resolution features. These are known as skip connections, and is what enables the decoder to incrementally recover the spatial information lost in the encoder. Two consecutive convolutions combine the features from the previous and current step, localizing spatially the encoded features. The resulting architecture has good convergence properties, allows dense segmentation of the input and provides good localization due to its use of multi resolution features.

The challenges for deep learning in medical imaging revolve around the low availability of data combined with the demand for high precision models within the clinical setting. One of the main challenges is the lack of ground truth segmentations, which are time consuming to perform and have to be done by expert physicians. Another issue which contributes to the lack of data is that of patient privacy and data protection, which increases the complexity and obstacles for data sharing between clinicians and researchers. Thus, when training a deep learning model with millions of parameters on not enough data, there is a risk of performing overfitting. In this scenario, the model can essentially memorize the little amount of training data within its parameters without distilling generalizable features and, hence, its performance will be poor on images outside of the training set.

Deep learning segmentation methods involving brain CT or MRI have more specific issues and challenges related to image acquisition and the nature of each segmentation task. One big issue is that of domain generalization, by which a model suffers from degraded performance on images having different characteristics than those of the training set, such as changes in image appearance or population characteristics. Differences in intensity or contrast can be caused by variations in the acquisition parameters or by the use of different scanner models or brands with diverse hardware configurations and image processing pipelines. Another important factor that can influence the domain generalization of a deep learning model is when the pathological characteristics or distribution of the training dataset are different than those of the evaluation dataset. For example, a model trained on stroke patients having mostly large lesions will not be likely to perform as well on cases with small ones. The ideal scenario for a deep learning method to obtain an unbiased and precise model is to have a training dataset with an equal and varied representation of pathological subtype, longitudinal evolution, lesion location, age, sex, race... In this sense, any representation imbalance of the training set will be captured as a bias in the deep learning model towards a better performance on the most represented cases.

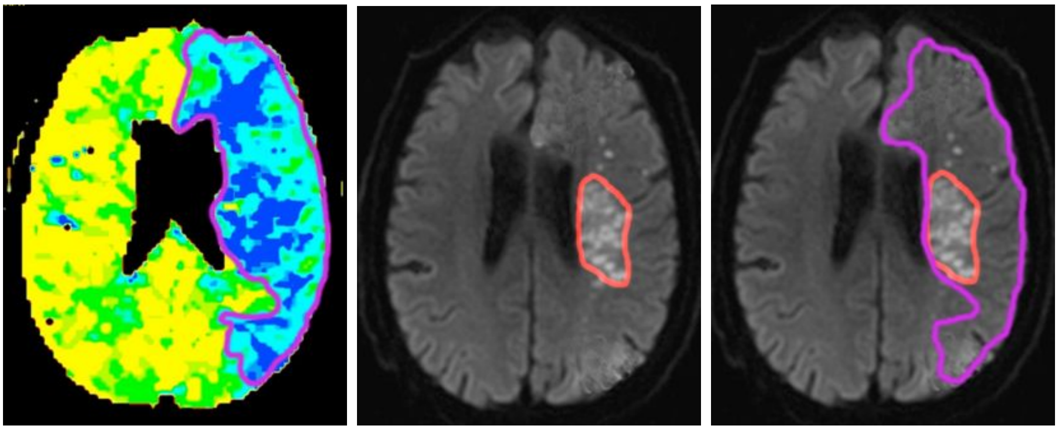
In the field of brain lesion segmentation specifically, one of the largest issues is that of class imbalance due to brain lesions being typically much smaller than the rest of healthy brain. A disproportionally imbalanced class representation would cause the network weights to be trained much more from the misclassification of healthy tissue than from lesions. In this way, the model is mostly specialized to correctly segment the healthy tissue and not enough on segmenting the specific particularities of brain lesions.

2.3 Ischemic stroke lesion segmentation

When a patient with suspected stroke is admitted in a hospital, the protocol for diagnosis and treatment decision will typically include a non-contrast CT (NCCT), CT perfusion (CTP) and CT angiography (CTA). NCCT is fast, inexpensive and readily available, however, it has limited sensitivity in the acute setting for direct response, and is mainly used to exclude other pathologies as well as to inform further course of action. Complementarily, CTP is used to provide more accurate diagnostic as well as helping in selecting patients for re-perfusion therapy. CTP allows to identify the combined extent of the penumbra, the region which is affected by the stroke but can be potentially salvaged, and the lesion core, composed of irreversibly damaged tissue. Finally, CTA is an imaging technique by which the exact location of the thrombus within an intracranial vessel can be obtained, and may guide intra-arterial thrombolysis or clot retrieval. In ischemic stroke, MRI is more time consuming and less available than CT but has significantly higher sensitivity and specificity in the diagnosis of acute ischemic infarction in the first few hours after onset. More specifically, DWI demonstrates increased signal of the infarct core within minutes of arterial occlusion. Generally, after 6 hours, high T2 signal will be detected, which continues to increase over the next day or two. Thus, MRI is usually reserved for the most challenging and complex cases where CT imaging is not enough to arrive at a diagnostic or treatment decision.

Segmentation tasks of clinical interest for ischemic stroke are similar in both CT and MRI based assessments. In CT imaging, the main tasks of interest revolve around accurately segmenting the stroke lesion core from NCCT images. In CTP derived maps, an abnormally appearing region which shows the combined extent of core and penumbra can be quantified or simply visually assessed. However, differentiating between core and penumbra is of critical importance to assess the precise extent of the tissue that can still be salvaged with treatment or surgery. For this, the core must be independently segmented from the NCCT image also acquired within the initial assessment and which shows only the lesion core. This is challenging due to the low sensitivity of the lesion core in NCCT images, where it is typically seen as a subtly hypodense region, and which is not easily assessed visually even by experienced radiologists.

Within MRI for ischemic stroke, the segmentation tasks of interest also revolve around core and penumbra differentiation. For this, a concept typically called PWI/DWI mismatch might hold the key for effective ischemic stroke assessment in MRI. The assumption is that DWI shows the irreversibly damaged tissue of the lesion core, the swelling dead cells, while PWI shows all tissue with abnormal vascularity, which includes both core and penumbra. The mismatch of abnormalities between PWI and DWI opened the door to the differentiation between salvageable and non-salvageable tissue (see Fig. 2.9). Several studies [17, 18, 19, 20] explored the prognostic potential of the DWI–PWI mismatch, although questions surrounding its validation remain. The lack of standard acquisition parameters, different post-processing algorithms and the broad definition of the regions to segment make difficult to draw conclusions on its potential for widespread adoption.



(A) PWI (B) DWI (C) PWI/DWI mismatch

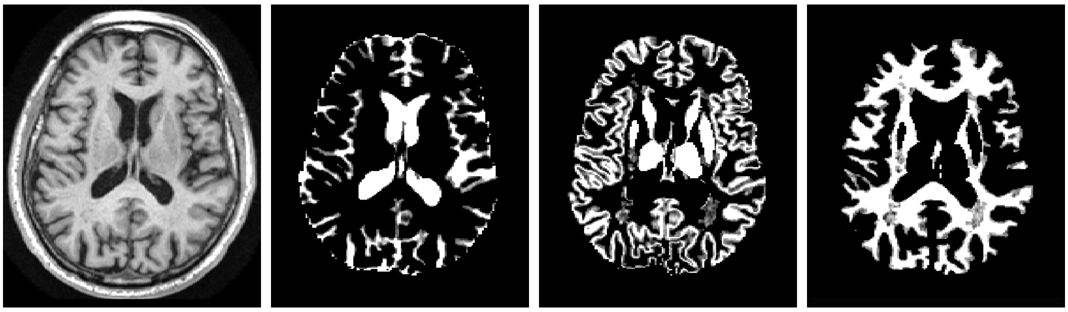
FIGURE 2.9: Example of the PWI/DWI mismatch. The penumbra (outlined in purple), tissue with abnormal vascularity that can still be salvaged, can be clearly seen in PWI images. The lesion core (outlined in red), irreversibly damaged tissue, appears clearly in DWI images due to its with restricted diffusivity.

Imaging tasks involving stroke lesions are complex due to its relation with the vascular system. First, the appearance of stroke lesions varies significantly over time, not only between, but even within the clinical phases of stroke development. Second, stroke lesions can appear at any location in the brain and take on any shape. They may or may not be aligned with the vascular supply territories and multiple lesions can appear at the same time, e.g. caused by an embolic shower. Some lesions may have a radius of few millimeters while others encompass almost a complete hemisphere. Third, lesion structures may not appear as homogeneous regions; instead, their intensity can vary significantly within the lesion territory. In addition, automatic stroke lesion segmentation in the acute setting is complicated by the possible presence of other stroke-similar pathologies, such as chronic lesions or white matter hyperintensities (WMHs). Because of the many nuances and issues with stroke imaging, less quality public datasets exist for stroke than for other neuroimaging tasks. However, the situation has evolved favorably for stroke imaging in the last few years with the introduction of the Ischemic Stroke Lesion Segmentation (ISLES) challenge, which constituted one of the first sources of high quality labelled public stroke datasets. The ISLES 2015 challenge [21] was introduced as part of the Medical Image Computing & Computer Assisted Intervention (MICCAI) international conference to demonstrate the potential for automatic segmentation and prediction methods in stroke research. The ISLES 2015 challenge focused on MRI and included the sub-acute ischemic stroke lesion segmentation (SISS) and the acute stroke outcome/penumbra estimation (SPES) subtasks. The following two editions of the ISLES challenge in 2016 and 2017 focused on prediction of chronic lesion outcome and prediction of the degree of disability 90 days after a stroke incidence (clinical outcome) from sub-acute MRI. The ISLES challenge in 2018 changed course and instead focused on stroke core lesion segmentation from acute NCCT and CT Perfusion scans with manually outlined core lesions on MRI DWI scans acquired soon thereafter. By combining CT and MRI imaging, the region of DWI abnormality can act as a gold standard for an irreversible brain infarction segmentation task using only CT imaging as input. More recently, the ISLES 2022 challenge proposed the task of multimodal MRI stroke core lesion segmentation from DWI images. In this edition, there was a wider ischemic stroke disease spectra, involving variable lesion size and burden, more complex infarct patterns and variable anatomically located lesions in data from multiple centers.

Before the popularization of deep learning approaches, most of the state-of-the-art methods were based upon more classical machine learning classifiers such as fuzzy C-means (FCM) clustering [22] or Random Decision Forests (RDF) [23, 24, 25]. RDF classifiers have excellent generalization properties, which has made them popular for difficult tasks with few training samples such as stroke lesion segmentation. However, they are essentially a cascade of simple classifiers acting on hand crafted features, and see their potential severely limited by the quality of the given features, which may vary for different tasks. Moreover, RDFs are not capable of explicitly using or combining the given features to find better or composite correlations between them. Deep learning does not have these limits by design but is still restricted by the architectural design, the amount of data and the training procedure. Deep learning techniques took hold of state-of-the-art methods for stroke lesion segmentation quickly after their mainstream application to medical imaging sparked by the U-Net architecture of Ronneberger, Fischer, and Brox [16] proposed in 2015. Among the ISLES 2015 challenge participants, only a single deep learning approach [26] was among the top three submissions, while in the ISLES 2016 and 2017, all top three submissions were based on deep learning approaches using the U-Net architecture.

2.4 Brain atrophy quantification

Brain atrophy assessed on structural MRI has been demonstrated as a valid marker on post-mortem histology for the neurodegeneration seen in AD patients [27, 28]. Other studies dealing with MS patients have provided evidence of brain atrophy as a marker for clinical outcomes and treatment response [29, 30, 31, 32]. Currently, MRI derived measures of atrophy provide a non-invasive way to quantify the longitudinal evolution of patients. Atrophy measures rely on the precise quantification of the volume of relevant structures or tissue that can be seen with good contrast across several images taken years apart. In general, the tissue of the brain parenchyma is composed of two major tissue types: gray matter (GM) and white matter (WM). Gray matter is made of mostly unmyelinated neurons in charge of nerve connections and general processing. In contrast, white matter is made of mostly myelinated neurons which transmit nerve signals much faster and act as an information highway between distant parts of the brain and body. The remaining space in the intracranial cavity is filled with cerebrospinal fluid (CSF), a clear fluid in which the brain parenchyma and spinal cord are suspended. CSF exerts pressure on the outside of the brain and spinal cord and acts as a stabilizer and shock absorber. In addition, small CSF-filled cavities inside of the brain called ventricles are used to compensate



(A) T1w (B) CSF (C) GM (D) WM

FIGURE 2.10: Example of a brain tissue segmentation.

the outer CSF pressure and maintain the shape of the soft brain tissue. Within the brain parenchyma, the tissue can be further differentiated into specialized distinct structures, such as the sub-cortical deep gray matter structures. These are located beneath the cerebral cortex and involved in complex activities such as memory, emotion, pleasure and hormone production. It has been shown that different brain pathologies are associated with characteristic longitudinal changes on the morphology and volume of specific brain tissues or individual structures. To obtain precise measurements, a larger difference in brightness between the considered tissues or structures and a high signal to noise ratio is desired. In general, T1-weighted MRI is typically used for structural analysis since it offers a good contrast between the WM, GM and CSF components, as seen in Figure 2.10.

Although many technological improvements on MRI scanners and processing software have been made in recent years, atrophy quantification methods are still affected by different confounding factors related to image acquisition and processing as well as to different physiological and pathological phenomenons [33]. These issues include, but are not limited to:

Partial volume effect (PVE). In brain atrophy quantification from MRI, an assumption is made that each tissue within the parenchyma has a particular intensity distribution that can be characterized for segmentation. However, due to the insufficient resolution of the images, several types of tissue can be captured in the intensity value of a single voxel, as shown in Figure 2.11. In such cases, the voxel intensity depends not only on the imaging sequence and tissue properties, but also on the proportions of each tissue type present in the voxel, which especially affects the segmentation of interfaces between tissues or structure boundaries. Thus, very precise and robust estimation of the intensity distribution of each tissue is needed for correctly estimating the tissue mixture in each voxel. In this sense, measures of volume from MRI should be done from the estimated partial volume probabilities and never from categorical segmentations that assign a single class to each voxel. The main challenge for atrophy quantification methods is to correctly estimate the partial volume mixture in the presence of noise or other image artifacts that distort the voxel intensities.

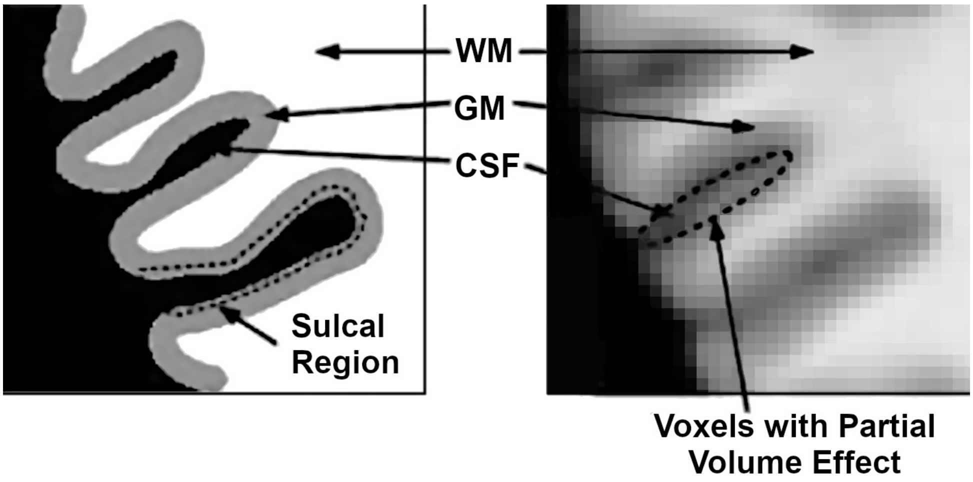


FIGURE 2.11: The partial volume effect.

Noise. MRI noise follows a Rician distribution [34] which, unlike additive Gaussian noise, is signal dependent and makes its removal a difficult task. MRI noise directly affects the measurement of volumes since it complicates the estimation of partial volumes, which heavily rely on the intensity of each voxel to estimate its tissue mixture. Although an extensive amount of methods and approaches have been proposed for MRI denoising, they are not used in any of the state-of-the-art brain atrophy quantification techniques.

Bias field. Because of inhomogeneities in the scanner magnetic field, a low frequency smooth multiplicative signal is observed that makes the same tissue have varied intensity distributions across different regions of the image. Although this effect has been reduced in modern scanners, it is still a source of error and specific preprocessing steps are taken in state-of-the-art tissue segmentation methods to remove it.

Movement artifacts. Due to the way MRI is acquired in the frequency domain, a small movement of the subject during acquisition can introduce intensity artifacts across the entire image in the spatial domain. Similarly to noise, these artifacts shift the intensity of voxels and interfere with the precise estimation of partial volume mixtures. Although modern scanners are equipped with movement correction strategies, these only work for small displacements and some image artifacts can still be introduced in the acquired images.

Gradient distortion. Changes in position of the head relative to the magnet isocenter of the scanner can introduce spatial distortions in the image of up to 5mm [35] (see Figure 2.12). The morphological distortions can persist even after gradient correction depending on the calibration accuracy of the scanner. Thus, in longitudinal studies, special care should be taken in repositioning the patient relative to the magnet of the scanner.

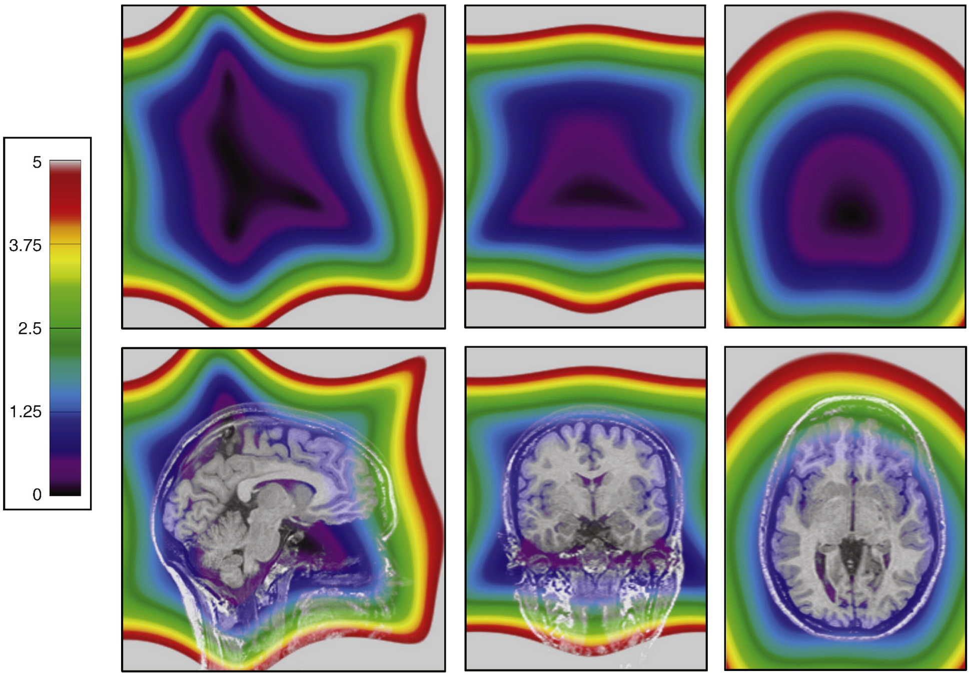
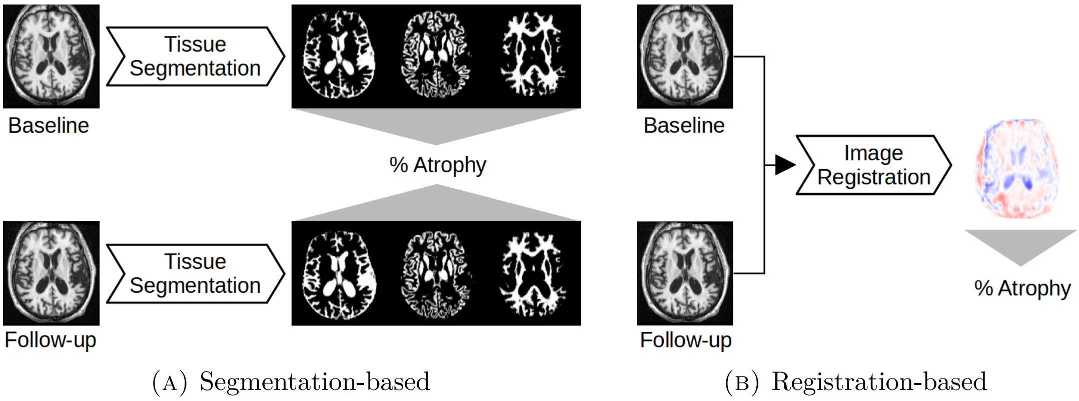


FIGURE 2.12: Example of an MRI gradient distortion field from the work of Caramanos et al. [35]. The color scale represents the distance (in millimeters) that a voxel moves because of gradient distortion between its "real" and its "apparent" location on such an MRI scan that has not been corrected for the gradient distortion.

Changes in scanner equipment. In clinical practice, several scanner models of different vendors might coexist within a single site. In chronic neurodegenerative pathologies, it is challenging to ensure the same patient is scanned with the same equipment and acquisition parameters across long time scales. Even when utilizing the same equipment and acquisition protocols, system upgrades, such as sequence innovations or hardware improvements, are inevitable and will have an effect on the image contrast. This change directly affects the intensities of tissues and interferes with the consistency of partial volume estimation in longitudinal image processing.

Pathological effects. The appearance of brain lesions within the brain parenchyma is a common co-occurrence in pathologies where brain atrophy is also an interesting image marker. For instance, the WM lesions typically observed in MS patients appear in MRI as abnormally appearing hyperintense or hypointense regions. The presence of these lesions can interfere with the characterization of the intensity distribution of normally appearing tissue and shift the partial volume mixture estimation towards over or underestimating a certain tissue type. Another pathology induced effect is that of pseudoatrophy, in which treatments using disease-modifying drugs and steroids have been associated with a counterintuitive acceleration of brain volume loss caused by their anti-inflammatory effect, that reduces apparent tissue volume. This phenomenon complicates the interpretation of observed changes in brain volume and atrophy rates when evaluating the effect of a given treatment.



(A) Segmentation-based (B) Registration-based

FIGURE 2.13: Main types of brain atrophy quantification methods.

Physiological effects. A series of confounding factors might be introduced simply by normal physiological changes between acquisitions. For instance, a shift in measured brain volume has been observed depending on the time of day at which a patient was scanned [36]. Other studies have found that the hydration status induces significant ventricular volume changes [37].

Evaluation of brain atrophy quantification methods is not an easy task, mainly due to the nonexistence of sufficiently accurate manual ground truths for direct evaluation of their performance [38]. Typically, manual methods of volumetry made by physicians are confined to line measurements such as bicaudate ratio, brain width, the midbrain to pons ratio, the lateral width, and the third ventricle width. These simple measurements are already a time-consuming task for highly experienced physicians and provide limited information. Thus, automated computational methods for brain atrophy quantification are limited to indirect evaluation metrics, such as short interval imaging error or correlation to known clinical differences between populations. A measure of quantification error can be obtained by evaluating the atrophy measured between scan-rescan images, two consecutive MRI acquisitions using the same protocol and without repositioning of the patient, under the assumption that an ideal atrophy quantification method should measure zero change between them. However, this metric does not really consider the changes due to actual atrophy or introduced by repositioning of the patient between longitudinal scans. For this, the sensitivity of an atrophy quantification method to longitudinal changes can be assessed by quantifying the differences between two subject populations known to have different rates of brain atrophy. The assumption being that better quantification methods would see larger and more pronounced differences between the longitudinal change measures of these two populations.

In general, longitudinal brain atrophy quantification methods can be classified into either segmentation-based or registration-based techniques, as exemplified in Figure 2.13. In segmentation-based methods, a target structure or tissue is cross-sectionally segmented in each of the longitudinal scans and atrophy is quantified from differences in the measured volumes. In contrast, registration-based techniques derive measures of atrophy from the observed spatial deformation of structures or tissue between two longitudinal scans. Segmentation-based methods are typically regarded as less accurate and more variable than its registration-based counterparts [33] and their use is discouraged for longitudinal studies. Although several segmentation-based methods for cross-sectional brain volumetry from T1-w MRI have been proposed in recent literature, only SIENA-XL [39] has been purposefully built for longitudinal imaging. Registration-based methods are usually preferred for longitudinal change analysis since they typically have lower quantification error and better sensitivity to atrophy changes [33]. SIENA [40] is a well-known and widely used registration-based atrophy quantification method based on the boundary shift integral (BSI) [41]. Within SIENA, atrophy is measured between two linearly registered scans from the surface displacement of the interface between GM and WM, which is obtained from tissue segmentations of each scan made with FAST [42] . More recently, measures of atrophy based on Jacobian integration have shown further improvements, such as larger effect sizes and lower quantification error [43]. These methods measure volume changes through integrating the determinant of the Jacobian of a non-linear transformation between two longitudinal scans. The region for integration is typically obtained from a cross-sectional segmentation of a structure or tissue in one of the scans. It is worth noting that, even within registration-based methods, some form of cross-sectional segmentation of tissue or structures is still required.

2.4.1 Effect of WM lesions

As previously discussed, the WM lesions typically seen in MS patients interfere with brain tissue segmentation methods and bias the measured volumes. WM lesions appear as an additional intensity distribution that intersects with those of normally appearing tissue and shift their estimated intensity profile. The presence of these abnormal intensities alters the classification not only of neighboring but also of distant brain tissue. The error introduced by WM lesions can vary depending on its volume and intensity profile [44]. In MS, new lesions can appear as a result of disease activity and existing ones may grow or change over time, which means that different levels of error can be obtained for images taken at different timepoints.

Typically, the effect of WM lesions in brain tissue segmentation is reduced by performing lesion inpainting as a preliminary step before segmentation. These techniques fill the voxels within a WM lesion mask, made either manually or automatically, with intensities resembling the normally appearing WM (NAWM) of that image, which is characterized in a tissue segmentation made prior to the inpainting. More recently, methods using deep learning have been proposed based on the use of Convolutional Neural Networks (CNNs) of Generative Adversarial Networks (GANs) for lesion inpainting. These data-driven methods achieve a significantly more realistic and accurate inpainting of the abnormal intensities than previous state-of-the-art techniques.

2.5 Datasets

In this section, we introduce the datasets used in this PhD thesis and provide a brief overview of their aims and image characteristics. Since we have dealt with several imaging markers and pathologies, the datasets are divided into sections according to their intended aim or task.

2.5.1 Stroke lesion segmentation datasets

The Ischemic Stroke Lesion Segmentation (ISLES) challenge was started in 2015 with the aim of providing a platform for fair and direct comparison of automated segmentation methods as well as to increase the number of publicly accessible datasets to promote and facilitate scientific progress in this area. Automated methods for ischemic stroke lesion segmentation are of great clinical interest since they can be used to more objectively inform the diagnosis and treatment decision workflows as well as to help in the improvement or validation of predictive and prognostic markers.

ISLES 2015 challenge

The 2015 edition of the ISLES challenge [21] focused on acute and sub-acute ischemic stroke lesion segmentation tasks using multi-spectral MRI data. The challenge proposed two sub-tasks dealing with different lesion parts, developmental stages and MRI modalities.

SISS sub-task. The sub-acute ischemic stroke segmentation (SISS) sub-task covered the segmentation of ischemic stroke lesions from multi-spectral MRI sequences acquired in the sub-acute stroke development stage. The provided dataset included 28 training and 36 testing cases acquired in the first week after onset. For the training images, the provided gold standard, the whole lesion extent, was manually segmented by an experienced medical doctor. Images were acquired with either a 1.5T (Siemens Magnetom Avanto) or 3T MRI system (Siemens Magnetom Trio). For each case, 4 co-registered multimodal images were provided including anatomical (T1, T2, FLAIR) and diffusion (DWI) MRI. The images were acquired as 3D volumes of $230 \times 230 \times 153$ dimensions at $1 \times 1 \times 1$ mm voxel spacing.

SPES sub-task. The acute stroke penumbra estimation sub-task (SPES) focused on the segmentation of acute ischemic stroke lesions for outcome prediction. The provided dataset included 30 training and 20 testing cases acquired within the first day after onset of symptoms. For the training images, the gold standard segmentation, the penumbra label, was obtained as the mismatch between whole lesion extent and the core delineated in perfusion and diffusion images respectively. Images were acquired with a 3T Phillips systems on two centers. For each case, 7 co-registered modalities were provided including anatomical (T1 contrast, T2), diffusion (DWI) and perfusion (CBF, CBV, TTP, Tmax) MRI. The images were acquired as 3D volumes of $96 \times 110 \times 71$ dimensions at $2 \times 2 \times 2$ mm spacing.

ISLES 2018 Challenge

The ISLES 2018 challenge [45, 21] focused on the task of acute stroke lesion core segmentation solely from CT and CT perfusion images, taken within 8 h of symptom onset. CT derived imaging is actively used in routine clinical practice to triage stroke patients, because of its speed, availability, and lack of contraindications. However, additional MRI is typically needed to inform the treatment decision process since CT imaging does not provide enough sensitivity to distinguish between the lesion parts. MRI using diffusion and perfusion imaging is used to clearly distinguish between the irreversibly damaged tissue ("core") and hypoperfused lesion tissue ("penumbra") that could be salvaged by the treatment. Although the combined extent of core and penumbra can be clearly observed on CT imaging, it is very challenging to differentiate between them without an additional MRI. Thus, for the ISLES 2018 challenge, the provided gold standard segmentation of the lesion core was manually drawn on additional magnetic resonance DWI trace images taken within 3 h of the initial CT scan. Since the goal of the challenge is to segment the lesion core from CT imaging alone, the DWI scans used for generating the ground truth were not distributed as part of the dataset. The provided dataset included 94 labeled training images and 62 unlabeled testing images. For each case, a CT scan, a raw CT perfusion time series (CT-PWI) and four derived perfusion maps (CBF, CBV, MTT and Tmax) were provided. The images were acquired as slabs with a variable number of axial slices, ranging from 2 to 22 depending on the patient, with 5 mm spacing and a resolution of 256 × 256. The raw perfusion time series include between 40 and 63 volumes, acquired 1–2 s apart, of the same dimensions as the CT for each patient.

2.5.2 Effect of WM lesions on brain tissue segmentation

Calgary-Campinas Public Brain MR Dataset

The Calgary-Campinas dataset [46] is an open, multi-vendor, multi-field strength brain 3D MR dataset resulting from a collaborative effort between researchers at the Vascular Imaging Lab located at the University of Calgary and the Medical Image Computing Lab located at the University of Campinas (UNICAMP). The dataset is composed of 359 T1-weighted brain scans from 359 healthy adults with an average age of 53.5 ± 7.8 years, ranging between 29 to 80 years. Images were acquired on scanners from three vendors (GE, Philips, and Siemens) at two different magnetic field strengths of 1.5 T and 3 T, approximately 60 scans were obtained per vendor. Most scans in this dataset have a voxel size of 1.0 × 1.0 × 1.0 mm except for sixty scans acquired at 0.89 × 0.89 × 0.89 mm and another sixty acquired at 1.33 × 1.0 × 1.0 mm. The dataset also includes silver standard brain masks generated through a consensus of several state-of-the-art automatic skull stripping methods. Manual brain mask segmentations for twelve randomly selected subjects performed by an expert are also provided.

MSSEG Challenge

The multiple sclerosis lesions segmentation (MSSEG) challenge [47] was hosted at the MICCAI 2016 international conference. It provided a multicentric database for training consisting of 15 multimodal (T1-w, T1-w gadolinium, T2-w, FLAIR and PD) MR images obtained from MS patients with an average lesion load of 20.8 ± 19.9 ml. Images were acquired on three different scanners at different voxel sizes: five images from a Philips Ingenia 3T scanner at $0.7 \times 0.74 \times 0.74$ mm, five images from a Siemens Verio 3T scanner at $1.1 \times 0.5 \times 0.5$ mm and the remaining five images from a Siemens Aera 1.5T scanner at $1.25 \times 1.03 \times 1.03$ mm. The MR images were rigidly coregistered to the FLAIR scan, which was manually annotated by 7 independent experts, and a consensus gold standard WM lesion segmentation approach was built.

ISBI 2015 Longitudinal MS Lesion Segmentation Challenge

The longitudinal MS lesion segmentation challenge [48] was conducted at the 2015 International Symposium on Biomedical Imaging (ISBI) and focused on evaluation of automatic segmentation methods on data acquired at multiple time points from MS patients. The training data was composed of 21 multimodal scans (T1-w, T2-w, FLAIR and PD) from five MS patients, approximately four timepoints per subject, having an average lesion load of 11.6 ± 10.5 ml. Images were acquired on a 3T MRI Philips scanner with a voxel size of $0.82 \times 0.82 \times 1.17$ mm. Manual delineations were made by two experts identifying and segmenting white matter lesions on the MR images. The MR images from each subject as well as the expert WM lesion delineations were rigidly coregistered to the T1-w scan.

WMH Challenge 2017

The white matter hyperintensity (WMH) segmentation challenge [49] aimed at providing a platform for direct comparison of methods for automatic segmentation of WMH of presumed vascular origin. The provided training set included 60 sets of brain 3D MR images (3D T1 and 2D multislice FLAIR) from 60 subjects of memory clinics showing cognitive impairment with an average lesion load of 17.5 ± 17.1 ml. Images were taken with five different 3T MR scanners from three different vendors (Siemens, Philips and GE) with voxel sizes of $1.0 \times 1.0 \times 1.0$ mm and $0.94 \times 0.94 \times 1.0$ mm. The FLAIR scans from each subject were resampled and coregistered to the 3D T1 scan via an affine transform. The provided gold standard was made with manual annotations of WMHs made by experts.

2.5.3 Brain atrophy quantification datasets

MIRIAD challenge

The Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) challenge [50] was created to test and compare atrophy measurement techniques in dementia. The provided dataset consists of a publicly accessible series of longitudinal T1 MRI scans of 46 mild–moderate Alzheimer's subjects and 23 healthy controls with an average age of 69.5 ± 7.1 years old. The longitudinal scans were taken at intervals of 2, 6, 14, 26, 38 and 52 weeks and 18 and 24 months from baseline, as well as rescan images at three of the timepoints, for both AD and controls. The rescan images were taken during three of the scanning sessions (0, 6 and 38 weeks) without repositioning of the subject. All scans were taken by the same radiographer on the same 1.5 T Signa MRI scanner (GE Medical systems, Milwaukee, WI, USA) with a voxel size of $0.9375 \times 1.5 \times 0.9375$ mm and total image dimensions of $256 \times 124 \times 256$. In our study, we consider both the rescan image pairs and the baseline to 2-weeks image pairs to have a tissue similarity prior that can be used for regularization. From the original dataset, some images were discarded due to poor scan quality or movement artifacts.

ADNI data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) aims at providing multisite consistent longitudinal three-dimensional T1-weighted MRI data to validate biomarkers for use in Alzheimer's disease clinical treatment trials. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. In this thesis, we consider a subset of subjects originally included in the "ADNI1: Complete 1Yr 1.5T" standardized data collection and use similarly preprocessed scans with corrected gradient nonlinearity and B1 and N3 nonuniformity correction. In total, we use 1063 scans from 105 AD patients and 145 healthy control subjects, having 250 pairs of baseline and 1 year of follow-up with 541 scan-rescan images taken at both timepoints. Within this cohort, scans were taken with varied voxel sizes ranging from $0.94 \times 0.94 \times 1.2$ mm to $1.3 \times 1.3 \times 1.2$ mm. In total, 8 different scanners from 2 manufacturers (GE and Siemens) were used for image acquisition. In 45 of the 250 subjects, a different scanner model from the same manufacturer was reported for the 1-year follow-up scan.

Chapter 3

Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks

In this chapter, we propose a deep learning method for the Ischemic Stroke Lesion Segmentation (ISLES) 2018 challenge, which focused on ischemic stroke lesion core segmentation solely from NCCT and CT perfusion images. Our 2D patch-based deep learning approach incorporates techniques to minimize the effect of class imbalance, exploit the use of bilateral symmetry features and post-processing based on uncertainty estimation and lesion morphology. Better segmentation overlap is obtained by alleviating the class imbalance with a balanced training patch sampling along with a class balancing loss function. False positive rate reduction and better borders are obtained by using bilateral symmetry features between the brain hemispheres which are incorporated through the use of symmetric modality augmentation. Lesion border segmentation and volume estimation is improved by performing uncertainty filtering based on test-time dropout averaging. The method has shown state-of-the-art performance in a blind testing set evaluation performed through the official online platform of the ISLES 2018 challenge. The presented approach has been published in the following paper:

Paper published in **Computers in Biology and Medicine** (CBM) OPEN ACCESS
Volume: 115, Published: December 2019
DOI: 10.1016/j.compbiomed.2019.103487
JCR MCB IF 3.434, Q1(8/59)

Computers in Biology and Medicine 115 (2019) 103487

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks

Albert Clèrigues [*], Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, Xavier Lladó

Institute of Computer Vision and Robotics, University of Girona, Spain

ARTICLE INFO

Keywords:
Acute ischemic stroke
Brain
CT
CT perfusion
Automatic lesion segmentation
Convolutional neural networks
CNN

ABSTRACT

The use of Computed Tomography (CT) imaging for patients with stroke symptoms is an essential step for triaging and diagnosis in many hospitals. However, the subtle expression of ischemia in acute CT images has made it hard for automated methods to extract potentially quantifiable information. In this work, we present and evaluate an automated deep learning tool for acute stroke lesion core segmentation from CT and CT perfusion images. For evaluation, the Ischemic Stroke Lesion Segmentation (ISLES) 2018 challenge dataset is used that includes 94 cases for training and 62 for testing. The presented method is an improved version of our workshop challenge approach that was ranked among the workshop challenge finalists. The introduced contributions include a more regularized network training procedure, symmetric modality augmentation and uncertainty filtering. Each of these steps is quantitatively evaluated by cross-validation on the training set. Moreover, our proposal is evaluated against other state-of-the-art methods with a blind testing set evaluation using the challenge website, which maintains an ongoing leaderboard for fair and direct method comparison. The tool reaches competitive performance ranking among the top performing methods of the ISLES 2018 testing leaderboard with an average Dice similarity coefficient of 49%. In the clinical setting, this method can provide an estimate of lesion core size and location without performing time costly magnetic resonance imaging. The presented tool is made publicly available for the research community.

1. Introduction

Stroke is the third largest cause of death and the biggest source of acquired disability worldwide [1]. This condition is caused by a fatally low blood supply in a region of the brain. A shorter time to treatment since onset is strongly linked to a better outcome [2]. The stroke lesion is initially divided in two areas: the infarct core, composed of irreversibly damaged tissue, and the penumbra, tissue at risk that can still be recovered if blood flow is restored. Localization and quantification of the acute core or penumbra is of great clinical interest since it can help evaluate the amount of tissue that could be recovered with different treatments and take better informed decisions.

Non-contrast computed tomography (CT) imaging is fast, inexpensive, ubiquitous and is already used by clinicians as an essential first step for triage, diagnosis and treatment assessment of acute ischemic stroke [3]. Additionally, the information in these images has good prognostic potential, but are difficult to interpret. The infarct core is seen through subtle texture and intensity changes, also called parenchymal hypoattenuation, often masked by artifacts, noise or other tissue abnormalities [4]. Additionally, CT perfusion (CTP) can be used to assess the blood perfusion in the brain. To acquire CTP images, first an intra-venous contrast agent is injected and then repeated scans are made as it spreads through the brain. While CT shows the lesion core, CTP more clearly shows all areas with abnormal perfusion including both core and penumbra. The combination of both is also fast to acquire and might provide enough reliable information for automatic analysis.

Early work on supervised methods for acute stroke detection and segmentation using exclusively CT images relied on hand-crafted features exploiting texture and intensity [5–8].

Recent developments on Convolutional Neural Networks (CNN) [9] have given rise to methods with superior results that are present in the majority of state-of-the-art biomedical segmentation frameworks [10–13]. This trend can also be seen in the most recent methods for stroke lesion segmentation from MR images [14–16]. More specifically, U-shaped architectures based on the U-Net [10] are well suited for dense semantic segmentation. These kind of architectures have seen a number of recent improvements such as their extension for 3D volumetric segmentation [11,17] or the introduction of long and short residual skip connections [15,18]. Stroke lesion segmentation on CT images shares many of the same challenges as MR imaging, but still poses an

[*] Correspondence to: Ed. P-IV, Campus Montilivi, University of Girona, 17003 Girona, Spain.
E-mail address: albert.clerigues@udg.edu (A. Clèrigues).

https://doi.org/10.1016/j.compbiomed.2019.103487
Received 2 July 2019; Received in revised form 30 September 2019; Accepted 5 October 2019
Available online 9 October 2019

inherently different learning problem. Despite the promising results of deep learning applied to brain lesion segmentation, it still presents limitations for real world scenarios that severely limit its applicability. The most critical issues include typically small size of annotated datasets to train, domain and task dependent training procedures, highly unbalanced class extent (i.e. much less lesion tissue than healthy) and overfitting to the training images.

Deep learning has only been recently applied to CT imaging for acute stroke with the 2018 edition of the Ischemic Stroke Lesion Segmentation (ISLES) challenge. This challenge started in 2015 to provide a platform for a fair and direct comparison of automated methods for stroke imaging. The fourth edition in 2018 provides the first public acute stroke dataset using CT and CTP images. From the five challenge finalists, all deep learning based methods, four report the use of CNNs based on the U-Net architecture [10], one of which corresponds to our workshop challenge approach [19]. In these works, the issue of class imbalance was alleviated mainly with the use of cost sensitive loss functions, either class weighting [20,21] or difficulty weighting [22], or using patches with balanced sampling strategies [19].

In this work, we present and evaluate an automated deep learning tool for acute stroke lesion core segmentation from CT and CTP images. The presented tool is a simpler and improved version of the method initially submitted to the ISLES 2018 challenge, which already ranked among the challenge finalists, referred to as the workshop challenge approach. It achieves state-of-the-art performance while offering an easy training procedure and fast inference times. For alleviating class imbalance, both a patch based method with a balanced sampling strategy and a hybrid class weighted loss function are used. The deep learning architecture is an asymmetric encoder–decoder using long and short residual connections as done in recent state-of-the-art networks for dense segmentation [15,23]. Additionally, symmetric modality augmentation is performed that allows to exploit the brain symmetry property between hemispheres to find more robust image features. The introduced improvements with respect to our workshop challenge submission are quantified by crossvalidation on the ISLES 2018 training set. The proposed methodology is evaluated against other state-of-the-art methods with the blind challenge testing set submission, ranking among the top out of 41 entries. In the treatment decision workflow, this tool could provide a fast estimate of the lesion core location and volume without having to perform costly MR imaging. We release this tool to the community, available at https://github.com/NIC-VICOROB/stroke-core-ct-segmentation.

2. Materials

2.1. Data

The ISLES 2018 challenge tackled the segmentation of stroke lesion core from acute CT scans, taken within 8 h of stroke onset. The provided dataset (Kistler, 2013; Maier, 2017) includes 94 labeled training images and 62 unlabeled testing images. For each case, a CT scan, a raw CT perfusion time series (CT-PWI) and four derived perfusion maps (CBF, CBV, MTT and Tmax) are provided. The images were acquired as slabs with a variable number of axial slices, ranging from 2 to 22 depending on the patient, with 5 mm spacing and a resolution of 256 × 256. The raw perfusion time series include between 40 and 63 volumes, acquired 1–2 s apart, of the same dimensions as the CT for each patient. The provided gold standard was manually drawn on additional magnetic resonance DWI trace images not included in the challenge testing set, where the infarct core is seen more clearly, taken within 3 h of the initial CT scan.

2.1.1. Pre-processing

From the provided modalities, we only consider the use of CT and the four derived CT perfusion maps (CBF, CBV, MTT and Tmax), omitting the raw CT-PWI time series. Image pre-processing is then applied to the provided images in two steps: Firstly, the CT image is skull stripped and, secondly, a modality augmentation to exploit the symmetry of brain hemispheres is performed.

CT Skull stripping. The brain mask for skull stripping is obtained from the non-zero values of the sum of the four provided perfusion images, which did not take any value on the skull. Finally, the mask is multiplied with the CT image, which leaves only the desired brain tissue.

Symmetric modality augmentation. The use of the symmetry property showed significant improvements on chronic stroke lesion segmentation in MR images [24]. Since typically only one hemisphere of the brain is affected by the stroke, the brain mid-sagittal symmetry can be exploited to assess differences between both hemispheres and locate the lesion more accurately. In our case, we take advantage of the symmetry property by creating a symmetric version of each provided modality. In this way, a single patch will include information from the same spatial location of both hemispheres. To generate the symmetric modalities we first flip the CT images by the mid-sagittal axis. Since the images are not perfectly centered in the volume and some are slightly rotated, the opposing hemispheres might not be correctly aligned after the flip. Hence, we use FSL FLIRT [25] constrained to an axial affine transformation to linearly register both images and roughly align opposing hemispheres. In this case, a linear registration is sufficient since the symmetry features are not expected to rely on fine differences but rather on overall differences of patch intensity, parenchyma and/or perfusion statistics. Finally, the provided modality volumes are merged with the symmetrically augmented and used together for segmentation as an image with ten modality volumes. In this way, a single patch will additionally include bilateral information of all modalities. Fig. 1 shows an example case with the provided and augmented modalities.

3. Method

The proposed approach is a 2D patch based deep learning approach for segmentation of the acute stroke lesion core from CT perfusion images. Since the lesion core class represents around 5% of the brain tissue in the training set, class imbalance is an issue that needs to be dealt with. If no deliberate action is taken, the training set would include fewer examples of lesion than healthy tissue, which would bias the learning and worsen segmentation performance. Additionally, overfitting to the training set is likely, considering the small quantity of data, which would cause bad performance for other images. To minimize this effect, the training is regularized by using: (a) data augmentation with elastic deformation fields, (b) dropout layers that introduce noisy updates during training and (c) early stopping that interrupts training when no more generalizable knowledge can be learned. Finally, a combination of classification uncertainty estimation and use of highly overlapping patches further reduces outliers and segmentation artifacts.

3.1. Class imbalance

The most common techniques to alleviate this issue for deep learning methods are three: cost sensitive loss functions, which assign different cost to misclassification of examples from different classes [26]; the use of patches with deliberate sampling, typically aiming to overrepresent the minority class, or multi-phase training, where a part of the network is retrained with a different class distribution. In this work, we propose the use of both a balanced patch sampling and a cost sensitive loss function to alleviate the imbalance.

The employed sampling strategy is an extension of a recent proposal for brain lesions in general [12]. The strategy has been extended to take into account the anatomy and pathophysiology of acute stroke. In practice, a target number of patches is set for each patient. Then, half of the patches are extracted centered on lesion voxels and the other half on healthy ones. These are sampled in regular spatial steps to ensure all parts of the volume are uniformly represented. For the lesion class sampled voxels have a random offset applied in the x and y axis before

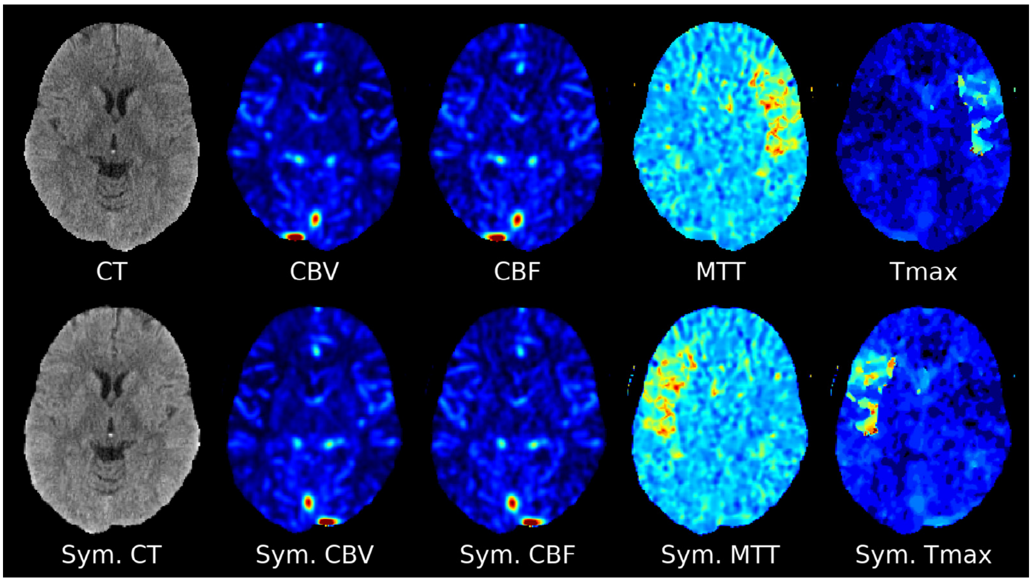


Fig. 1. Top row: Provided CT and derived CT perfusion maps. Bottom row: Resulting symmetrically augmented modalities.

patch extraction, as done by Guerrero et al. [15]. This offset is sampled from a random uniform distribution and is limited to half of the patch size to ensure the originally sampled voxel is inside the finally extracted patch. This increases the representation of areas adjacent to the core label, the penumbra region, while providing a degree of translational data augmentation. The patches will be extracted centered on these voxels. For patients with smaller lesions, several patch extractions from the same lesion voxel and data augmentation are applied, using the elastic deformation described in [27] with parameters α = 2.5 and γ = 0.12, to reach the target number of patches per patient. In this way, only if the number of lesion voxels is smaller than the target number, they will be repeated and augmented using elastic deformation. On average, the augmented patches amount to 5% of the training set. The use of this patch sampling strategy raised the lesion voxel fraction in the training set from 5% to 12%.

Additionally, we use a cost sensitive loss function that is the sum of the Generalized Dice Loss (GDL) [28] and the crossentropy loss to further minimize the effects of class imbalance. While the crossentropy loss is minimized with correct confident predictions, the GDL is minimized by maximizing the relative overlap between prediction and ground truth. In practice, jointly minimizing both terms provides the crossentropy convergence properties with the balancing class weighting of the GDL.

However, despite the use of both techniques, the overlap segmentation is decreased when bigger patch sizes are considered due to worsened imbalance, since larger patch sizes will tend to include a bigger ratio of healthy to lesion voxels. After empirical testing with several patch sizes ranging from 16 × 16 to 96 × 96, we choose a patch size of 64 × 64 that offers the best compromise between a large receptive field and worsened class imbalance.

3.2. Deep learning architecture

The employed network, depicted in Fig. 2, is a 2D asymmetric residual encoder–decoder that produces whole patch predictions. It is based on recent state-of-the-art networks for chronic stroke [15] and related biomedical tasks [11]. The network has five resolution steps with 8 base filters, which are doubled in each step, resulting in a latent space with 128 feature maps of 4 × 4 resolution. It has long and short residual connections to ease gradient flow, which improves convergence properties and allows for better accuracy [29]. The asymmetry comes from the reduced number of parameters found in the decoder branch. It has been shown that the role of the decoder is not as critical and its complexity can be reduced without damaging the performance [23]. In this way, the residual blocks have two convolutional layers in the encoder and one in the decoder, resulting in 75% and 25% of the parameters in each respectively. Additionally, it includes prediction dropout layers that will be used for estimating the uncertainty in classification to minimize outliers.

3.3. Pipeline overview

In this section we will briefly describe the different parts of the training and testing pipeline to train the network and use it to segment the desired images.

Training. In the training phase, the randomly initialized network weights are trained with patch training and validation sets built from the provided images. A total of 376,000 patches, 4000 from each case, of size 64 × 64 are extracted using the sampling strategy described in Section 3.1 to create the training set. The sum of the Generalized Dice and Crossentropy loss is used as the objective function. During training, the weights are updated with the Adadelta optimizer [30], which requires no manual tuning of learning rate. After several empirical tests, we use a batch size of 64 patches during training since it provides a good compromise between sensitivity and overfitting. The batch size determines the number patches whose gradients will be averaged before a network weight update during training. A bigger batch size averages the gradients of more patches, which improves the overall accuracy while giving less weight to errors in individual samples. To further minimize overfitting, early stopping with a patience of ten epochs is performed when the sum of error rate and L1 loss on the validation set reaches a global minimum. We set the low number of ten patience epochs to avoid excessive overfitting to the validation set, given the small size of the dataset. Although further training might still improve the validation metrics it could be at the cost of overfitting to the validation images and worsening the performance with testing images. In practice, the networks are trained for a maximum of 100 epochs or until the early stopping condition has been met, storing the network weights with the best validation metrics. The number of training epochs ranges from 20 to 40 for the reported experiments.

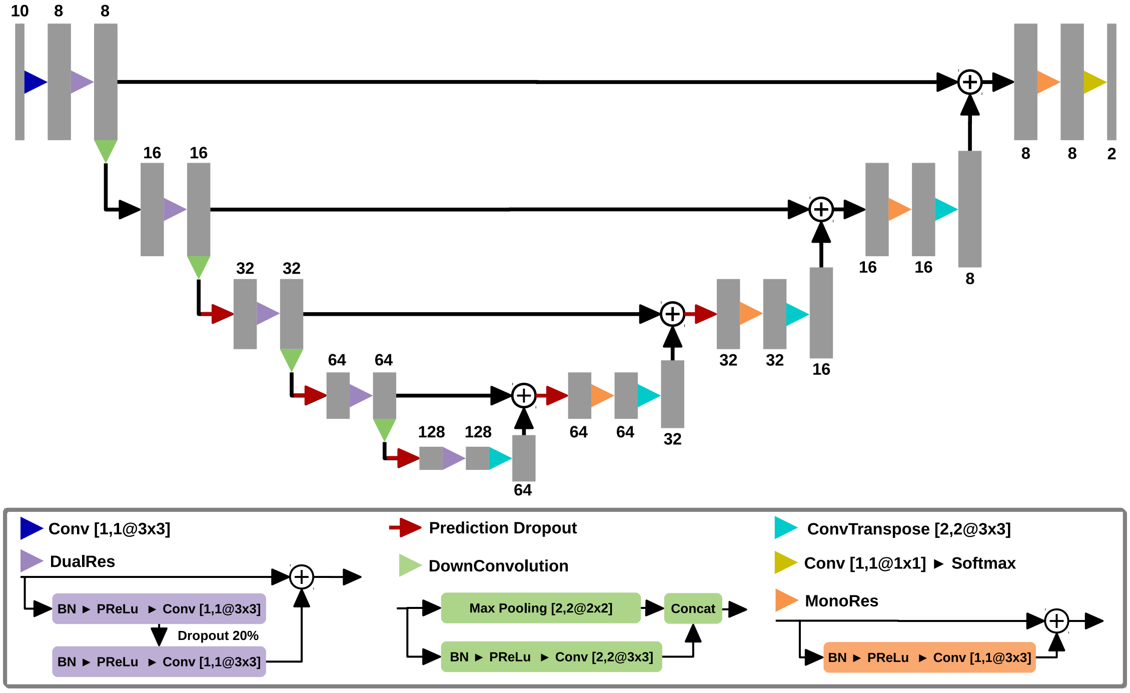


Fig. 2. Diagram of the employed deep learning architecture, an asymmetrical residual encoder–decoder CNN. Gray rectangles represent feature maps with the number of features indicated either on top or bottom. For the convolutional layers, [S_x,S_y@K_x×K_y] indicates the strides and kernel size in each axis respectively. Red arrows mark the location where dropout is applied at prediction time for average uncertainty filtering.

Testing. In the testing phase we predict the class probability distribution for each voxel of a given image with the trained network. Firstly, patches are extracted from the whole image at regular spatial intervals to make sure all parts of the volume are represented. Furthermore, a degree of overlap is considered to improve spatial label coherence and minimize boundary artifacts. Each extracted patch is forward passed through the network and its predicted probabilities accumulated into a common space, preserving its spatial location. Finally, the average of accumulated probabilities in each voxel is made. Additionally, uncertainty filtering by averaging is applied to each patch forwarded through the network. It has been shown that a patch predicted while using dropout can be considered a Monte Carlo sample from the unknown classification probability distribution [31]. In our case, for each patch, 3 forward passes are performed with a voxel-wise prediction dropout rate of 10%. As suggested by [32], dropout in prediction is only performed in the deepest resolution steps as seen in Fig. 2. Finally, the probability distribution is computed as the voxel-wise average of the three noisy predictions.

Post-processing. In the post-processing phase, a binary segmentation is produced from the predicted probability maps. It is performed in two steps, first the probabilities are binarized according to a threshold T and then a connected component filtering removes lesions smaller than S_{min} voxels. The parameters T and Smin that optimize the DSC and HD of the tool are found through grid search for each evaluation. More specifically, we test 9 different thresholds T, from 0.1 to 0.9 in 0.1 steps, and 6 minimum lesion sizes (Smin) ranging from 10 to 500 voxels. Each combination of these parameters is then used to binarize the predicted probability maps and compute segmentation metrics. We select the T and Smin that jointly optimize the DSC and HD metrics, the ones used to rank the ISLES 2018 challenge workshop participants.

3.4. Implementation details

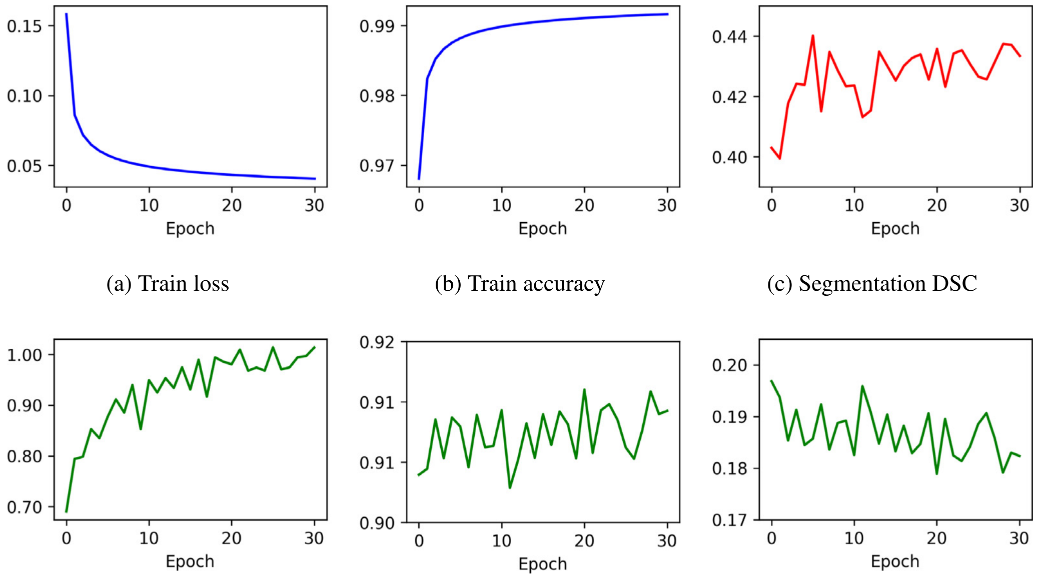
The proposed method has been implemented with Python, using the Torch scientific computing framework [33]. All experiments have been run on a GNU/Linux machine running Ubuntu 18.04 with 64 GB of RAM memory and an Intel® Core™ i7-7800X CPU. The network training and testing has been done with an NVIDIA TITAN X GPU (NVIDIA corp, United States) with 12 GB G5X memory.

4. Evaluation and results

The proposed methodology is evaluated with a crossvalidation experiment showing the improvements against our initial workshop challenge approach and with an external blind evaluation against state-of-the-art methods using the testing set. The evaluation metrics for both experiments include the Dice similarity coefficient (DSC) [34] and Hausdorff distance (HD), the ones considered to rank the workshop challenge participants. Additionally, we also consider other metrics more relevant to the clinical setting such as positive predictive value (PPV), sensitivity and coefficient of determination (COD), also called R^2, between the predicted and true core volume. Finally, we consider the dependent t-test for paired samples to assess the statistical significance of differences between the evaluation results.

4.1. Crossvalidation experiment

The purpose of this experiment is to quantitatively assess the improvements introduced to the proposed method with respect to our workshop challenge approach (the baseline). Mainly, the improvements come from a more regularized network training procedure, symmetric modality augmentation and uncertainty filtering. Additionally, a single network is used in contrast with the two networks in cascade configuration of the baseline. Thanks to the added improvements we can avoid the use of the second model, which simplifies the training procedure and reduces inference times. The current more regularized training procedure uses the sum of GDL and crossentropy as loss function and the sum of L1 loss and error rate for early stopping. However, for the baseline approach [19] the networks were trained using crossentropy as loss function and a probabilistic Dice loss [28] for early stopping. Additionally, we are able to use bigger 64 × 64 patches without a decrease in segmentation performance as it happened with the baseline,



(a) Train loss

(b) Train accuracy

(c) Segmentation DSC

(d) Validation loss

(e) Validation accuracy

(f) Validation L1 + error rate

Fig. 3. Loss and accuracy plots for a single cross-validation fold.

where using 64 × 64 patches resulted in 4% lower DSC than using 48 × 48. The bigger patch size of 64 × 64 offers a bigger receptive field from which to learn features.

The experiment consisted of four evaluations, first the baseline and then three with the incremental improvements that comprise the proposed method. Each evaluation is performed in 5 crossvalidation folds across the 94 labeled images of the ISLES 2018 dataset, having 75 training and 19 validation images for each fold. Since some scans correspond to different regions of the same patient, we ensure that all the same patient scans are within the same set. In each fold, a single network is trained with the training patches and then the validation volumes are predicted, resulting in a probability map for each case. After all five folds have finished there will be one predicted probability map for each of the 94 training images. Finally, the probability maps are postprocessed using T and Smin, found through grid search, that achieve the best segmentation metrics across all folds of the crossvalidation.

Fig. 3 shows loss and accuracy plots for a single cross-validation fold, since the other folds were of similar nature. Additionally, it shows the early stopping metric value, the L1 loss plus error rate, and the segmentation DSC of the validation images. The figure shows how the loss function evaluated on the validation set increases, instead of decreasing, while the validation accuracy improves. For this reason, we do not use the validation loss and instead use the sum of L1 loss and Error rate as a monitored metric on the patch validation set for early stopping, since it is more correlated with segmentation DSC of the validation images. In this case, the early stopping metric reaches a global minimum in epoch 20 where the segmentation DSC begins to stabilize. Although further training might still improve the validation metrics it might be at the cost of overfitting to it and worsening the performance with testing images.

Table 1 shows the evaluation metrics obtained from the baseline and incremental improvements. Compared with the baseline, the regularized training procedure with a single model significantly improves the DSC and sensitivity ($p < 0.02$). When augmented modalities are additionally considered, the PPV significantly improves although the sensitivity is reduced ($p < 0.05$). Moreover, when uncertainty filtering is considered the HD is significantly reduced at the expense of a lower sensitivity ($p < 0.03$). In general, all introduced improvements raise the COD, meaning that the estimated volume is closer to the gold standard. In summary, the proposed tool provides significantly better DSC, HD and PPV ($p < 0.05$) than the baseline with a marginal higher sensitivity.

Fig. 4 shows qualitative evaluations of the incremental improvements for three representative cases. As compared with the baseline, the regularized training achieves better sensitivity and specificity in all cases, reducing the amount of false positives and negatives. The addition of symmetric modalities overall improves lesion localization but can reduce the sensitivity for some samples. For instance, the use of symmetric modalities increases the false positives in the middle row case. Finally, the bottom row is a good example of the effect of uncertainty filtering in the majority of cases, improving lesion localization and estimated volume. However, in some cases it may also introduce additional outliers as seen in the top row case, where false positives appear in the upper part of the lesion.

4.2. ISLES 2018 testing evaluation

For segmentation of the 62 unlabeled testing images from the ISLES 2018 dataset, we used all five networks, one from each fold, that were trained for the crossvalidation evaluation with all improvements. An averaging approach is used where each patch is passed through the five trained models and the five predictions are averaged together to produce a single patch prediction. In this way, bootstrap aggregation [35] is performed, where each network is trained with a different subset of training data. Finally, the resulting class probability maps of the testing images are binarized using the previously computed optimal parameters $T = 0.2$ and $S_{min} = 200$ from the crossvalidation experiment. Table 2 shows ongoing benchmark leaderboard of the ISLES 2018 testing set sorted by average DSC, where the proposed methodology ranks among the top entries out of 41 participants.

5. Discussion

The results of the ISLES 2018 testing set evaluation show that the proposed methodology achieves state-of-the-art performance ranking 2nd in the ongoing benchmark leaderboard among 41 submissions. The approach by Song et al. [20] manages to achieve a 2% higher DSC by additionally using the 40 or more volumes that comprise each raw perfusion time series (CT-PWI) to further extract features for segmentation. The use of the raw perfusion time series would involve an increase in memory requirements and processing time, additionally making the training procedure more complex. In our case, we still use some of the

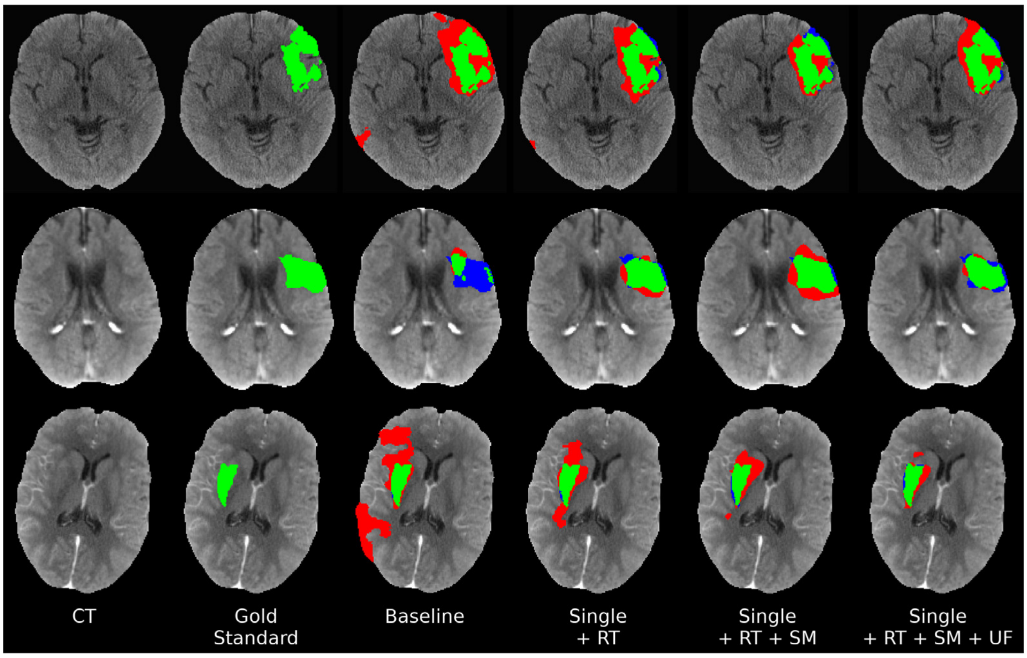


Fig. 4. Lesion core segmentation masks of the baseline and incremental improvements. True positives are denoted in green, false positives in red and false negatives in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Evaluation metrics of the crossvalidation experiment in the ISLES 2018 training set. The baseline results correspond to our workshop challenge approach while Single refers to the current approach using a single network. The evaluated improvements are three: the regularized training (RT), symmetric modality augmentation (SM) and the uncertainty filtering (UF).

Method	*T*	S_{min}	DSC (%)	PPV (%)	Sens. (%)	HD	R^2
Baseline	.4	200	49.0 ± 23.6	46.9 ± 29.5	57.2 ± 26.7	29.5 ± 18.9	.67
Single +RT	.1	300	53.5 ± 24.6	51.8 ± 28.9	66.0 ± 25.5	29.8 ± 23.9	.74
Single +RT +SM	.2	200	54.8 ± 24.8	58.3 ± 29.8	63.7 ± 25.3	26.6 ± 19.9	.78
Single +RT +SM +UF	.2	200	54.7 ± 24.2	57.8 ± 29.1	60.9 ± 25.0	23.5 ± 15.8	.82

Table 2
Top 10 entries of the ongoing benchmark leaderboard, last accessed 25/06/2019, of ISLES 2018 testing set as ranked by average DSC. The entry of the presented tool is highlighted in bold. *Values divided by 1,000,000.

Rank	User	DSC	PPV	Sensitivity	HD*
1	songt1 [20]	0.51 ± 0.31	0.55 ± 0.36	0.55 ± 0.34	19.4 ± 39.5
2	**clera2 (ours)**	**0.49 ± 0.31**	**0.51 ± 0.36**	**0.57 ± 0.35**	**11.3 ± 31.6**
3	pengl1 [21]	0.49 ± 0.31	0.56 ± 0.37	0.53 ± 0.33	19.4 ± 39.5
4	zhans10	0.49 ± 0.32	0.53 ± 0.35	0.54 ± 0.35	17.7 ± 38.2
5	cheny11 [36]	0.48 ± 0.32	0.59 ± 0.38	0.46 ± 0.33	9.7 ± 29.6
6	lilic2	0.48 ± 0.32	0.48 ± 0.34	0.6 ± 0.36	17.7 ± 38.2
7	liliy8	0.48 ± 0.31	0.5 ± 0.36	0.55 ± 0.34	19.4 ± 39.5
8	liliy2	0.47 ± 0.32	0.53 ± 0.36	0.47 ± 0.32	16.1 ± 36.8
9	xiaoh3 [22]	0.47 ± 0.31	0.56 ± 0.37	0.49 ± 0.33	19.4 ± 39.5
10	zhuoj2	0.47 ± 0.32	0.51 ± 0.36	0.54 ± 0.36	11.3 ± 31.6

information obtained from the absorption curve parametrization of the raw perfusion time series in the 4 perfusion parameter maps (CBF, CBV, MTT and Tmax). Despite the potential performance improvement of also processing the raw time series as shown by Song et al. [20], we avoid it in favor of reducing the training complexity and provide faster inference times.

The crossvalidation experiment shows the big influence that class imbalance and training regularization can have on segmentation performance. For instance, the class weighting properties of the focal loss allow the use of bigger 64 × 64 patches without worsened imbalance and provides a DSC improvement of 4.5% over the baseline. However, this patch size is too small to fit both brain hemispheres simultaneously and makes implausible exploiting symmetrical features. The use of symmetric modality augmentation allows learning of these features without having to use bigger patches that would worsen class imbalance. Despite the overall improvement from augmented modalities, some cases are actually worsened, as seen in the middle row of Fig. 4 with a lower PPV that increases false positives. Finally, we noted that the use of uncertainty filtering significantly reduced outliers but also harmed segmentation performance with bigger dropout rates. We found that averaging the output of several passes with a low dropout rate of 10% in prediction was enough to reduce outliers without significantly harming the overlap performance. Despite the marginally worsened DSC, PPV and sensitivity that uncertainty filtering provides, we believe the significantly reduced HD and better estimation of the core volume are more desirable properties in the clinical setting. Additionally, since each patch will require the average of three noisy predictions, this effectively triples the network inference time. However, even when considering the pre-processing step, segmentation of the largest images typically takes under two minutes in our system.

6. Conclusions

In this work, we presented and evaluated an automated method for acute stroke lesion core segmentation from CT and CTP images. The presented tool achieves state-of-the-art performance while using a simple training procedure with a single network. The training requires minimal tuning of parameters thanks to the Adadelta optimizer and a robust class imbalance handling using balanced patch sampling and a class weighting loss function. We improve segmentation performance with a novel way of using the symmetry property of brain hemispheres in patch based methods. We also explore the use of prediction dropout layers to reduce outliers and improve lesion core volume estimation, a predictor of clinical severity and outcome in ischemic stroke [37]. This tool can provide with an estimate of core location and volume without acquiring time costly MR images. In the clinical setting, this estimate can be used to guide treatment decisions or help assess the need for further MR imaging. A trainable implementation of the presented tool is freely released for the research community.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Albert Clèrigues holds an FPI grant from the Ministerio de Ciencia, Innovación y Universidades with reference number PRE2018-083507. This work has been partially supported by Retos de Investigación TIN2015-73563-JIN and DPI2017-86696-R from the Ministerio de Ciencia, Innovación y Universidades. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research.

References

[1] C.L. Sudlow, C.P. Warlow, Comparable studies of the incidence of stroke and its pathological types: Results from an international collaboration. International stroke incidence collaboration., Stroke 28 (3) (1997) 491–499.

[2] Sunil A. Sheth, Reza Jahan, Jan Gralla, Vitor M. Pereira, Raul G. Nogueira, Elad I. Levy, Osama O. Zaidat, Jeffrey L. Saver, Time to endovascular reperfusion and degree of disability in acute stroke, Ann. Neurol. 78 (4) (2015) 584–593, http://dx.doi.org/10.1002/ana.24474.

[3] Michael H. Lev, Jeffrey Farkas, Joseph J. Gemmete, Syeda T. Hossain, George J. Hunter, Walter J. Koroshetz, R. Gilberto Gonzalez, Acute stroke: Improved nonenhanced CT detection—Benefits of soft-copy interpretation by using variable window width and center level settings, Radiology 213 (1) (1999) 150–155, http://dx.doi.org/10.1148/radiology.213.1.r99oc10150.

[4] Rafał Józwiak, Artur Przelaskowski, Grzegorz Ostrek, Conceptual improvements in computer-aided diagnosis of acute stroke, J. Med. Inf. Technol. 17 (2011).

[5] Andrius Ušinskas, Romualdas Dobrovolskis, Bernd F. Tomandl, Ischemic stroke segmentation on CT images using joint features, Informatica 15 (2) (2004) 283–290.

[6] M. Chawla, S. Sharma, J. Sivaswamy, L.T. Kishore, A method for automatic detection and classification of stroke from brain CT images, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009, pp. 3581–3584, http://dx.doi.org/10.1109/IEMBS.2009.5335289.

[7] Fuk-hay Tang, Douglas K.S. Ng, Daniel H.K. Chow, An image feature approach for computer-aided detection of ischemic stroke, Comput. Biol. Med. 41 (7) (2011) 529–536, http://dx.doi.org/10.1016/J.COMPBIOMED.2011.05.001.

[8] N. Hema Rajini, R. Bhavani, Computer aided detection of ischemic stroke using segmentation and texture features, Measurement 46 (6) (2013) 1865–1874, http://dx.doi.org/10.1016/J.MEASUREMENT.2013.01.010.

[9] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, http://dx.doi.org/10.1162/neco.1989.1.4.541.

[10] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241, arXiv:1505.04597.

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, Olaf Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2016, pp. 424–432, http://dx.doi.org/10.1007/978-3-319-46723-8_49.

[12] Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, Ben Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78, http://dx.doi.org/10.1016/J.MEDIA.2016.10.004.

[13] Jose Dolz, Christian Desrosiers, Ismail Ben Ayed, 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study, NeuroImage 170 (2017) 456–470, http://dx.doi.org/10.1016/J.NEUROIMAGE.2017.04.039.

[14] Liang Chen, Paul Bentley, Daniel Rueckert, Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks, NeuroImage: Clin. 15 (2017) 633–643, http://dx.doi.org/10.1016/J.NICL.2017.06.016.

[15] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M.C. Valdés-Hernández, D.A. Dickie, J. Wardlaw, D. Rueckert, White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks, NeuroImage: Clin. 17 (2018) 918–934, http://dx.doi.org/10.1016/J.NICL.2017.12.022.

[16] Rongzhao Zhang, Lei Zhao, Wutao Lou, Jill M. Abrigo, Vincent C.T. Mok, Winnie C.W. Chu, Defeng Wang, Lin Shi, Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional densenets, IEEE Trans. Med. Imaging 37 (9) (2018) 2149–2160, http://dx.doi.org/10.1109/TMI.2018.2821244.

[17] Fausto Milletari, Nassir Navab, Seyed-Ahmad Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, 2016, arXiv:1606.04797.

[18] Ke Zhang, Miao Sun, Tony X. Han, Xingfang Yuan, Liru Guo, Tao Liu, Residual networks of residual networks: Multilevel residual networks, 2016, arXiv:1608.02908. http://dx.doi.org/10.1109/TCSVT.2017.2654543.

[19] Albert Clèrigues, Sergi Valverde, Jose Bernal, Kaisar Kushibar, Mariano Cabezas, Arnau Oliver, Xavier Lladó, Ensemble of convolutional neural networks for acute stroke anatomy differentiation, in: International MICCAI Brainlesion Workshop, 2018.

[20] Tao Song, 3D Multi-scale U-Net with atrous convolution for ischemic stroke lesion segmentation, in: International MICCAI Brainlesion Workshop, 2018.

[21] Pengbo Liu, Stroke lesion segmentation with 2D convolutional neutral network and novel loss function in: International MICCAI Brainlesion Workshop, 2018.

[22] Xiaojun Hu, Weilin Huang, Sheng Guo, Matthew R. Scott, StrokeNet: 3D Local refinement network for ischemic stroke lesion segmentation, in: International MICCAI Brainlesion Workshop, 2018.

[23] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, Eugenio Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, 2016, arXiv preprint arXiv:1606.02147. arXiv:1606.02147.

[24] Yanran Wang, Aggelos K. Katsaggelos, Xue Wang, Todd B. Parrish, A deep symmetry convnet for stroke lesion segmentation, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 111–115, http://dx.doi.org/10.1109/ICIP.2016.7532329.

[25] Mark Jenkinson, Peter Bannister, Michael Brady, Stephen Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, NeuroImage 17 (2) (2002) 825–841.

[26] Charles Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence, 2001, pp. 973–978.

[27] P.Y. Simard, D. Steinkraus, J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings, vol. 1, IEEE Comput. Soc, pp. 958–963. http://dx.doi.org/10.1109/ICDAR.2003.1227801.

[28] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Sebastien Ourselin, M. Jorge Cardoso, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, Lecture Notes in Comput. Sci. 10553 LNCS (2017) 240–248, http://dx.doi.org/10.1007/978-3-319-67558-9_28, arXiv:1707.03237.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[30] Matthew D. Zeiler, ADADELTA: An adaptive learning rate method, 2012, arXiv preprint arXiv:1212.5701, abs/1212.5. http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.

[31] Yarin Gal, Zoubin Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2015, pp. 1050–1059. arXiv:1506.02142.

[32] Tanya Nair, Doina Precup, Douglas L. Arnold, Tal Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 655–663. http://dx.doi.org/10.1007/978-3-030-00928-1_74.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, Adam Lerer, Automatic differentiation in pytorch, in: Neural Information Processing Systems, 2017.

[34] Lee R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302, http://dx.doi.org/10.2307/1932409.

[35] Leo Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140, http://dx.doi.org/10.1007/BF00058655.

[36] Yu Chen, Yuexiang Li, Yefeng Zheng, Ensembles of modalities fused model for ischemic stroke lesion segmentation, in: International MICCAI Brainlesion Workshop, 2018.

[37] Karl-Olof Löuvbld, Alison E. Baird, Gottfried Schlaug, Andrew Benfield, Bettina Siewert, Barbara Voetsch, Ann Connor, Cara Burzynski, Robert R. Edelman, Steven Warach, Ischemic lesion volumes in acute stroke by diffusion-weighted magnetic resonance imaging correlate with clinical outcome, Ann. Neurol. 42 (2) (1997) 164–170, http://dx.doi.org/10.1002/ana.410420206.

Chapter 4

Acute and sub-acute stroke lesion segmentation from multimodal MRI

In this chapter, we propose a deep learning based approach for both the sub-acute ischemic stroke lesion segmentation (SISS) and the acute stroke penumbra estimation (SPES) sub-tasks of the Ischemic Stroke Lesion Segmentation (ISLES) 2015 challenge. Our 3D patch-based deep learning approach achieves state-of-the-art results by utilizing techniques for class imbalance handling, bilateral feature extraction and morphology based post-processing. To mitigate the class imbalance, we use a balanced training patch sampling along with a difficulty weighted loss function. The use of augmented symmetric modalities allows the network to exploit features based on the symmetry of brain hemispheres despite the use of small patches. The proposed approach demonstrated state-of-the-art performance by achieving first position at the ongoing blind testing set evaluation leaderboards of both tasks, SISS and SPES, from the ISLES 2015 challenge. The presented methodology has been published in the following paper:

Paper published in **Computer Methods and Programs in Biomedicine** (CMPB)
Volume: 194, Published: October 2020
DOI: 10.1016/j.cmpb.2020.105521
JCR CSTM IF 3.424, Q1(15/104)

Computer Methods and Programs in Biomedicine 194 (2020) 105521

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Acute and sub-acute stroke lesion segmentation from multimodal MRI

Albert Clèrigues*, Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, Xavier Lladó

Institute of Computer Vision and Robotics, University of Girona, P-IV, Campus Montilivi, 17003 Girona Spain

ARTICLE INFO

Article history:
Received 25 April 2019
Revised 30 March 2020
Accepted 23 April 2020

Keywords:
Brain
MRI
Ischemic stroke
Automatic lesion segmentation
Convolutional neural networks
Deep learning

ABSTRACT

Background and objective. Acute stroke lesion segmentation tasks are of great clinical interest as they can help doctors make better informed time-critical treatment decisions. Magnetic resonance imaging (MRI) is time demanding but can provide images that are considered the gold standard for diagnosis. Automated stroke lesion segmentation can provide with an estimate of the location and volume of the lesioned tissue, which can help in the clinical practice to better assess and evaluate the risks of each treatment.

Methods. We propose a deep learning methodology for acute and sub-acute stroke lesion segmentation using multimodal MR imaging. We pre-process the data to facilitate learning features based on the symmetry of brain hemispheres. The issue of class imbalance is tackled using small patches with a balanced training patch sampling strategy and a dynamically weighted loss function. Moreover, a combination of whole patch predictions, using a U-Net based CNN architecture, and high degree of overlapping patches reduces the need for additional post-processing.

Results. The proposed method is evaluated using two public datasets from the 2015 Ischemic Stroke Lesion Segmentation challenge (ISLES 2015). These involve the tasks of sub-acute stroke lesion segmentation (SISS) and acute stroke penumbra estimation (SPES) from multiple diffusion, perfusion and anatomical MRI modalities. The performance is compared against state-of-the-art methods with a blind online testing set evaluation on each of the challenges. At the time of submitting this manuscript, our approach is the first method in the online rankings for the SISS (DSC=0.59 ± 0.31) and SPES sub-tasks (DSC=0.84 ± 0.10). When compared with the rest of submitted strategies, we achieve top rank performance with a lower Hausdorff distance.

Conclusions. Better segmentation results are obtained by leveraging the anatomy and pathophysiology of acute stroke lesions and using a combined approach to minimize the effects of class imbalance. The same training procedure is used for both tasks, showing the proposed methodology can generalize well enough to deal with different unrelated tasks and imaging modalities without hyper-parameter tuning. In order to promote the reproducibility of our results, a public version of the proposed method has been released to the scientific community.

1. Introduction

Stroke is a medical condition by which an abnormal blood flow in the brain causes the death of cerebral tissue. Stroke is the third cause of morbidity worldwide, after myocardial infarction and cancer, and the most prevalent cause of acquired disability [1]. The affected tissue in the acute phase can be divided into three concentric regions depending on the potential for recovery, also referred as salvageability: core, penumbra and benign oligemia [2]. The core, located at the center, is formed by irreversibly damaged tissue from a fatally low blood supply. The penumbra, located around the core, represents tissue at risk but that can still be recovered if blood flow is quickly restored. Finally, the benign oligemia is the outer most ring whose vascularity has been altered but is not at risk of damage. Once the symptoms of stroke have been identified, a shorter time to treatment is highly correlated with a positive outcome [3]. Mechanical thrombectomy is a strongly recommended option for eligible patients [4]. However, this surgery is not free of risks. An overall complication rate of 15.3% was observed in a year long study [5]. In the treatment decision context, an estimate of the salvageable tissue can aid physicians take more informed treatment decisions.

* Corresponding author.
E-mail address: albert.clerigues@udg.edu (A. Clèrigues).

https://doi.org/10.1016/j.cmpb.2020.105521
0169-2607/© 2020 Elsevier B.V. All rights reserved.

The Ischemic Stroke Lesion Segmentation (ISLES) challenge started in 2015 to provide a platform for fair and direct comparison of automated methods. It included two sub-tasks, the sub-acute ischemic stroke lesion segmentation (SISS) and the acute stroke penumbra estimation (SPES). The following ISLES 2016 and 2017 editions changed its focus from lesion segmentation to chronic lesion outcome prediction from MRI. In the 2015 ISLES workshop results, the top three methods in the SPES sub-task all used Random Decision Forests (RDFs) [6] using hand-crafted features [7–9]. RDFs were typically used in methods for stroke lesion segmentation due to their excellent generalization properties, which make them well suited for difficult tasks with few training samples [10]. Recent advances on convolutional neural networks (CNNs) [11] have achieved superior results and are currently replacing RDFs in most state-of-the-art methods. In contrast with RDFs, CNNs enable the joint learning of optimal features and classification criteria at training time for the specific task. However, these kind of networks are still restricted by the architectural design, the amount and quality of available data and the training procedure. Recently, advances in regularization techniques and data imbalance handling allow for increased CNN generalization performance in brain lesion segmentation that rivals that of RDFs. The best method in the SISS sub-task of the 2015 ISLES workshop employed a deep learning strategy consisting of a dual path encoder network with a conditional random field (CRF) post-processing [12]. More recently, Zhang et al [13] achieved similar results in the ISLES 2015 testing set by using a similar CNN trained with a deep supervision technique and a multi-scale loss function. Similarly, the work of Karthik et al. [14] further improved results in ISLES 2015 reaching a DSC of 0.70 on the SISS training set, but does not perform an evaluation on the publicly available blind test set. Despite the good results of these kind of networks, the U-Net architecture [15], an encoder-decoder network, is replacing other state-of-the-art architectures for stroke lesion segmentation. This is clearly seen in the submissions for the ISLES 2017 challenge, where 10 out of the 14 participating methods, including the top three, used CNNs based on the U-Net architecture [16]. Recent approaches for stroke lesion segmentation from MR imaging also used these kind of networks. The work by Olivier et al. [17] used a U-Net based network with a two phase training first using whole brain images and then in the second phase also adding small patches of wrongly segmented regions from the first phase. More recently, Xue et al. [18] used a multi-path 2.5D dual U-Net using brain symmetry modality augmentation with a late fusion strategy on the ATLAS dataset of chronic stroke patients [19].

In this work, we present a deep learning approach for acute and sub-acute stroke lesion segmentation from multimodal MRI images. We use a 3D asymmetric encoder-decoder network based on the U-Net architecture with global and local residual connections. Within our approach, the class imbalance issue is alleviated with the use of small patches with balanced training patch sampling strategies and a dynamically weighted loss function. Additionally, we pre-process the provided images to facilitate using the symmetry property of brain hemispheres. In contrast to the work of Xue et al. [18] the symmetry features in our approach are fused earlier, before the network input, allowing the encoder to extract joint features between the original and symmetric modalities. The methodology is evaluated by cross-validation with the training images and with a blind online testing set evaluation against other state-of-the-art methods. The proposed approach demonstrates state of the art performance by ranking first in the testing leaderboard of both challenges [20] without any dataset specific tuning.

2. Data

For evaluation of the proposed methodology we use the public datasets provided for the two sub-tasks of the 2015 ISLES challenge [21]. They both encompass stroke lesion segmentation tasks from MRI imaging but using different imaging modalities and acquisition time since onset.

2.1. SISS dataset

For the sub-acute ischemic stroke segmentation (SISS) sub-task, a dataset was provided with 28 training and 36 testing cases acquired in the first week after onset [21]. The stroke MRI was performed on either a 1.5T (Siemens Magnetom Avanto) or 3T MRI system (Siemens Magnetom Trio). For each case, 4 co-registered multimodal images were provided including anatomical (T1, T2, FLAIR) and diffusion (DWI) MRI. The images were acquired as 3D volumes of 230 × 230 × 153 dimensions at 1 × 1 × 1 mm spacing. All four MRI modalities were used for evaluation of the proposed approach. For the training images, the provided gold standard, the whole lesion extent, was manually segmented by an experienced medical doctor.

2.2. SPES dataset

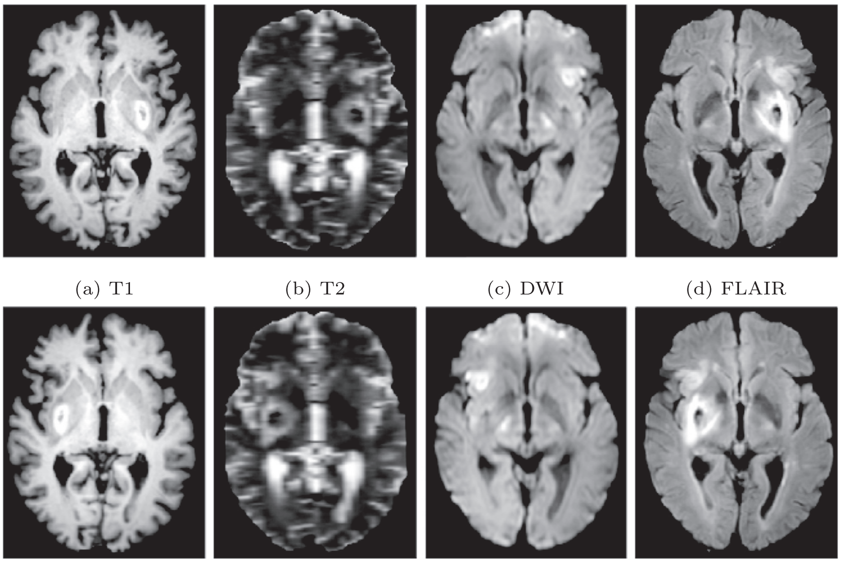
The acute stroke penumbra estimation sub-task (SPES) included 30 training and 20 testing cases acquired in the first day after onset with 3T Phillips systems on two centers [21]. For each case, 7 co-registered modalities were provided including anatomical (T1 contrast, T2), diffusion (DWI) and perfusion (CBF, CBV, TTP, Tmax) MRI. The images were acquired as 3D volumes of 96 × 110 × 71 dimensions at 2 × 2 × 2 mm spacing. All seven MRI modalities were used for evaluation of the proposed approach. For the training images, the gold standard segmentation, the penumbra label, was obtained as the mismatch between whole lesion extent and the core delineated in perfusion and diffusion images respectively.

3. Methodology

We propose a 3D patch based deep learning method using an asymmetrical residual CNN based on the U-Net architecture [15]. Within our approach, the class imbalance issue is addressed with a combination of techniques including the use of small patches (24 × 24 × 16) and a weighted loss function. We also regularize the training procedure with dropout [22], data augmentation and early stopping. For image segmentation, the use of whole patch predictions with a high degree of overlap minimizes the need for additional post-processing. In the following, we briefly describe the main components and implementation details of our methodology.

3.1. Data pre-processing

The given images are first pre-processed with a symmetric modality augmentation to allow learning of features based on the symmetry of brain hemispheres despite the small receptive field of the used patches. Explicit symmetry information was already shown to improve results for chronic stroke lesion segmentation [23]. In our case, instead of using one patch per hemisphere in a multi-path network we use a single joint patch with a single-path network. In practice, we augment the provided modality images with symmetric versions that swap the left and right hemispheres. We first flip one of the images along the mid-sagittal axis and then we apply FSL FLIRT [24] to perform a linear registration between the original and flipped image. Finally, the rest of modalities are



(a) T1 (b) T2 (c) DWI (d) FLAIR

(e) Sym. T1 (f) Sym. T2 (g) Sym. DWI (h) Sym. FLAIR

Fig. 1. Provided and symmetrically augmented modalities from case 2 of the SISS training images.

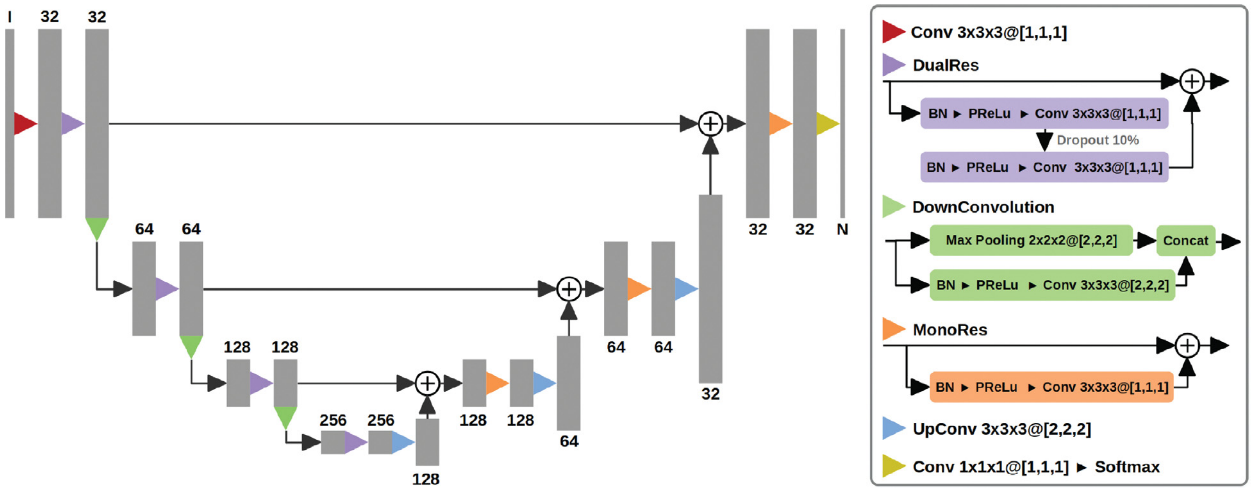

I
32
32
64
64
128
128
256
256
128
128
128
64
64
64
32
32
32
N
Conv 3x3x3@[1,1,1]
DualRes
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
Dropout 10%
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
DownConvolution
Max Pooling 2x2x2@[2,2,2]
Concat
BN ▸ PReLu ▸ Conv 3x3x3@[2,2,2]
MonoRes
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
UpConv 3x3x3@[2,2,2]
Conv 1x1x1@[1,1,1] ▸ Softmax

Fig. 2. Employed U-Net based architecture using 3D convolutions, 4 resolution steps and 32 base filters. The architecture consists of an asymmetrical encoder-decoder network using long and short residual connections. For the convolutional layers, $K_x \times K_y \times K_z@[S_x,S_y,S_z]$ indicates the kernel and stride dimensions in each axis. The number of channels is indicated above or under each feature map. In the input and output feature maps, I and N denote the number of image modalities and segmentation classes respectively.

registered using the same transformation. Fig. 1 shows an example of the resulting symmetrically augmented modalities. These are then appended to the provided ones, effectively doubling the number of images for each patient. In this way, a single extracted patch will also include intensity information from the opposite hemisphere.

3.2. CNN architecture

The used architecture, illustrated in Fig. 2, is a 3D asymmetrical encoder-decoder network based on the U-Net [15] architecture and its 3D extension, the 3D U-Net [25]. Additionally, we also use short and long residual connections as used by the 2D uResNet architecture [26] for chronic stroke in MRI. The asymmetry comes from the number of parameters found in the encoder and decoder branches, with 75% and 25% of the parameters respectively. We use residual blocks with two convolutional layers in the encoder and with a single convolutional for the decoder. It has been shown that the decoder's role is not as critical for segmentation, mainly upsampling the work of the encoder and fine-tuning the details [27]. Additionally, instead of the more typical rectified linear unit (ReLU) [28] we use in our residual blocks a parametric version, the PReLU non-linearity [29], as suggested by Paszke et al [27]. We perform downsampling in each resolution step by concatenating the result of a max pooling operation and strided convolution as proposed by Szegedy et al [30]. This strategy avoids representational bottlenecks while keeping the number of parameters contained. Finally, upsampling in the decoder branch is performed with the use of transposed convolutions.

3.3. Class imbalance handling

The class imbalance issue is caused by the typically smaller extent of the lesion class as compared with the rest of healthy tissue class. If no deliberate action is taken, the training set will be composed mostly from examples of healthy tissue and few from the lesion. This would induce a biased learning that would harm the segmentation performance. To alleviate this issue, we use a com-

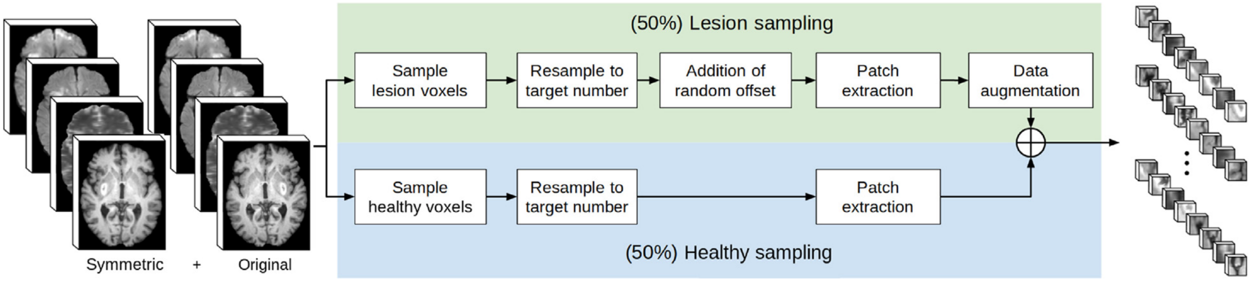


Fig. 3. Diagram of the used training patch sampling strategy for acute stroke related tasks that considers the anatomy and pathophysiology of stroke lesions. The extracted patches are of size 24 × 24 × 16 and include all input modalities.

bination of small patches with a balanced training patch sampling and a difficulty weighted loss function. The employed loss function, the Focal loss [31], is a dynamically weighted extension of the crossentropy loss defined as:

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1-p_\mathrm{t})^\gamma \log(p_\mathrm{t}) \tag{1}$$

where p_t and α_t are the predicted probability and weight for class t respectively. This function is dynamically weighted inversely proportional to the prediction confidence, so the network learns less from confident classifications and more from misclassified examples. In this way, class imbalance is alleviated as the network stops learning from the larger amount of healthy examples while still learning from the less common lesioned tissue. We use the Focal loss default parameters as suggested by Lin et al [31], with scaling factor $\gamma = 2$ and class weights $\alpha_0 = 0.25$ and $\alpha_1 = 0.75$ for the healthy and lesion classes respectively.

The use of patches allows using a training sampling strategy that can undersample the healthy class and oversample the lesion for a more balanced class representation. The employed strategy is an extension of two recently proposed ones for brain lesions [32] and chronic stroke [26]. In practice, a goal number of patches to extract is set per patient, as we aim to have a balanced patch representation of each case. Then, 50% of the training patches are extracted centered on voxels corresponding to healthy tissue and the other 50% on lesion. These are sampled at regular spatial steps to ensure that all parts of the brain are equally represented. The voxels sampled from the lesion class have a random offset added to increase representation of the region surrounding the lesion, the benign oligemia. As suggested by Guerrero et al [26], the offset is limited to half of the patch size to ensure the originally sampled voxel remains in the final extracted patch. For patients with smaller lesions, a combination of several patch extractions from the same lesion voxel and data augmentation is done to ensure the number is reached. The same sampled voxel will actually produce different patches since lesion voxels have a random offset applied. Finally, patches are extracted centered on these voxels. Additionally, for the lesion sampled patches, data augmentation is applied with five anatomically feasible operations including sagittal reflections and 90°, 180° and 270° axial rotations. A diagram summarizing the described strategy is depicted in Fig. 3.

Despite the balancing effects of the Focal loss and training patch sampling, the segmentation performance is still reduced when bigger patch sizes are considered. Since there are much fewer lesion voxels than healthy ones, bigger patches tend to include more healthy class voxels and further worsen class imbalance in the training set. The employed patch size of 24 × 24 × 16, determined empirically, offers the best compromise between receptive field and worsened imbalance for the considered datasets.

3.4. Network training

For training the randomly initialized network weights, we first extract patches to build the training and validation sets. As stated in Section 3.3, we use patches of size 24 × 24 × 16 sampled with a balanced patch sampling strategy. During training, we use the Focal loss [31] along with the Adadelta optimizer [33], to avoid costly grid search of a learning rate, with a batch size of 16 patches. This optimizer requires no manual tuning of parameters and appears robust to noisy gradient information, different model architecture choices, various data modalities and selection of hyper-parameters. Moreover, to prevent overfitting we use the early stopping technique by monitoring the performance on a validation set at the end of each epoch. In this way, the training is interrupted when the monitored metric reaches a local minimum, which means no more generalizable knowledge is being learned from the training images. The sum of the L1 loss and error rate on the validation set is used as the monitored metric with a patience of 8 epochs.

3.5. Segmentation and post-processing

Once the network weights have been trained, to segment a new volume patches are first extracted from every part of the image and forward passed through the network. These are sampled uniformly with a regular extraction step of 4 × 4 × 1 so that all parts of the brain are predicted. The resulting patch probabilities are then combined in a common space preserving their original spatial location to produce the whole volume probability map. In our case, the combination is performed per voxel by averaging the class probabilities of the various patches. Furthermore, some degree of overlap between the extracted patches is used since the extraction step is smaller than the patch size. Therefore, the same voxel is Doncs si no voleu sentir els meus laments, i us voleu posar taps, en trobareu a la nesspresso del pequatre.labeled seen in different neighborhoods and the resulting class probabilities are averaged. This technique reduces the need for post-processing steps as it provides coherently spatial labels without block artifacts.

Finally, the probability maps are binarized by thresholding the lesion class probabilities and then performing a connected component filtering by lesion volume. The variable threshold T_h can compensate over/under confident networks while the minimum lesion size S_{min}, measured in number of voxels, takes advantage of lesion priors to minimize false positives. In practice, the probability maps are binarized using the same threshold and minimum lesion size for each evaluation. These are found through grid search after all networks have been trained to offer the best compromise between the desired evaluation metrics.

3.6. Implementation details

The proposed method has been implemented with Python, using the Torch scientific computing framework [34]. All experiments

Table 1
Cross-validation experiment evaluation metrics on the SISS and SPES sub-tasks. The post-processing parameters T_h and $S_{\min}$ are found through grid search to maximize the score defined in Eq. (1). Significant differences of the reported metrics to the Baseline and Balanced approaches are marked with ▲ and ⋆ respectively.

Approach	T_h	$S_{\min}$	DSC	PPV	Sensitivity	HD
SISS sub-task						
Baseline	0.4	50	0.64 ± 0.22	0.69 ± 0.27	0.68 ± 0.21	43.7 ± 32.6
Balanced	0.4	200	0.67 ± 0.21	0.73 ± 0.22	0.69 ± 0.23	▲30.9 ± 28.9
Proposed	0.5	200	▲⋆ 0.71 ± 0.19	▲ 0.78 ± 0.20	0.67 ± 0.22	▲ 29.5 ± 29.5
SPES sub-task						
Baseline	0.6	500	0.80 ± 0.17	0.82 ± 0.21	0.82 ± 0.19	11.1 ± 6.9
Balanced	0.4	500	0.82 ± 0.15	0.84 ± 0.14	▲ 0.85 ± 0.17	12.4 ± 7.6
Proposed	0.5	200	▲ 0.82 ± 0.16	0.85 ± 0.13	▲ 0.85 ± 0.17	⋆ 11.2 ± 7.3

have been run on a GNU/Linux machine running Ubuntu 18.04 with 64GB of RAM memory and an Intel$\hat{\text{A}}$® Core$^{\text{TM}}$ i7-7800X CPU. The network training and testing has been done with an NVIDIA TITAN X GPU (NVIDIA corp, United States) with 12GB G5X memory.

4. Evaluation and results

We perform a quantitative and qualitative evaluation with both a cross-validation experiment and a blind external evaluation using the challenge web platform. The metrics used in the quantitative evaluations will be the ones provided by the online platform. These include the Dice similarity coefficient (DSC) [35], sensitivity, positive predictive value (PPV) and Hausdorff distance (HD). The DSC measures the relative overlap of the segmentation with the ground truth and is used as a measure of segmentation performance. The sensitivity and PPV measure different properties relative to the lesion class segmentation. On the one hand, the sensitivity is the fraction of gold standard lesion that has been correctly labeled as such in the resulting segmentation. On the other hand, the PPV measures the fraction of all predicted lesion labels that are correct. In this way, a segmentation that correctly labeled a small fraction of the lesion with few false positives would have a high PPV but low sensitivity. Finally, the Hausdorff distance can be intuitively seen as a measure of the *largest* border error between the segmentation and ground truth. To assess the statistical significance of differences between the crossvalidation results on the training set we consider the *t*-test for paired samples.

4.1. Cross-validation experiment with training images

The purpose of the cross-validation experiment is to quantitatively asses the main introduced improvements of the proposed methodology against a Baseline approach without them. For the Baseline approach, we use the proposed methodology without the class imbalance handling nor the data pre-processing step. Instead we use the crossentropy loss and training patch sampling as described in [32], using 24 × 24 × 16 patches without any addition of a random offset. We then evaluate the effects of a Balanced approach that only uses the class imbalance handling described in Section 3.3, without performing symmetric modality augmentation. Finally, the Proposed approach also adds the data pre-processing step to implement the complete proposed methodology.

Each evaluation is performed in 4 folds, adjusting the number of cases per fold accordingly, with the same training procedure for both the SISS and SPES datasets. To build the patch training set for each fold, 10 000 patches per case are extracted from the training images summing approximately 260 000 patches in total. Once the networks from each fold have been trained and the probability maps generated for all training images, the post-processing parameters T_h and $S_{\min}$ are found through grid search to optimize the desired metrics across all folds. We consider the range of thresholds T_h from 0.1 to 0.9 and minimum lesion size $S_{\min}$ from 10 to 1000 voxels. More specifically, we choose the parameter combination that jointly maximizes the average DSC and HD, the two metrics used to determine the 2015 ISLES workshop results. In practice, a combined score is computed as:

$$\text{Score} = \frac{\text{DSC} * \left(1 - \dfrac{\text{HD}}{\text{HD}_{\max}}\right)}{\text{DSC} + \left(1 - \dfrac{\text{HD}}{\text{HD}_{\max}}\right)} \tag{2}$$

where $\text{HD}_{\max}$, set to 200 voxels, is used to normalize the HD metric to the range between 0 and 1.

4.1.1. SISS sub-task results

The evaluation metrics of the cross-validation experiment using the SISS dataset can be found in the upper part of Table 1. With respect to the Baseline, the Balanced approach significantly improves the Hausdorff distance ($p < 0.01$) with marginal improvements in other metrics. When the symmetrically augmented modalities are further considered, the Proposed approach achieves significantly better DSC, PPV and HD ($p < 0.02$) as compared with the Baseline. However, despite the improvement in evaluation metrics, the Proposed approach needs a more restrictive minimum lesion size of 200 voxels to maximize the score as compared with the Baseline, which only filtered lesions smaller than 50 voxels. Additionally, Fig. 4 suggests that the performance of the network on the SISS dataset is independent of lesion size or count. It also shows a case where the lesion was completely missed due to its small size and unusual location at the cerebelum.

Representative examples of the qualitative results from the proposed method can be found in Fig. 5. Cases 9 and 15 represent the overall results of the proposed methodology, correctly detecting the lesions in most cases with an outline that approximates the provided gold standard. Among the observed limitations are inaccurate borders and over/under segmentation of certain regions. For instance, cases 13 and 17 are two of the worst results on the SISS dataset where false positive lesions are detected due to the existence of other chronic lesions with a similar appearance.

4.1.2. SPES sub-task results

The evaluation metrics of the cross-validation experiment using the SPES dataset can be found in the bottom part of Table 1. The class imbalance handling used in the Balanced approach significantly improves the sensitivity ($p < 0.01$) while providing marginal increase on the rest except the Hausdorff distance. When both improvements are simultaneously considered in the Proposed approach, it achieves a significantly better DSC and sensitivity ($p < 0.01$) than the Baseline. Additionally, the augmented modalities reduce the minimum lesion size $S_{\min}$ from 500 to a less restrictive 200 voxels. Additionally, Fig. 6 shows the performance on

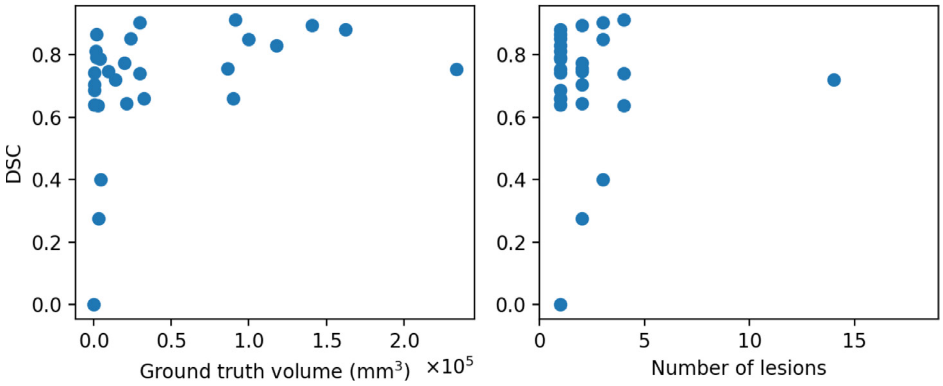


Fig. 4. SISS cross-validation results by lesion volume and count.

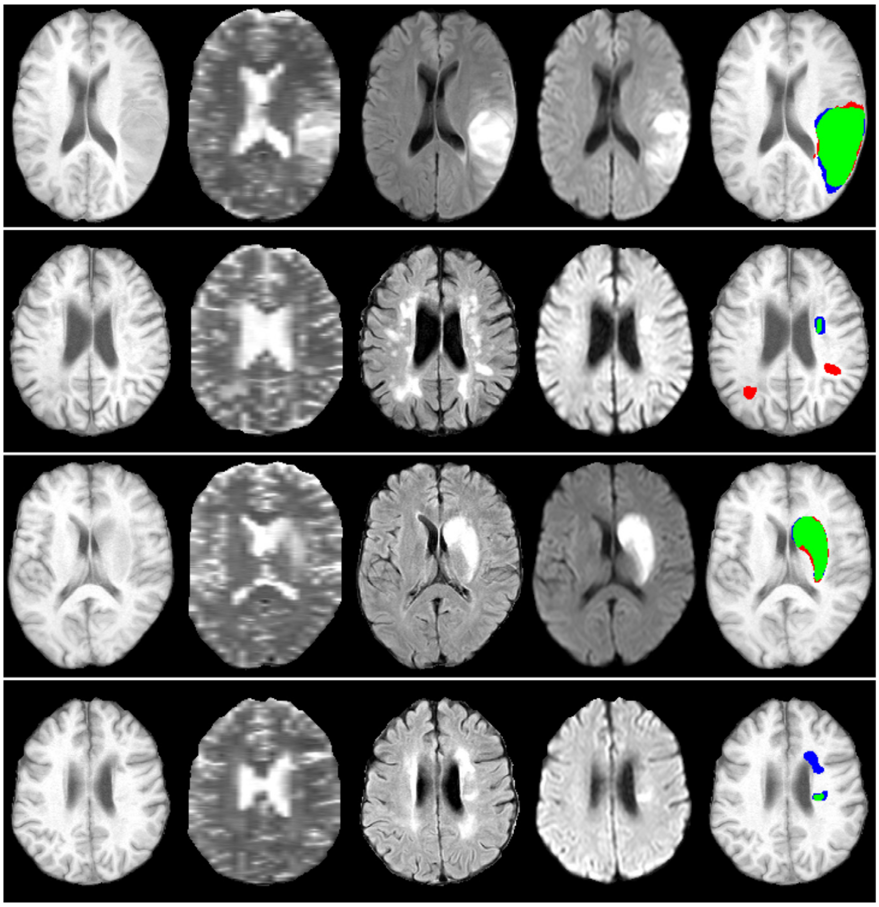


Fig. 5. Images and output masks of representative cases with good and bad performance (top to bottom, cases 9, 13, 15 and 17) from the training images of ISLES 2015 SISS dataset. From left to right, the displayed MR modalities are T1, T2, FLAIR, DWI and output segmentation. On all displayed segmentation results, true positives are denoted in green, false positives in red and false negatives in blue.

the SPES dataset with respect to the lesion size and count showing no apparent correlation between them.

Fig. 7 shows qualitative results of four representative segmentation examples from the proposed method. In general, the majority of the lesion is correctly segmented with minor border and small hole inaccuracies as seen in cases 2 and 11. Other less typical errors include under or oversegmentation of the lesion, as seen in case 15 where false positives are found on the upper part of the lesion. In the example of case 18, the lesion is clearly undersegmented probably due to a confounding unusual appearance of the TTP modality.

4.2. Blind challenge evaluation

To compare the proposed methodology against other state-of-the-art methods for acute stroke we submit our final approach for blind external evaluation in the ISLES 2015 challenge framework. The web platform used to hold the 2015 ISLES workshop [20] remains open for later submission and maintains an ongoing challenge leaderboard where the average testing set results are publicly displayed. Since the gold standard is hidden for the testing images, a fair and direct method comparison is possible. For evaluation in the challenge framework of ISLES 2015, we use the four networks

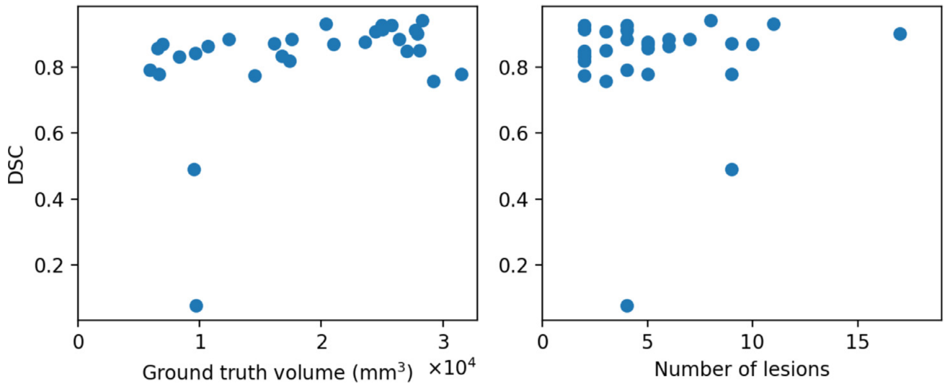


Fig. 6. SPES cross-validation results by lesion volume and count.

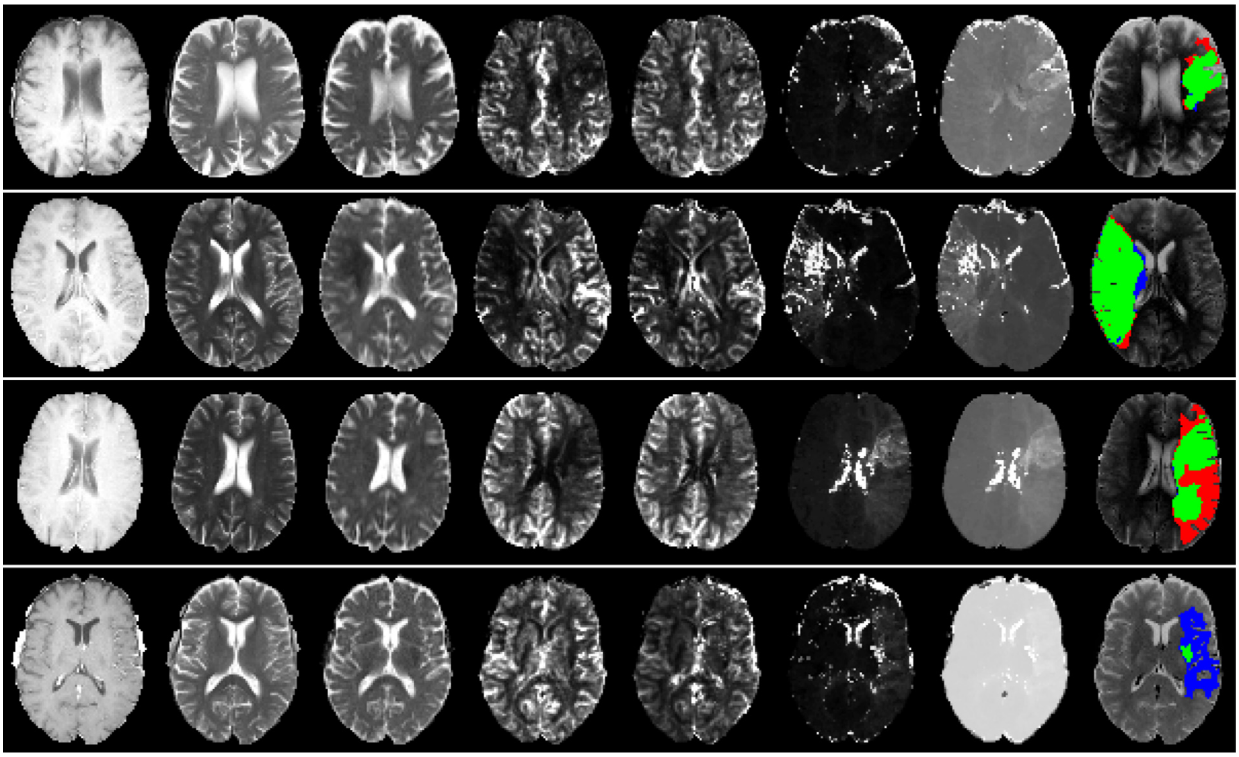


Fig. 7. Images and output masks of representative cases with good and bad performance (top to bottom, cases 2, 11, 15 and 18) from the training images of ISLES 2015 SPES dataset. From left to right, the displayed MR modalities are T1c, T2, DWI, CBF, CBV, TMax, TTP and output segmentation. On all displayed segmentation results, true positives are denoted in green, false positives in red and false negatives in blue.

trained for the Proposed approach during the cross-validation experiment, one from each fold, and average their outputs to produce a single testing patch prediction. The testing images are then segmented as described in Section 3.5 using the T_h and $S_{\min}$ set in the cross-validation experiment. In this way, the challenge results are produced with the same networks trained in the cross-validation experiment.

4.2.1. Challenge results

Tables 2 and 3 shows the top five entries as ranked by DSC of the ongoing testing leaderboard results for the SISS and SPES subtasks respectively. The proposed methodology achieves state-of-the-art performance in both sub-tasks, ranking first out of 74 entries in the SISS leaderboard and first out of 41 entries in the SPES leaderboard. As compared with the next best entries, we achieve similar or higher DSC with a 12% and 28% lower Haussdorf distance in the SISS and SPES sub-tasks respectively. Additionally, in the SPES dataset we also obtain an 8% higher sensitivity.

5. Discussion

We have performed both qualitative and quantitative evaluations of the proposed methodology in two different tasks without any dataset specific tuning of training hyper-parameters. The methodology has been shown to perform equally well for the acute or sub-acute stages and with different combinations of MRI modalities. The results are improved with respect to the Baseline thanks to the combined approach to alleviate data imbalance and also through the explicit learning of features based on the brain symmetry. Additionally, the method is fast in inference, taking under 3 min to pre-process and predict a new image.

The proposed methodology demonstrates state-of-the-art performance ranking 1st by average DSC while having a smaller HD as

Table 2
Top 5 out of 74 entries of the ongoing SISS testing leaderboard [20] as ranked by average DSC.

Rank	User	DSC	PPV	Sensitivity	HD
1	**clera2 (ours)**	**0.59 ± 0.31**	**0.65 ± 0.35**	**0.60 ± 0.30**	**34.7 ± 28.9**
2	kamnk1 [12]	0.59 ± 0.31	0.68 ± 0.33	0.60 ± 0.27	39.6 ± 30.7
3	zhanr6 [13]	0.58 ± 0.31	0.60 ± 0.33	0.68 ± 0.24	38.9 ± 35.3
4	lianl1	0.57 ± 0.29	0.58 ± 0.30	0.64 ± 0.29	43.0 ± 30.5
5	saliz1	0.57 ± 0.31	0.54 ± 0.31	0.67 ± 0.29	41.1 ± 36.7

Table 3
Top 5 entries out of 41 of the ongoing SPES testing leaderboard [20] as ranked by average DSC.

Rank	User	DSC	PPV	Sensitivity	HD
1	**clera2 (ours)**	**0.84 ± 0.10**	**0.82 ± 0.15**	**0.89 ± 0.06**	**20.7 ± 13.9**
2	mckir1 [8]	0.82 ± 0.08	0.83 ± 0.10	0.82 ± 0.14	29.0 ± 16.3
3	cheng5	0.81 ± 0.11	0.81 ± 0.12	0.81 ± 0.14	22.7 ± 12.6
4	maieo1 [7]	0.81 ± 0.09	0.84 ± 0.08	0.80 ± 0.14	23.6 ± 13.0
5	ghosp1	0.80 ± 0.11	0.80 ± 0.15	0.83 ± 0.11	57.1 ± 25.4

compared with the next best method in both challenges. Moreover, we are the first U-Net based approach in the online testing leaderboard to outperform the best 2015 ISLES workshop entries. In the SISS sub-task, we obtain a similar DSC but with lower HD than the next best method. In contrast with the approach by Kamnitsas et al [12], we can avoid the use of the additional post-processing step with conditional random fields that needs several image dependent configurable parameters. In our case, the use of a U-Net based architecture that provides whole patch predictions allows performing highly overlapped segmentations without a large increase of inference time or introducing additional configurable parameters. In the SPES sub-task, we obtain a higher sensitivity with a lower HD as compared with the next best method by McKinley et al [8] that used a random decision forest classifier with several hand-crafted features over 3 × 3 × 3 and 5 × 5 × 5 neighborhoods including local texture features, mean intensity, skewness, etc. By using a deep learning based method, the feature representation is learned at training time without having to rely on manually testing and finding the most appropriate ones for each specific task. Additionally, deep learning based methods can be integrated as part of a bigger diagnostic and prognostic processing pipelines where lesion or penumbra segmentation could be used as prior information. They can also benefit from related techniques such as continued training with newer images, fine tuning to improve performance as part of a composite pipeline, transfer learning to learn a similar unrelated task with few training examples [36] or performing domain adaptation between sites [37].

Despite the good relative performance, the qualitative results show that the proposed methodology is still limited by inaccurate borders, missing lesion parts and other confounding factors. Furthermore, while the found minimum lesion size maximize the desired metrics along all training images they might still filter out some small lesions at testing time.

6. Conclusions

In this work, we have presented a methodology that achieves state-of-the-art performance in two different stroke lesion segmentation tasks. To the best of our knowledge, the proposed methodology is the first to obtain competitive results in both the ISLES 2015 SISS and SPES sub-tasks with the same approach. We have achieved these results by doing both regularization of the training procedure and providing additional meaningful information for lesion segmentation. Useful features using the brain symmetry could not be learned as the employed patch size is too small to include both hemispheres. The proposed symmetric modality augmentation facilitates using the similarity between hemispheres to improve lesion localization without using larger patches that would worsen class imbalance. Moreover, we have shown the big influence class imbalance can have in reducing distant outliers and false positives that provide a lower Hausdorff distance at testing time. By using a combined approach we achieve a less biased segmentation with a better balance between sensitivity and specificity.

In the clinical setting, treatment decisions for ischemic stroke patients need to be fast, as the ischemic brain ages the equivalent of 3.6 years each hour without treatment [38], and well justified given the risk of complications involved in surgical interventions. Accurate segmentation of the lesion and/or penumbra from MR images can provide a fast quantitative estimate of the extent and location of the penumbra region, tissue that can be recovered if flow is restored. This estimate could be used to assess if the treatment risks outweigh the potential benefits and allow for faster and better informed decisions. Furthermore, these kind of lesion segmentation methods can be used to carry out correlation studies between lesion location and chronic disability status which would make this estimate even more informative in the treatment decision context. The proposed methodology is made publicly available for the scientific community [39].

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgements

Albert Clèrigues holds an FPI grant from the Ministerio de Ciencia, Innovación y Universidades with reference number PRE2018-083507. This work has been partially supported by Retos de Investigación TIN2015-73563-JIN and DPI2017-86696-R from the Ministerio de Ciencia, Innovación y Universidades. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research.

References

[1] J. Redon, M.H. Olsen, R.S. Cooper, O. Zurriaga, M.A. Martinez-Beneito, S. Laurent, R. Cifkova, A. Coca, G. Mancia, Stroke mortality and trends from 1990 to 2006 in 39 countries from Europe and Central Asia: implications for control of high blood pressure, Eur. Heart J. 32 (11) (2011) 1424–1431, doi:10.1093/eurheartj/ehr045.

[2] I. Rekik, S. Allassonnière, T.K. Carpenter, J.M. Wardlaw, Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal, NeuroImage 1 (1) (2012) 164–178, doi:10.1016/J.NICL.2012.10.003.

[3] S.A. Sheth, R. Jahan, J. Gralla, V.M. Pereira, R.G. Nogueira, E.I. Levy, O.O. Zaidat, J.L. Saver, Time to endovascular reperfusion and degree of disability in acute stroke, Ann. Neurol. 78 (4) (2015) 584–593, doi:10.1002/ana.24474.

[4] B.C. Campbell, P.J. Mitchell, L. Churilov, M. Keshtkaran, K.-S. Hong, T.J. Kleinig, H.M. Dewey, N. Yassi, B. Yan, R.J. Dowling, Endovascular thrombectomy for ischemic stroke increases disability-free survival, quality of life, and life expectancy and reduces cost, Front. Neurol. 8 (2017) 657, doi:10.3389/fneur.2017.00657.

[5] R.K. Singh, V.A. Chafale, R.S. Lalla, K.C. Panchal, A.P. Karapurkar, S.V. Khadilkar, P.K. Ojha, Y. Godge, R.K. Singh, R. Benny, Acute ischemic stroke treatment using mechanical thrombectomy: a study of 137 patients., Ann. Indian Acad. Neurol. 20 (3) (2017) 211–216, doi:10.4103/aian.AIAN_158_17.

[6] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, IEEE Comput. Soc. Press, 1995, pp. 278–282, doi:10.1109/ICDAR.1995.598994.

[7] O. Maier, M. Wilms, H. Handels, Random forests for acute stroke penumbra estimation, in: Proceedings of ISLES 2015 challenge, 2015, pp. 77–80.

[8] R. Mckinley, L. Häni, R. Wiest, M. Reyes, Segmenting the ischemic penumbra: a spatial random forest approach with automatic threshold finding, in: Proceedings of ISLES 2015 Challenge, 2015, pp. 69–73.

[9] D. Robben, D. Christiaens, J.R. Rangarajan, J. Gelderblom, P. Joris, F. Maes, P. Suetens, O. Maier, M. Reyes, D. Robben, D. Christiaens, J. Raman Rangarajan, J. Gelderblom, P. Joris, F. Maes, P. Suetens, ISLES Challenge 2015: a voxel-wise, cascaded classification approach to stroke lesion segmentation, in: Proceedings of ISLES 2015 Challenge, Springer, Cham, 2015, pp. 27–31, doi:10.1007/978-3-319-30858-6_22.

[10] O. Maier, C. Schröder, N.D. Forkert, T. Martinetz, H. Handels, Classifiers for ischemic stroke lesion segmentation: a comparison study, PLoS ONE 10 (12) (2015) e0145118, doi:10.1371/journal.pone.0145118.

[11] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, doi:10.1162/neco.1989.1.4.541.

[12] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, B. Glocker, Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI, in: Proceedings of ISLES 2015 Challenge, 2015, pp. 13–17.

[13] R. Zhang, L. Zhao, W. Lou, J.M. Abrigo, V.C.T. Mok, W.C.W. Chu, D. Wang, L. Shi, Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional densenets, IEEE Trans. Med. Imaging 37 (9) (2018) 2149–2160, doi:10.1109/TMI.2018.2821244.

[14] R. Karthik, U. Gupta, A. Jha, R. Rajalakshmi, R. Menaka, A deep supervised approach for ischemic lesion segmentation from multimodal MRI using fully convolutional network, Appl. Soft Comput. J. 84 (2019) 105685, doi:10.1016/j.asoc.2019.105685.

[15] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[16] S. Winzeck, A. Hakim, R. McKinley, J.A. Pinto, V. Alves, C. Silva, M. Pisov, E. Krivov, M. Belyaev, M. Monteiro, A. Oliveira, Y. Choi, M.C. Paik, Y. Kwon, H. Lee, B.J. Kim, J.-H. Won, M. Islam, H. Ren, D. Robben, P. Suetens, E. Gong, Y. Niu, J. Xu, J.M. Pauly, C. Lucas, M.P. Heinrich, L.C. Rivera, L.S. Castillo, L.A. Daza, A.L. Beers, P. Arbelaezs, O. Maier, K. Chang, J.M. Brown, J. Kalpathy-Cramer, G. Zaharchuk, R. Wiest, M. Reyes, ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI, Front. Neurol. 9 (2018), doi:10.3389/fneur.2018.00679.

[17] A. Olivier, O. Moal, B. Moal, F. Munsch, G. Okubo, I. Sibon, V. Dousset, T. Tourdias, Active learning strategy and hybrid training for infarct segmentation on diffusion MRI with a U-shaped network, J. Med. Imaging 6 (4) (2019) 1, doi:10.1117/1.jmi.6.4.044001.

[18] Y. Xue, F.G. Farhat, O. Boukrina, A.M. Barrett, J.R. Binder, U.W. Roshan, W.W. Graves, A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images, NeuroImage 25 (2020) 102118, doi:10.1016/j.nicl.2019.102118.

[19] S.L. Liew, J.M. Anglin, N.W. Banks, M. Sondag, K.L. Ito, H. Kim, J. Chan, J. Ito, C. Jung, N. Khoshab, S. Lefebvre, W. Nakamura, D. Saldana, A. Schmiesing, C. Tran, D. Vo, T. Ard, P. Heydari, B. Kim, L. Aziz-Zadeh, S.C. Cramer, J. Liu, S. Soekadar, J.E. Nordvik, L.T. Westlye, J. Wang, C. Winstein, C. Yu, L. Ai, B. Koo, R.C. Craddock, M. Milham, M. Lakich, A. Pienta, A. Stroud, A large, open source dataset of stroke anatomical brain images and manual lesion segmentations, Sci. Data 5 (1) (2018) 1–11, doi:10.1038/sdata.2018.11.

[20] ISLES 2015 - SICAS Medical Image Repository, Leaderboard: SISS/SPES Testing., 2019(accessed: 3 April 2019).

[21] O. Maier, B.H. Menze, J. von der Gablentz, L. Häni, M.P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, D. Christiaens, F. Dutil, K. Egger, C. Feng, B. Glocker, M. Götz, T. Haeck, H.-L. Halme, M. Havaei, K.M. Iftekharuddin, P.-M. Jodoin, K. Kamnitsas, E. Kellner, A. Korvenoja, H. Larochelle, C. Ledig, J.-H. Lee, F. Maes, Q. Mahmood, K.H. Maier-Hein, R. McKinley, J. Muschelli, C. Pal, L. Pei, J.R. Rangarajan, S.M. Reza, D. Robben, D. Rueckert, E. Salli, P. Suetens, C.-W. Wang, M. Wilms, J.S. Kirschke, U.M. Krämer, T.F. Münte, P. Schramm, R. Wiest, H. Handels, M. Reyes, ISLES 2015 - A Public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI, Med. Image Anal. 35 (2017) 250–269, doi:10.1016/J.MEDIA.2016.07.009.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958, doi:10.1214/12-AOS1000.

[23] Y. Wang, A.K. Katsaggelos, X. Wang, T.B. Parrish, A deep symmetry convnet for stroke lesion segmentation, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 111–115, doi:10.1109/ICIP.2016.7532329.

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

[25] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2016, pp. 424–432, doi:10.1007/978-3-319-46723-8_49.

[26] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. Valdés-Hernández, D. Dickie, J. Wardlaw, D. Rueckert, White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks, NeuroImage 17 (2018) 918–934, doi:10.1016/J.NICL.2017.12.022.

[27] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, ENet: a deep neural network architecture for real-Time semantic segmentation, arXiv:1606.02147 (2016).

[28] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

[29] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2818–2826, doi:10.1109/CVPR.2016.308.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988, doi:10.1109/TPAMI.2018.2858826.

[32] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78, doi:10.1016/J.MEDIA.2016.10.004.

[33] M.D. Zeiler, ADADELTA: An adaptive learning rate method, arXiv:1212.5701 (2012). 10.1145/1830483.1830503

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, Neural Information Processing Systems, 2017.

[35] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302, doi:10.2307/1932409.

[36] S. Pang, Z. Yu, M.A. Orgun, A novel end-to-end classifier using domain transferred deep convolutional neural networks for biomedical images, Comput. Methods Programs Biomed. 140 (2017) 283–293, doi:10.1016/J.CMPB.2016.12.019.

[37] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J.C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, X. Lladó, One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks, NeuroImage 21 (2019) 101638, doi:10.1016/J.NICL.2018.101638.

[38] J.L. Saver, Time is brain–quantified, Stroke 37 (1) (2006) 263–266, doi:10.1161/01.STR.0000196957.55928.ab.

[39] Github, NIC-VICOROB/stroke-mri-segmentation, 2019, (accessed: 3 April 2019).

Chapter 5

Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry

In this chapter, we study the effect of WM lesions within our deep learning based framework for tissue segmentation and propose a data generation and training technique which learns from a reference method while embedding WM lesion effect reduction within the model itself. Typically, to avoid the effect of WM lesions from influencing classical brain tissue segmentation methods, the lesion is first segmented and then inpainted with normally appearing white matter intensities. Most works in the literature first evaluate the inpainting quality with appearance based metrics and then proceed to evaluate if the volumes measured by state-of-the-art tissue segmentation methods are less affected thanks to the inpainting. In our approach, we jointly optimize the tasks of lesion inpainting and tissue segmentation end-to-end within our deep learning system, which yields an inpainting model which is also optimized for the downstream segmentation task. This effectively couples both tasks and allows to obtain much lower errors on the measured tissue volumes.

Paper published in **Computerized Medical Imaging and Graphics** (CMIG)
Volume: 103, Published: January 2023
DOI: https://doi.org/10.1016/j.compmedimag.2022.102157
JCR RNMMI IF 7.422, Q1(14/314)

Computerized Medical Imaging and Graphics 103 (2023) 102157

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry

Albert Clèrigues [a,*], Sergi Valverde [b], Joaquim Salvi [a], Arnau Oliver [a], Xavier Lladó [a]

[a] Institute of Computer Vision and Robotics, University of Girona, Spain
[b] Tensor Medical, Girona, Spain

ARTICLE INFO

Keywords:
Magnetic resonance imaging
Brain tissue segmentation
White matter lesions
Lesion inpainting
Deep learning
Brain volumetry

ABSTRACT

Automated methods for segmentation-based brain volumetry may be confounded by the presence of white matter (WM) lesions, which introduce abnormal intensities that can alter the classification of not only neighboring but also distant brain tissue. These lesions are common in pathologies where brain volumetry is also an important prognostic marker, such as in multiple sclerosis (MS), and thus reducing their effects is critical for improving volumetric accuracy and reliability. In this work, we analyze the effect of WM lesions on deep learning based brain tissue segmentation methods for brain volumetry and introduce techniques to reduce the error these lesions produce on the measured volumes. We propose a 3D patch-based deep learning framework for brain tissue segmentation which is trained on the outputs of a reference classical method. To deal more robustly with pathological cases having WM lesions, we use a combination of small patches and a percentile-based input normalization. To minimize the effect of WM lesions, we also propose a multi-task double U-Net architecture performing end-to-end inpainting and segmentation, along with a training data generation procedure. In the evaluation, we first analyze the error introduced by artificial WM lesions on our framework as well as in the reference segmentation method without the use of lesion inpainting techniques. To the best of our knowledge, this is the first analysis of WM lesion effect on a deep learning based tissue segmentation approach for brain volumetry. The proposed framework shows a significantly smaller and more localized error introduced by WM lesions than the reference segmentation method, that displays much larger global differences. We also evaluated the proposed lesion effect minimization technique by comparing the measured volumes before and after introducing artificial WM lesions to healthy images. The proposed approach performing end-to-end inpainting and segmentation effectively reduces the error introduced by small and large WM lesions in the resulting volumetry, obtaining absolute volume differences of 0.01 ± 0.03% for GM and 0.02 ± 0.04% for WM. Increasing the accuracy and reliability of automated brain volumetry methods will reduce the sample size needed to establish meaningful correlations in clinical studies and allow its use in individualized assessments as a diagnostic and prognostic marker for neurodegenerative pathologies.

1. Introduction

Global and regional volumetry of the brain parenchyma is a promising biomarker that can improve prognosis for multiple sclerosis (MS) patients (Bendfeldt et al., 2009; Lansley et al., 2013; Pérez-Miralles et al., 2013). Brain volume loss has been shown to be a predictor of disease progression and disability status in MS patients (Di Filippo et al., 2010; Ghione et al., 2020). Moreover, the rate of brain volume loss is also used to evaluate the effectiveness of disease-modifying treatments in clinical studies as well as for individualized treatment response assessment (Sotirchos et al., 2020; Cortese et al., 2022). Magnetic resonance (MR) imaging offers a noninvasive way to perform indirect volume measurements on the brain parenchyma and its distinct cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM) components. In non-uniformity corrected T1-w MR images, these tissues are characterized by normally distributed intensity profiles with different means and variances. However, a characteristic of brain scans from MS patients is the presence of WM lesions appearing as a fourth intensity distribution that intersects with the brain tissue intensities to be measured. The presence of WM lesions can bias the characterization of

[*] Correspondence to: University of Girona, Ed. P-IV, Campus Montilivi, 17003 Girona, Spain.
E-mail address: albert.clerigues@udg.edu (A. Clèrigues).

https://doi.org/10.1016/j.compmedimag.2022.102157
Received 27 May 2022; Received in revised form 2 December 2022; Accepted 2 December 2022
Available online 13 December 2022

normal-appearing tissue intensities and interfere with brain tissue quantification methods (González-Villà et al., 2017). The error that WM lesions introduce is highly dependent on their aspect and size (Battaglini et al., 2012), which change over time, introducing varying levels of error in images taken at different timepoints. These lesions can especially affect the estimation of partial volumes found in the interfaces between brain tissues and have been observed to produce a boundary shifting effect (Magon et al., 2014). Reducing the error introduced by WM lesions in brain tissue segmentation is critical for improving the reliability and accuracy of cross-sectional and longitudinal brain volumetry methods.

Techniques that minimize the effect of WM lesions in brain tissue segmentation usually involve lesion inpainting as a preliminary step before segmentation. These techniques fill the lesioned voxels with intensities resembling the normal-appearing WM (NAWM) of that image. Chard et al. (2010) proposed the use of a Gaussian mixture model to characterize and sample the intensity distribution of NAWM in the whole image to fill lesioned voxels while also emulating scanner noise and nonuniformity. Battaglini et al. (2012) and Magon et al. (2014) both proposed similar local inpainting methods for preliminary brain tissue segmentation to then fill the lesion voxels with intensities similar to those of the NAWM adjacent to the lesioned voxels. Similarly, Valverde et al. (2014) also used preliminary tissue segmentation to characterize and sample the NAWM intensity distribution but did so on a slice-by-slice basis. Prados et al. (2016) proposed a non-local mean patch-based inpainting method that can work with longitudinal data and for any MR modality. More recently, data-driven methods using deep learning have been proposed based on the use of convolutional neural networks (CNNs) for lesion inpainting. Armanious et al. (2019) used a 2D conditional generative adversarial network (cGAN) to synthesize realistic looking intensities for a square patch removed from the input slice. Xiong et al. (2020) used 2D U-Net with a nonlesion attention module to inpaint lesioned voxels, while Zhang et al. (2020) proposed the use of 2D U-Net with edge priors as additional input to improve the inpainting quality. Manjón et al. (2020) proposed a 3D blind inpainting method that automatically inpaints any abnormal-looking voxels without requiring prior lesion segmentation, unlike the methods described previously that require a preliminary WM lesion mask. Tang et al. (2021) proposed an inpainting approach for MS lesions using dynamic learnable gate masks to improve the morphological and textural consistency of inpainted regions and reduce their effect on subsequent brain tissue segmentation. Most works cited above have been shown to improve the results for brain tissue segmentation methods by reducing segmentation differences between healthy and artificially lesioned image pairs. However, to the best of our knowledge, the recent deep learning-based brain tissue segmentation approaches (Rajchl et al., 2018; Guha Roy et al., 2019; Henschel et al., 2020) have not evaluated the effect of WM lesions.

In this work, we propose a 3D patch-based deep learning tissue segmentation framework for brain volumetry which learns from the outputs of a reference classical brain tissue segmentation method. In our approach, we improve the robustness on pathological cases having WM lesions by using small patches and a percentile-based input normalization. To further minimize the effect of WM lesions, we also propose the use of a multi-task double U-Net architecture performing end-to-end inpainting and segmentation. To train the proposed method as well as to evaluate the WM lesion effect, we use pairs of lesioned and non-lesioned versions of the same brain image. Since these pairs of images cannot be naturally obtained, artificial lesions are introduced into a set of scans from healthy subjects to obtain both versions of the same image. Our goal is to learn a segmentation model that can minimize the effect of WM lesions on the rest of the normal-appearing tissue in the image. During training, we use the artificially lesioned brain images as input and target the brain tissue probabilities of their originally healthy counterpart image as output. In this way, the system is trained to minimize the impact of WM lesion voxels on the segmentation of neighboring healthy tissue. In the proposed method, a preliminary WM lesion mask is used to occlude the lesioned voxels of the input patch by masking it with zeros. Then, a double chained U-Net architecture is used, where the first network inpaints the occluded lesion voxels and the second performs brain tissue segmentation from the inpainted patch. Both networks are trained end-to-end so that the inpainter network is also trained to aid in the segmentation task.

We evaluate the effect of WM lesions on our deep learning framework as well as on FAST (Zhang et al., 2001), the brain tissue segmentation method used to generate the training targets, which is implemented in the FSL package of analysis tools for structural MR brain imaging data. In the evaluation, we quantify the tissue volume differences between healthy and artificially lesioned versions of the same image for each of the considered tissue segmentation methods. Without performing lesion inpainting, our deep learning framework already shows significantly smaller and more localized volume differences due to the presence of WM lesions than the reference method. We then evaluate the extent to which the lesion effect minimization techniques reduce the error introduced on the measured tissue volumes. The FSL package also provides a WM lesion inpainting method (Battaglini et al., 2012), which is typically used along with FAST (Zhang et al., 2001). The FSL pipeline doing WM lesion inpainting and brain tissue segmentation is used as a baseline to compare against our deep learning approach. Additionally, we also compare against the case where we first inpaint the WM lesions with the FSL method and then perform the brain tissue segmentation with our deep learning approach. The proposed method doing end-to-end inpainting and tissue segmentation is faster and obtains significantly lower volume differences, especially when considering larger WM lesions. Even when the FSL_inpainting method is used to preprocess the image, our deep learning based tissue segmentation model still achieves significantly lower error and better performance on large WM lesions than the FSL pipeline. Thanks to the use of data-driven techniques, we are able to learn from a reference method while minimizing the WM lesion effect on the measured tissue volumes to almost negligible levels. The development framework is available to the research community at https://github.com/NIC-VICOROB/LITS.

2. Materials

Two different kinds of image datasets are used to train and evaluate the proposed method, healthy brain scans and lesioned brain scans with manually delineated WM lesion masks comprising small and large lesions from patients with multiple sclerosis (MS) and other pathologies. These brain images are used to generate artificially lesioned and healthy image pairs for training and evaluation. The location and morphology of artificial WM lesions introduced in the T1-w healthy images are taken from the WM lesion masks of lesioned brain scans, while their appearance is simulated by sampling intensities between the means of GM and WM tissue, similar to the work of Battaglini et al. (2012).

2.1. Healthy brain dataset

Calgary-Campinas Public Brain MR Dataset (Souza et al., 2018). This dataset is composed of 359 T1-weighted brain scans from 359 healthy adults with an average age of 53.5 ± 7.8 years, ranging between 29 and 80 years. Images were acquired on scanners from three vendors (GE, Philips, and Siemens) at two different magnetic field strengths of 1.5 T and 3 T, approximately 60 scans were obtained per vendor. Most scans in this dataset have a voxel size of $1.0 \times 1.0 \times 1.0$ mm^3 except for sixty scans acquired at $0.89 \times 0.89 \times 0.89$ mm^3 and another sixty acquired at $1.33 \times 1.0 \times 1.0$ mm^3. The dataset also includes silver standard brain masks generated through a consensus of several state-of-the-art automatic skull stripping methods.

2.2. Lesioned brain datasets

MSSEG Challenge (Commowick et al., 2018). The MSSEG Challenge hosted at the MICCAI 2016 international conference provided a multicentric database for training consisting of 15 multimodal (T1-w, T1-w gadolinium, T2-w, FLAIR and PD) MR images obtained from MS patients with an average lesion load of 20.8 ± 19.9 ml. Images were acquired on three different scanners at different voxel sizes: five images from a Philips Ingenia 3 T scanner at 0.7 × 0.74 × 0.74 mm^3, five images from a Siemens Verio 3 T scanner at 1.1 × 0.5 × 0.5 mm^3 and the remaining five images from a Siemens Aera 1.5 T scanner at 1.25 × 1.03 × 1.03 mm^3. The MR images were rigidly coregistered to the FLAIR scan, which was manually annotated by 7 independent experts, and a consensus gold standard WM lesion segmentation approach was built.

ISBI 2015 Longitudinal MS Lesion Segmentation Challenge (Carass et al., 2017). This challenge provided a multimodal (T1-w, T2-w, FLAIR and PD) training dataset with 21 longitudinal scans from five MS patients with an average lesion load of 11.6 ± 10.5 ml. Images were acquired on a 3 T MRI Philips scanner with a voxel size of 0.82 × 0.82 × 1.17 mm^3. Manual delineations were made by two experts identifying and segmenting white matter lesions on the MR images. The MR images from each subject as well as the expert WM lesion delineations were rigidly coregistered to the T1-w scan.

WMH Challenge 2017 (Kuijf et al., 2019). The training set provided 60 sets of brain MR images (3D T1 and 2D multislice FLAIR) from 60 subjects of two memory clinics showing cognitive impairment of presumed vascular origin with an average lesion load of 17.5 ± 17.1 ml. Images were taken with five different 3 T MR scanners from three different vendors (Siemens, Philips and GE) with voxel sizes of 1.0 × 1.0 × 1.0 mm^3 and 0.94 × 0.94 × 1.0 mm^3. The FLAIR scans from each subject were resampled and coregistered to the 3D T1 scan via an affine transform. The provided gold standard was made with manual annotations of white matter hyperintensities (WMHs) made by experts in accordance with the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE) criteria (Wardlaw et al., 2013).

2.3. Preprocessing

In the image preprocessing stage, we generate the healthy and lesioned image pairs that are used for training and evaluation. The location and morphology of artificial WM lesions are obtained from WM lesion masks of the three lesioned brain datasets that are registered to the healthy dataset scans. In practice, all the available WM lesion masks from lesioned datasets are registered to each of the T1-w healthy images, allowing the generation of several artificially lesioned scans from a single healthy scan. The registered WM masks are then used to generate artificial lesions in the healthy T1-w brain scans with the intensities located within the GM/WM interface. The preprocessing steps are explained in detail in the following sections.

2.3.1. Skull-stripping

The healthy scans belonging to the Calgary–Campinas dataset images need to be skull-stripped before segmenting with FAST to consider only the intensities corresponding to the intracraneal cavity. For this, we use the provided silver brain masks, which are applied to generate the skull-stripped images. For the lesioned brain datasets, two of them (MICCAI 2016 MS lesion segmentation challenge and ISBI 2015 Longitudinal MS Lesion Segmentation Challenge) were already skull-stripped, while the WMH Challenge 2017 dataset is processed using ROBEX (Iglesias et al., 2011) on the T1-w images.

2.3.2. Lesion mask registration

In this step, all the available lesioned scans are linearly registered to each of the healthy images, obtaining several artificial WM lesion mask instances in the space of each healthy scan. This process is performed independently for the training and evaluation image sets. To avoid performing a large number of registrations, we first register all the healthy and lesioned images to a common space and then combine these transforms to obtain the desired transforms. Linear affine registration is performed with the skull-stripped T1-w images from both healthy and pathological datasets to the MNI ICBM 152 nonlinear 6th Generation Symmetric Average Brain template using FSL FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002) with default parameters. This results in a linear transform matrix T(*I,MNI*) for each image *I*, which can also be inverted to obtain T(*MNI,I*). Then, for any pair of healthy *H* and lesioned *L* images, we can compute T(*L,H*) using the previously computed transforms to the MNI as follows:

$$\mathrm{T}(L,H) = \mathrm{T}(L,MNI) \circ \mathrm{T}(MNI,H) \quad (1)$$

T(*L,H*) is computed for each lesioned and healthy control image pair and then applied to the binary WM lesion mask using nearest neighbor interpolation. Finally, we ensure that the registered lesions are introduced only to the WM of healthy images. For this, we use FAST (Zhang et al., 2001) to obtain a binary WM mask for each healthy image and keep only the voxels from registered lesion masks that are also classified as WM in the healthy image.

2.3.3. Artificial WM lesions

The registered WM lesion masks are then used to generate several artificially lesioned images from each healthy image. The artificial lesion intensities are filled as in the work of Battaglini et al (Battaglini et al., 2012)., which presented and evaluated the lesion inpainting method we use as a baseline. In their approach, a preliminary FAST (Zhang et al., 2001) tissue segmentation is used to estimate the mean intensities of GM and WM and is then used to generate the intensity distribution for artificial lesions. These are then filled with intensities between the normally appearing GM and WM, with a mean equal to the average of the GM and WM means and a standard deviation equal to a fourth of the interval between the GM and WM means (Battaglini et al., 2012).

During training and inference of the proposed methodology, the artificial lesion intensities are effectively ignored as they are occluded by filling them with zeros. Hence, the intensities of artificial lesions are only useful for evaluating the WM lesion effect of tissue segmentation methods when no inpainting is used.

3. Methods

The proposed deep learning brain tissue segmentation framework consists of a 3D patch-based approach which learns from the outputs of FAST (Zhang et al., 2001), an automatic brain tissue segmentation method implemented in the FSL package. The backbone of our framework consists of a 3D network, depicted in Fig. 1, which is derived from the U-Net architecture (Ronneberger et al., 2015) and uses residual convolution blocks and skip connections. The convolutional layers use 3 × 3 × 3 kernels and are always preceded, except for the input and output nodes, by a batch normalization (BN) layer (Ioffe and Szegedy, 2015) and a parametric rectified linear unit (PReLu) activation (Nair and Hinton, 2010). The parameter distribution of the model is asymmetrical with respect to the residual blocks of the encoder using two convolutional layers, while a single layer is used in the decoder. The network has 4 resolution levels where the feature maps are downsampled by 2 × 2 × 2 in each level of the encoder and upsampled by the same factor in the decoder. Downsampling is performed by concatenating the result of a max pooling operation and strided convolution as proposed by Szegedy et al (Szegedy et al., 2016)., while upsampling is performed with a transposed convolution that learns the upsampling operator for each feature map.

Within our patch-based deep learning framework, the introduction of a WM lesion in a healthy brain scan produces segmentation differences at both global and local levels. Global differences appear when the modification of a small part of the input has an effect on the output

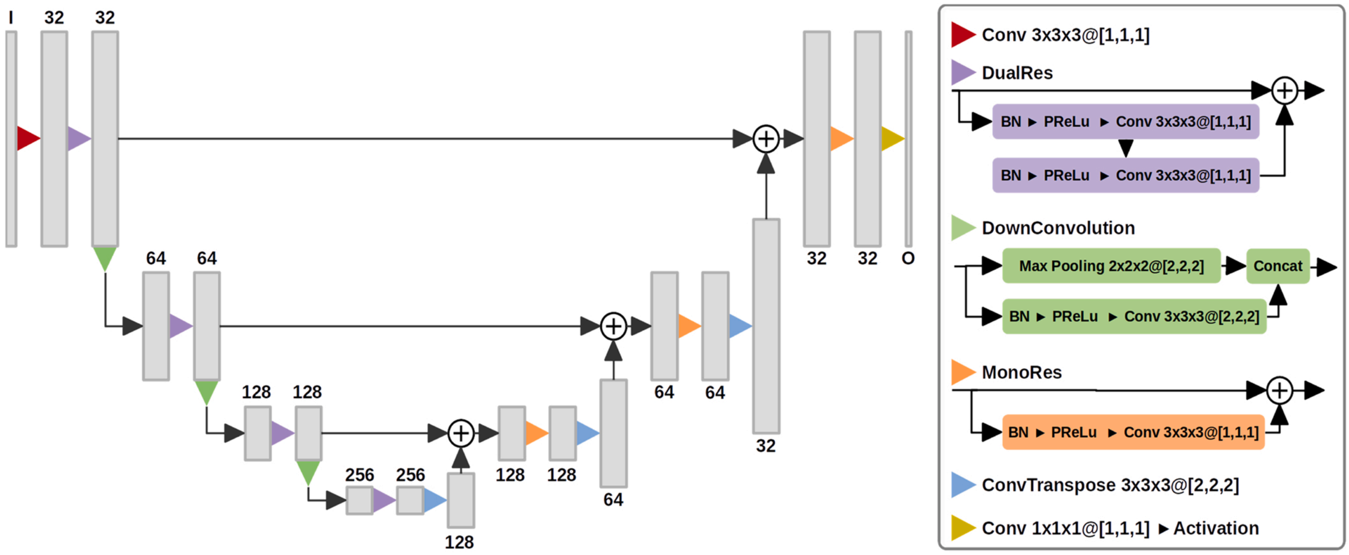

I
32
32
64
64
128
128
256
256
128
128
128
64
64
64
32
32
32
O
Conv 3x3x3@[1,1,1]
DualRes
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
DownConvolution
Max Pooling 2x2x2@[2,2,2]
Concat
BN ▸ PReLu ▸ Conv 3x3x3@[2,2,2]
MonoRes
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
ConvTranspose 3x3x3@[2,2,2]
Conv 1x1x1@[1,1,1] ▸ Activation

Fig. 1. Diagram of the U-Net derived model used as the backbone of our deep learning brain tissue segmentation framework. The network consists of a 3D U-Net model using residual convolution blocks and skip connections. The parameter distribution is asymmetrical, with the residual blocks of the encoder using two convolutional blocks while a single block is used in the decoder. In the convolutional layers (Conv), K_x K_y K_z@[S_x,S_y,S_z] indicates the kernel and stride dimensions in each axis. The gray boxes represent the feature maps with the number of channels indicated above or under it. The numbers of input and output feature maps are denoted I and O, respectively.

segmentation of the whole image. On the other hand, local differences are those where only the altered region and its neighborhood are affected. The main source of global segmentation differences in the proposed method is caused by input normalization which is applied to the T1-w scan. Input normalization is a technique used to homogenize the range and statistics of neural network inputs so that the variation is reduced and the model can be more finely tuned to the expected input values. Since our aim is to correctly segment the healthy tissue regardless of any intensity changes caused by the development and evolution of WM lesions, we want an input normalization procedure that is invariant to these appearance changes. Due to the combinatorial nature of neural networks, small perturbations in the input values can cause large output differences; thus, a small shift in the normalization parameters could have a measurable effect on the segmented tissue volumes. To minimize this, we propose the use of a minmax input normalization operation for T1-w MR images that maps intensities between the 0.05% and 99.95% percentiles to the $[-1,1]$ interval and then clamps to that same interval to clip any outliers within the desired range. This kind of normalization has much less variability between the healthy and artificially lesioned images than other tested techniques, such as z score normalization (zero mean and unit standard deviation) or intensity rescaling to the 0–1 range. Local segmentation differences due to the appearance of WM lesions are introduced not only to the lesioned voxels and their neighborhood but also to the whole patch where a lesioned voxel appears. Within the proposed patch-based approach, the introduction of artificial lesions in part of a patch will affect the output probabilities of the whole patch. During inference, the input image is sliced into patches to be segmented and then recombined for whole image segmentation. A larger patch size means that a larger proportion of patches contain lesioned voxels which introduce segmentation differences further from the lesioned voxels. Consequently, patch size is an important parameter for mitigating the local effect of WM lesions in patch-based brain tissue segmentation. To select the patch size, we empirically tested 5 isotropic patch sizes between $8 \times 8 \times 8$ and $40 \times 40 \times 40$. The best compromise between tissue segmentation performance and reducing the aforementioned differences is achieved by using a patch size of $16 \times 16 \times 16$.

To minimize the WM lesion effect within the proposed deep learning segmentation framework, we propose a multi-task double U-Net architecture, depicted in Fig. 2, where the first network performs inpainting and the second network segments the brain tissues. The aim is to obtain a segmentation model that can minimize the effect that a WM lesion has on its healthy neighborhood so that it can be correctly segmented despite the adjacent abnormal intensities. The proposed method takes a skull-stripped brain scan along with its binary WM lesion segmentation and outputs a probability distribution of brain tissue (CSF, GM and WM) for each input voxel. The lesioned area is occluded with zero-valued voxels before input to the network. First, the inpainter network inpaints any occluded lesion voxels in the input patch and tries to reconstruct the originally healthy intensities. The inpainted patch is then masked before tissue segmentation, keeping only the inpainted voxels from the first network and taking the original intensities for the rest of nonlesioned voxels. Finally, the second U-Net performs brain tissue segmentation from the inpainted masked patch and outputs a brain tissue probability distribution for each input voxel. During training, we input artificially lesioned images and target the tissue segmentation of the originally healthy image as output both networks are trained simultaneously in an end-to-end manner to allow the segmentation loss gradients from the second model to also backpropagate through the inpainter. This regularizes the inpainter toward inpainting in a way that should also help the tissue segmenter to more accurately approximate the healthy tissue probabilities. In this way, the goal of the inpainter is not to faithfully and accurately approximate the healthy T1 intensities, but rather, we want the tissue segmentation model to better approximate the healthy tissue probabilities regardless of any occluded zero-valued regions.

In the double chained U-Net configuration, the input of the first U-Net is a T1-w patch with WM lesions occluded and the binary WM lesion mask. The output of the inpainter is activated by a hyperbolic tangent function (tanh) to map the range of output intensities within the same $[-1,1]$ interval of input normalization. The input of the second network, the segmenter U-Net, is an inpainted T1-w patch and its output is activated using the Softmax function to obtain a tissue probability distribution for each input voxel.

3.1. Training

The double U-Net system is trained end-to-end using both the healthy

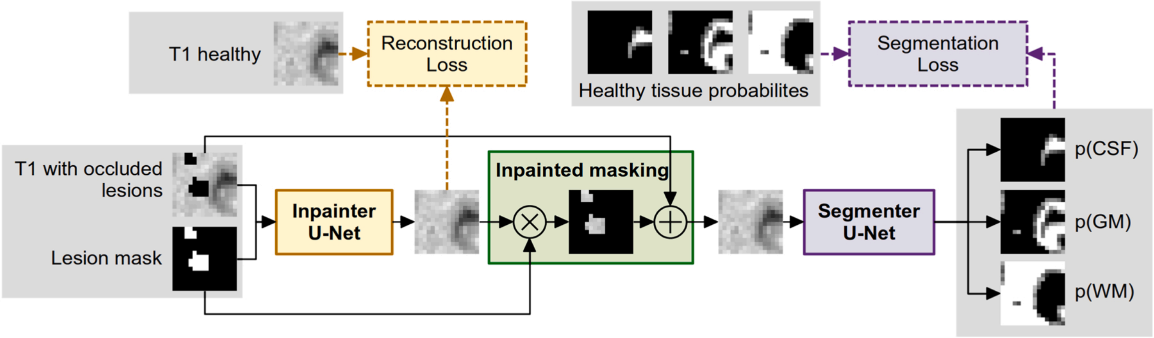


Fig. 2. Overview of the proposed patch-based double chained 3D U-Net architecture performing end-to-end inpainting and brain tissue segmentation. The binary WM lesion mask is used to occlude the lesion from the input patch with zero-valued voxels and is also input to the inpainter. Inpainted masking takes the inpainter output only for lesioned voxels and uses the original intensities for the rest of nonlesioned voxels. The tissue segmenter receives the inpainted masked patch and outputs a probability distribution among the background, CSF, GM and WM classes.

and artificially lesioned images as inputs, targeting the healthy image tissue probabilities as output in both cases. For the artificially lesioned images, the parameters for input normalization are computed only from the nonlesioned brain voxels. To train the proposed patch-based method, we first generate patch training and validation sets. From the training image set, we use 90% of the subjects to build the training patch set and the remaining 10% for the validation patch set. In total, we extract 1 million patches, 900,000 for training and 100,000 for validation. These patches are extracted centered on a set of voxels sampled using a deliberate strategy to balance the representation of segmentation classes as well as the representation of patches with and without lesions half of the patches are extracted evenly from the healthy images without artificial lesions. For each image, the patch centers are sampled using a preliminary FAST (Zhang et al., 2001) tissue segmentation as a guide to obtain 10% centered on the background class and 30% each from the CSF, GM and WM classes. The other half are taken evenly from all the available artificially lesioned images, centered on occluded artificially lesioned voxels. A random 3D offset of up to half the patch size is applied to the healthy and lesion sampled centers to increase the representation of boundaries. The model is then trained end-to-end using the Adadelta optimizer (Zeiler, 2012) with a learning rate of 0.2 and a batch size of 32 patches. To prevent overfitting, early stopping is performed when the loss on the validation set does not improve for 8 consecutive epochs. The loss function used for training and validation is composed of a reconstruction loss, used to train the inpainter, and a segmentation loss that is used to train both the tissue segmenter and inpainter networks. The reconstruction loss uses the mean squared error (MSE) between the original healthy patch and the patch reconstructed by the inpainter. For the segmentation loss, we use a version of the crossentropy loss using probabilistic targets, the probabilistic crossentropy (PCE) loss. Given an output voxel classification y over C classes and a target probability distribution t, the PCE loss is defined as follows:

$$\mathrm{PCE}(y,t) = \sum_{i=1}^{C} - y_i \cdot t_i \cdot \ln\left(\sum_{j=1}^{C} \exp(y_j)\right) \tag{2}$$

By using probabilities as targets instead of categorical labels, we encourage output segmentation that approximates the partial volume probabilities between tissues instead of trying to maximize the probability of the most likely tissue class. Finally, the loss function L is defined as follows:

$$L(\widehat{I_L}, \widehat{S_L}, I_H, S_H) = \mathrm{MSE}(\widehat{I_L}, I_H) + \mathrm{PCE}(\widehat{S_L}, S_H) \tag{3}$$

where $\widehat{I_L}$ is the patch reconstructed by the inpainter, $\widehat{S_L}$ is the tissue probability predicted from the inpainted masked patch and I_H and S_H are the originally healthy image intensities and brain tissue probabilities, respectively.

3.2. Inference

Once the network weights are trained, inference is performed from the T1-w MR image and its WM lesion mask by extracting overlapping patches sampled uniformly with a step size of 5 × 5 × 5 and the same patch size of 16 × 16 × 16 used during training. Performing inference on highly overlapping patches helps reduce block boundary artifacts and improve the spatial coherence of the output probabilities. These patches are then passed through the trained network, obtaining a probability distribution of brain tissue type for each voxel in each input patch. The probability distributions of overlapping patches are then combined through averaging into a common output segmentation space. Finally, the output is normalized to ensure that the tissue probability distributions of each voxel add up to one.

3.3. Implementation details

The proposed method is implemented with Python, using the Torch scientific computing framework (Paszke et al., 2017). All experiments are done on a GNU/Linux machine running Ubuntu 18.04 with 128 GB of RAM and an Intel® Core ™ i7–7800X CPU. Network training and inference are performed with an NVIDIA 1080 Ti GPU (NVIDIA Corp., United States) with 12 GB of G5X memory. Within our method, each U-Net model has approximately 7.03 million trainable parameters, which add up to a total of 14.06 million in the multi-task double U-Net configuration. In our system, the total training time of the proposed method is 22.25 h with an average inference time of 55 s per image in all tests performed. The development framework is available to the research community at https://github.com/NIC-VICOROB/LITS.

4. Evaluation and results

In this section, we evaluate the segmentation performance of the proposed methodology as well as the influence of WM lesions with and without lesion effect minimization. First, the healthy and pathological datasets are randomly split into a training and validation image set to train the proposed methodology and a testing set exclusively used for evaluation of the reported experimental results. From the Calgary–Campinas dataset, 45 scans are used for testing, 15 from each scanner, and the remaining 312 are used for training. The 15 WM lesion masks of the MSSEG Challenge dataset are split into 12 for training and 3 for testing. From the ISBI 2015 Longitudinal MS Lesion Segmentation Challenge dataset, 13 WM lesion masks are taken from 3 subjects for training and 8 masks are taken from the other 2 subjects for testing. Finally, we split the WMH Challenge 2017 dataset into 54 masks for training and 6 for testing. In total, the training set contains 312 healthy brain scans, each with 79 registered WM lesion masks, which amounts to 24,648 healthy and artificially lesioned training image pairs. The testing set is composed of 45 healthy brain scans, each with 17 registered WM lesion masks, making a total of 720 healthy and artificially lesioned testing image pairs. In this way, we ensure that none of the healthy T1-w images or registered lesion masks used during training are used for the evaluation.

To quantitatively evaluate the segmentation differences between healthy and artificially lesioned images, we use the absolute volume difference metric defined in Eq. (4), which is computed separately for the volumes of segmented GM and WM tissues.

$$\text{Abs. volume difference } (\%) = 100\cdot\frac{\left|V_{\text{lesioned}} - V_{\text{healthy}}\right|}{V_{\text{healthy}}} \tag{4}$$

Within the proposed methodology, we also evaluate the differences that WM lesions introduce at local and global scales. In the proposed patch-based method, the introduction of artificially lesioned voxels has a local effect by altering the output probabilities of the whole patch in which they appear. At the global level, the artificially lesioned voxels can alter the input normalization parameters and shift the input values for the whole image, which leads to global output segmentation differences. To evaluate these two effects separately, we also compute the evaluation metrics in two regions of interest (ROIs) related to the lesion neighborhood and patch size. To study the WM lesion effect at a local scale, we define the *within lesion neighborhood* ROI as all the voxels that might appear along the artificial WM lesion in an input patch. More specifically, we include all normally appearing tissue within a patch side length, 16 voxels, of an artificially lesioned voxel in any of the three dimensions. To study the global WM lesion effect, we define the *outside lesion neighborhood* ROI that encompasses all normally appearing voxels at a distance of a patch side length, 16 voxels, or more from an artificially lesioned voxel in all three dimensions.

To assess the statistical significance of differences between the segmentation differences of the baseline and proposed approaches we consider the paired t-test for related samples.

4.1. Tissue segmentation

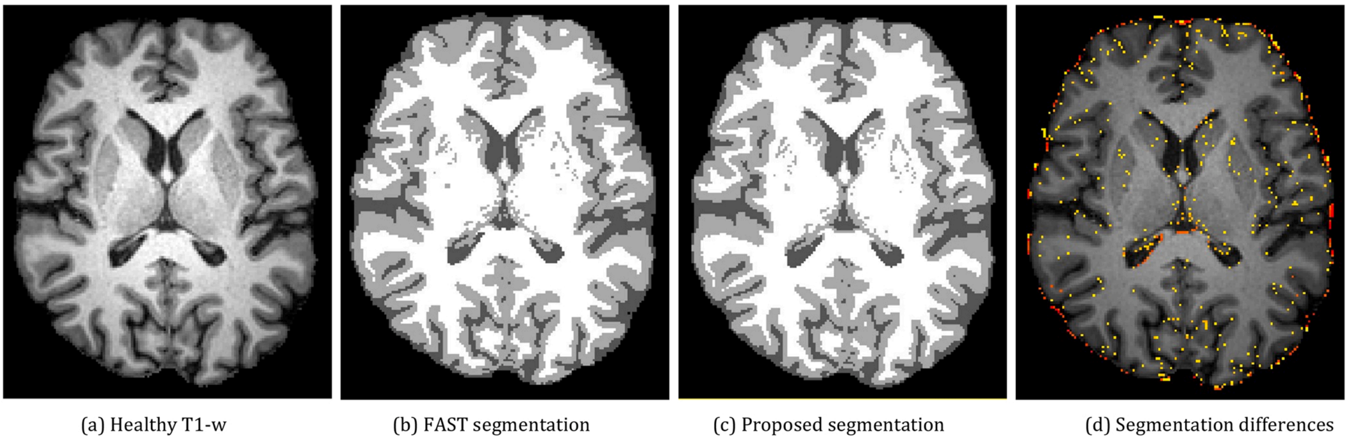
We evaluate the learned tissue segmentation model of the proposed approach by comparing it to FAST (Zhang et al., 2001), the reference method used during training. For this, we segment the 45 testing set images of the healthy dataset without artificially added lesions and compute the Dice similarity coefficient (DSC) with respect to the reference segmentations for the same images. The proposed approach obtains a DSC of 94.6 ± 2.5% in whole brain tissue (CSF + GM + WM) segmentation and a DSC of 99.0 ± 0.1% in parenchyma (GM + WM) segmentation. When individual tissues are considered, the DSCs are 94.6 ± 3.4% and 96.9 ± 1.6% for the GM and WM classes respectively. These results are in line with those of similar deep learning methods also using FAST segmentations as training targets (Rajchl et al., 2018).

Fig. 3 shows qualitative results of segmentation from FAST and the proposed approach as well as the differences between them, which are mainly located within tissue interfaces and in the brain mask edges the large segmentation differences located in the outer brain border appear because FAST assumes every nonzero voxel within the given brain mask has to be segmented as one of the tissues, which in this case is CSF. In contrast, the proposed approach does not make this assumption and mostly classifies voxels in the outer brain border as background instead of CSF. Although the interfaces between tissues with a strong partial volume effect are also a source of segmentation differences, Fig. 3d shows that the changes in classified tissues are due to quite small probability shifts that bias the most likely tissue class one way or the other. The probability differences are larger in the interfaces between WM and CSF, such as the ventricle border and, especially, in its lower left part. In these regions, the partial volumes between GM and CSF take an intensity value similar to that of the GM class and are mostly classified by FAST as GM, while the proposed deep learning approach tends to classify them as mostly WM.

4.2. Lesion effect

We evaluate the effect of WM lesions on tissue segmentation when no WM lesion effect minimization techniques are used. For this, we segment the healthy and artificially lesioned testing image pairs and compute the volume differences between each pair for GM and WM tissues. In this experiment, the inpainting network of the proposed method is essentially turned off, as empty WM masks are used for inference and artificial lesions are not occluded in the input images. We also evaluate the WM lesion effect on the FAST (Zhang et al., 2001) tissue segmentation method from the FSL package. Table 1 shows the absolute volume differences of GM and WM volumes for FAST and our deep learning segmentation method without inpainting. Overall, the proposed method is significantly less influenced by the presence of WM lesions at both local and global scales than FAST ($p < 0.01$). The proposed segmentation method shows an almost exclusively local influence, as nearly all the differences are located within the lesion neighborhood ROI. In contrast, the FAST segmentation method has a mostly global lesion influence, with high volume differences both within and outside the lesion neighborhood ROI.

Fig. 4 shows the tissue probability differences from a representative example of the lesion effect experiment for both tissue segmentation methods. In both cases, artificially lesioned voxels display large


(a) Healthy T1-w

(b) FAST segmentation

(c) Proposed segmentation

(d) Segmentation differences

Fig. 3. Comparison of segmentation results of FAST (b) and the proposed approach (c). (d) Absolute probability differences of voxels changing their most likely tissue class overlaid with a yellow to red colormap, where yellow corresponds to a difference of 0.02 and red corresponds to a difference of 1.0 or higher in the voxelwise sum of absolute probability differences. Differences between both methods are mainly located within tissue interfaces and in the outer brain border.

Table 1
Abs. volume differences (%) of the GM and WM of the healthy and artificially lesioned testing image pairs. The absolute volume differences of the proposed approach are all significantly lower ($p < 0.01$) than those of the baseline FAST method.

	FAST		Proposed (without inpainting)	
Tissue	mean ± std	median	mean ± std	median
(i) Whole brain				
GM	0.89 ± 1.14	0.27	0.07 ± 0.09	0.05
WM	1.22 ± 1.58	0.35	0.10 ± 0.11	0.07
(ii) Within lesion neighborhood				
GM	0.96 ± 1.23	0.34	0.13 ± 0.11	0.10
WM	1.10 ± 1.50	0.26	0.13 ± 0.13	0.11
(iii) Outside lesion neighborhood				
GM	0.70 ± 0.89	0.24	0.01 ± 0.03	0.00
WM	1.91 ± 2.66	0.54	0.01 ± 0.06	0.00

probability shifts caused by their newer darker intensities. The effect is not limited to these voxels and spreads to their neighborhood and even to the rest of the image. The FAST tissue segmentation method shows a large number of sparse small and medium probability differences mainly located in the interfaces between GM and WM tissue throughout the whole image. In contrast, the proposed patch-based deep learning approach displays groups of small probability shifts located around the artificially lesioned voxels and nearby structures. The differences are exclusively located in the *within lesion neighborhood* ROI, with no differences in the rest of the image. In contrast, the segmentation differences of FAST appearing within the whole image add up to a larger volume shift.

4.3. Lesion effect minimization

In this experiment, we evaluate how well the WM lesion effect minimization techniques reduce the GM and WM volume differences between segmentations of healthy and artificially lesioned images. In

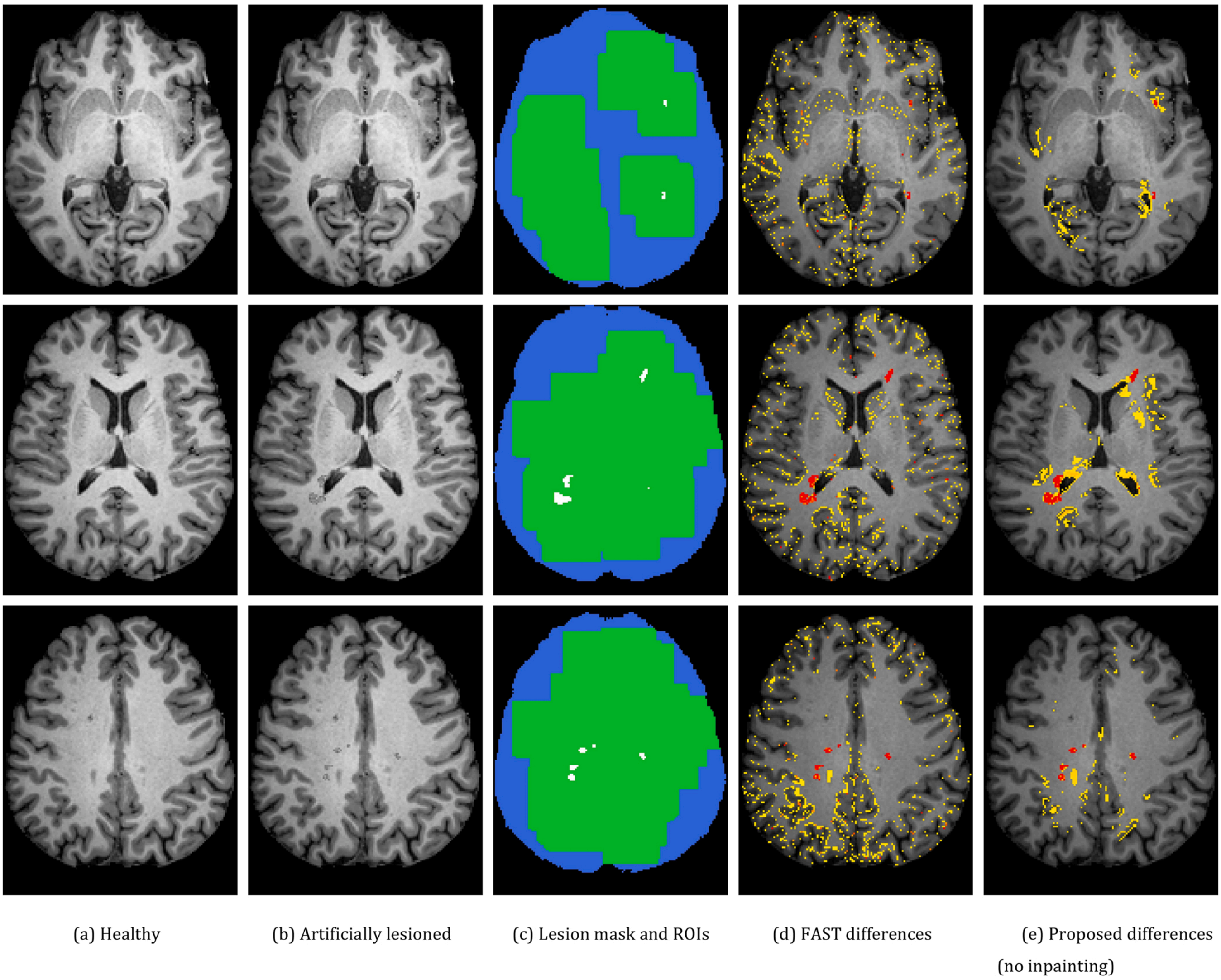

(a) Healthy
(b) Artificially lesioned
(c) Lesion mask and ROIs
(d) FAST differences
(e) Proposed differences (no inpainting)

Fig. 4. Representative example of the absolute segmentation differences between healthy and artificially lesioned brain tissue segmentations without WM lesion effect minimization. Columns 4a and 4b show three axial slices from the healthy and artificially lesioned images that were segmented. In 4c, the artificial lesion mask is shown in white, the *within lesion neighborhood* ROI is shown in green and the *outside lesion neighborhood* ROI is shown in blue. In 4d and 4e, the absolute probability differences are shown overlaid in a yellow to red colormap, where yellow corresponds to a difference of 0.02 and red corresponds to a 1.0 or greater difference in the voxelwise sum of both GM and WM absolute probability differences. While the proposed approach shows large clusters of small differences close to the artificially lesioned voxels, FAST is affected by a larger number of sparsely distributed differences over the whole image which, overall, add up to a larger shift in measured tissue volumes.

the proposed approach, we perform end-to-end inpainting and segmentation by occluding the artificial lesions with zero-valued voxels and providing the WM lesion mask as an additional input to our network. As a baseline comparison, we evaluate the WM lesion inpainting algorithm (FSL_inpainting) provided in the FSL package (Battaglini et al., 2012) to fill the lesion intensities before segmenting the brain tissues with FAST (Zhang et al., 2001). We also evaluate the use of FSL_inpainting to inpaint the WM lesions prior to performing brain tissue segmentation with our deep learning model which, in this case, is provided with empty WM lesion masks to avoid using the inpainter network.

The absolute volume differences of the segmentations of healthy and artificially lesioned versions of the same image are summarized in Table 2. Compared with the results in Table 1, the use of FSL_inpainting and our proposed method significantly reduce the volume differences for all methods ($p < 0.01$). Compared with the FSL_inpainting + FAST pipeline, the FSL_inpainting + Proposed method obtains significantly lower volume differences in all the considered ROIs ($p < 0.001$). This shows that the proposed deep learn- ing brain tissue segmentation framework is more robust to the error introduced by WM lesions even when using classical inpainting methods. Compared with FSL_inpainting + FAST, the proposed approach obtains significantly lower volume differences in all ROIs ($p < 0.001$). However, when comparing with the FSL_inpainting + Proposed approach, the proposed method obtains significantly lower *Whole brain* and *within lesion neighborhood* volume differences ($p < 0.001$), but significantly higher *outside lesion neighborhood* differences ($p < 0.01$). In this case, the input normalization parameters are much less affected by the intensities inpainted by FSL_inpainting than by the zeroes that are used to occlude those same voxels within the proposed approach. The *outside lesion neighborhood* ROI differences of the proposed approach increase significantly compared to those without performing inpainting in Table 1 ($p < 10^{-8}$). This is due to the occlusion with zeroes that we perform to the artificially lesioned voxels in the proposed approach, which slightly change the value of input normalization parameters and increase the segmentation differences for the whole image.

Fig. 5 shows the correlation between the artificial lesion volume and absolute GM and WM volume differences for the evaluated methods. Larger lesion loads tend to increase the segmentation differences for all methods, however, the ones using our deep learning based brain tissue segmentation model show a much lower error when larger lesion volumes are considered. This shows that the poor performance of the FSL pipeline on big lesions is not related to FSL_inpainting, since the proposed deep learning based tissue segmentation framework also takes in images preprocessed with FSL_inpainting and performs much better on larger lesion loads.

Table 2
Abs. volume differences (%) of the GM and WM of the segmentations of healthy and artificially lesioned testing image pairs when using lesion effect minimization techniques. Compared with the FSL inpainting + FAST method, both the Proposed and FSL_inpainting + Proposed approaches obtain significantly lower volume differences in all ROIs than the FSL_inpainting + FAST pipeline ($p < 0.001$). When comparing with the FSL_inpainting + Proposed approach, the proposed method obtains significantly lower *whole brain* and *within lesion neighborhood* volume differences ($p < 0.01$).

	FSL_inpainting + FAST		FSL_inpainting + Proposed		Proposed	
Tissue	mean ± std	median	mean ± std	median	mean ± std	median
(i) Whole brain						
GM	0.05 ± 0.09	0.014	0.02 ± 0.03	0.009	0.01 ± 0.03	0.004
WM	0.08 ± 0.14	0.020	0.03 ± 0.04	0.012	0.02 ± 0.04	0.005
(ii) Within lesion neighborhood						
GM	0.06 ± 0.10	0.019	0.04 ± 0.04	0.021	0.02 ± 0.03	0.008
WM	0.08 ± 0.14	0.018	0.04 ± 0.04	0.020	0.02 ± 0.03	0.007
(iii) Outside lesion neighborhood						
GM	0.04 ± 0.07	0.011	0.01 ± 0.02	0.000	0.01 ± 0.03	0.000
WM	0.13 ± 0.23	0.032	0.01 ± 0.04	0.000	0.03 ± 0.07	0.000

In terms of execution time, a brain tissue segmentation done with FAST within our system takes an average of 3.25 min per scan, while the FSL_inpainting part takes less than a second to complete. In total, the FSL pipeline doing WM lesion inpainting and tissue segmentation takes 7 min to process a scan since it requires two separate FAST executions, one to obtain the white matter segmentation mask required by FSL_inpainting and another to obtain the actual brain tissue segmentation from the inpainted image. In contrast, the proposed method doing end-to-end inpainting and tissue segmentation takes an average of 1 min to process a single scan.

5. Discussion

In this work, we focused on deep learning methods for brain tissue segmentation and performed the first study on the effect of WM lesions in this kind of approaches. We have proposed a deep learning based framework for brain volumetry which learns from a reference classical method and incorporates techniques to deal better with pathological cases having WM lesions. We have also proposed a multi-task double U-Net architecture, along with a training data generation procedure, to embed the WM lesion effect reduction within the brain tissue segmentation method itself. In our approach, instead of performing lesion inpainting in a previous separate step, we perform end-to-end WM lesion inpainting and brain tissue segmentation. By jointly optimizing both tasks, the inpainter is also trained to aid in the segmentation task through the gradient updates coming from the segmentation loss. In this sense, the actual quality or accuracy of inpainting in our framework is not important as long as the output segmentation more faithfully approximates the healthy tissue probabilities.

Without any kind of lesion inpainting, the tissue volumes provided by the proposed deep learning based framework are much less affected by the presence of WM lesions compared to the reference method used for training. Since the introduced artificial lesions affect the tissue probabilities of the patches where they appear, the use of a small patch size constrains the local effect to a smaller area around the lesion. Artificial lesions also change the estimated input normalization parameters which are calculated using all the image intensities. However, the proposed input normalization based on image percentiles is quite robust against these intensity changes and avoids any global segmentation differences in most cases. In comparison, FAST is affected by a larger number of sparse segmentation differences spread out over the whole image which, overall, add up to a larger shift in measured tissue volumes. This is most likely due to the initial k-means clustering step that FAST performs over the entire image to estimate the mean intensity of each tissue, which is later used during the estimation of partial volume probabilities. The introduction of artificial lesions biases the estimated mean intensity of each tissue which in turn biases the estimation of partial volume distributions, producing the observed segmentation differences in the interfaces between tissues.

In terms of WM lesion effect minimization, both the FSL_inpainting and our proposed approach significantly reduce their effect on the measured tissue volumes. However, we obtain significantly lower volume differences than the baseline FSL pipeline, especially when considering larger lesion loads. The results in Fig. 5 show that our deep learning tissue segmentation framework provides significant improvement even when using FSL_inpainting to preprocess the images. Furthermore, our proposed deep learning framework is faster, taking just under a minute to segment a whole brain scan while the baseline FSL pipeline takes an average of 7 min.

The main limitation of this study is that we cannot assess or evaluate

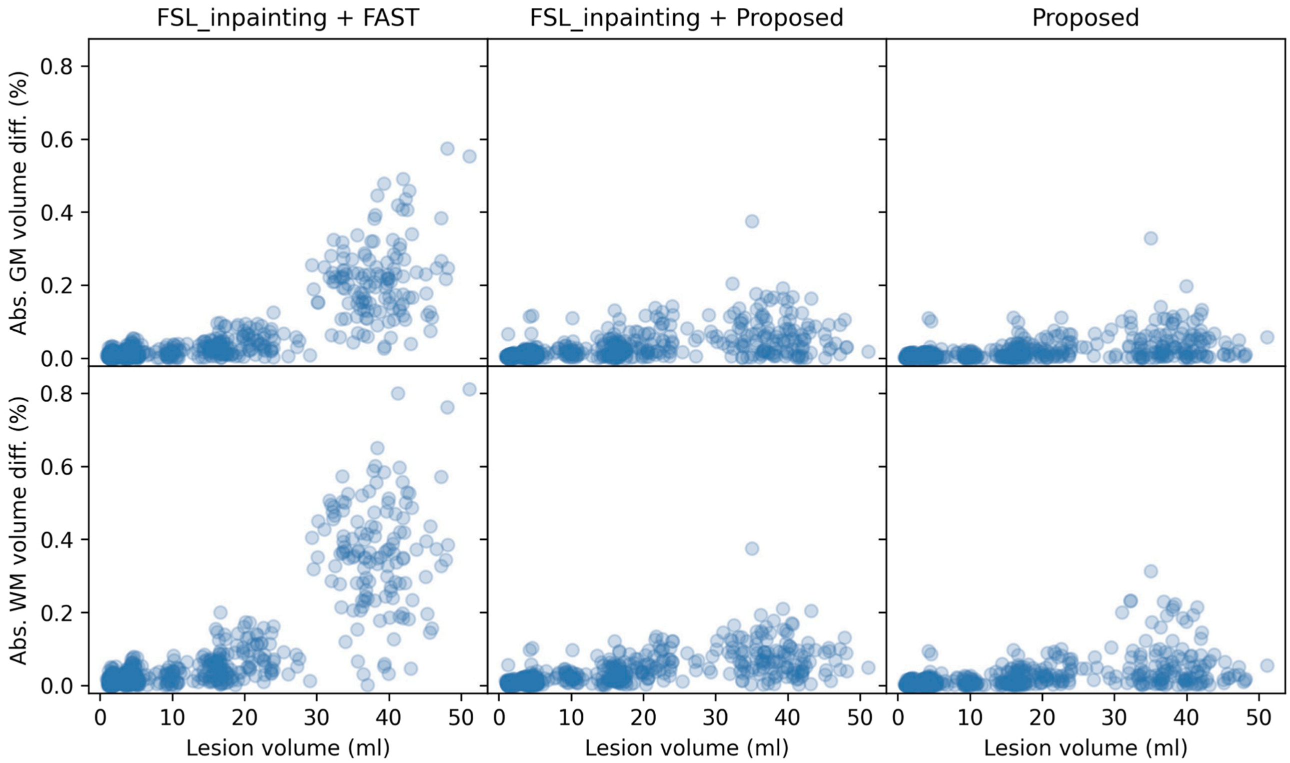


Fig. 5. Correlation of artificial lesion volume and the absolute volume differences (%) of the GM and WM of healthy and artificially lesioned images when using lesion inpainting.

the accuracy and precision of the learned tissue segmentation model and its lesion effect minimization performance on real WM lesions. Due to the way in which the proposed approach is trained, this requires a large database of MR images with manually annotated brain tissue and WM lesions of both healthy and pathological subjects. However, there is no such database, and our evaluation has therefore been limited to relative comparisons with FAST as the gold standard on artificially lesioned images. In this sense, our approach presents a lower WM lesion effect with and without inpainting with a Dice similarity coefficient of 94.6 ± 2.5% relative to FAST brain tissue segmentations. Unlike supervised learning methods using manually annotated segmentations for training, a higher DSC compared to that of the FAST segmentation is not indicative of better quality or accuracy, just of higher similarity. Unlike FAST, deep learning methods suffer from the domain adaptation issue where their performance is not guaranteed outside of the image domains used during training. In this sense, a different MR scanner or acquisition protocol than those used during training would likely lead to a decreased segmentation performance. In such cases, training a model from scratch on the target image domain only requires a set of healthy MR images from that domain to which WM lesion masks from publicly accessible pathological scans can be registered to train the proposed method. Another option is to use domain adaptation techniques that fine-tune pretrained network weights to optimize the model for the target domain.

In the proposed method, accurate WM lesion segmentation is required to obtain optimal results, and over or undersegmentation of the WM lesion would still introduce volume errors in the output segmentation. This could be an issue since manual lesion delineation or automated lesion segmentation is often performed on FLAIR MR images, while brain tissue volumetry is usually performed on T1-w MR images (Rovira et al., 2015). In this case, the FLAIR lesion segmentation mask is usually registered to the T1 image and might not encompass all abnormally appearing voxels in the target modality image. In the case of oversegmentation, the method can deal just as well with the inpainting and segmentation of larger occluded areas as long as they are to be segmented as WM. Due to the way the method was trained, any occluded voxel is assumed to be WM in its majority and will be segmented as such. If the WM lesion is undersegmented, the lesioned voxels are not inpainted, which introduces errors in neighboring tissue segmentation. However, the experimental results without inpainting show that the effect is still be smaller than that of FAST and confined to the undersegmented lesioned voxels neighborhood.

6. Conclusions

In this work, we focus on deep learning based tissue segmentation methods for brain volumetry and studied the error introduced by WM lesions. We have proposed a deep learning framework for brain tissue segmentation which is much less affected by WM lesions than the reference method used to train thanks to the use of small patches and a percentile-based input normalization. We have also proposed a multitask double U-Net architecture, along with a training data generation procedure, which performs lesion inpainting and tissue segmentation in an end-to-end manner and can reduce the WM lesion effect to almost negligible levels. Reducing the effect of WM lesions is critical for accurate and reliable cross-sectional volumetry or longitudinal brain atrophy quantification. Typically, state-of-the-art atrophy quantification approaches are based either on boundary shift integration (Smith et al., 2002) or Jacobian integration (Boyes et al., 2006), both of which rely on prior accurate segmentation of brain tissue which needs to be robust against the influence of WM lesions. Automated brain volumetry methods are currently only used to evaluate the efficacy of experimental therapies and to correlate with treatment outcomes in clinical studies. Improving their accuracy would either strengthen the statistical significance of correlations or reduce the sample sizes needed to establish them. In routine clinical practice, the use of brain volumetry methods is discouraged for prognosis, such as assessing patient progression in MS (Rovira et al., 2015). These methods are unreliable when applied to a single subject instead of a large population due to the inherent technical issues and other confounding factors that severely affect brain volumetry methods. Improving the accuracy and reducing the error from confounding factors such as WM lesions is critical to unlock brain volumetry as an imaging marker for the prognosis of patients with neurodegenerative diseases. In this sense, the proposed deep learning methodology is significantly less affected by WM lesions and can minimize the error they introduce in the measured tissue volumes.

CRediT authorship contribution statement

Albert Clèrigues: Conception and design of study, analysis and/or interpretation of data, Drafting the manuscript, Approval of the version of the manuscript to be published. **Sergi Valverde:** Conception and design of study, analysis and/or interpretation of data, Approval of the version of the manuscript to be published. **Joaquim Salvi:** Drafting the manuscript, revising the manuscript critically for important intellectual content, Approval of the version of the manuscript to be published. **Arnau Oliver:** Conception and design of study, Drafting the manuscript, revising the manuscript critically for important intellectual content, Approval of the version of the manuscript to be published. **Xavier Lladó:** Conception and design of study, Drafting the manuscript, revising the manuscript critically for important intellectual content, Approval of the version of the manuscript to be published.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

Albert Clèrigues holds an FPI grant from the Ministerio de Ciencia e Innovación with reference number PRE2018-083507. This work has been partially supported by DPI2020-114769RB-I00 from the Ministerio de Ciencia e Innovación. This work is also supported by ICREA under the ICREA Academia programme.

References

Armanious, K., Mecky, Y., Gatidis, S., Yang, B., 2019. Adversarial inpainting of medical image modalities. In: Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3267–3271.

Battaglini, M., Jenkinson, M., Stefano, N.D., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33, 2062–2071.

Bendfeldt, K., Kuster, P., Traud, S., Egger, H., Winklhofer, S., Mueller-Lenke, N., Naegelin, Y., Gass, A., Kappos, L., Matthews, P.M., Nichols, T.E., Radue, E.W., Borgwardt, S.J., 2009. Association of regional gray matter volume loss and progression of white matter lesions in multiple sclerosis — a longitudinal voxel-based morphometry study. NeuroImage 45, 60–67.

Boyes, R.G., Rueckert, D., Aljabar, P., Whitwell, J., Schott, J.M., Hill, D.L., Fox, N.C., 2006. Cerebral atrophy measurements using jacobian integration: comparison with the boundary shift integral. NeuroImage 32, 159–169.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Bazin, P.L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D. S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation data resource. Data Brief 12, 346–350.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32, 223–228.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Améli, R., Ferré, J.C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., McKinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Lladó, X., Santos, M.M., Santos, W.P., Silva-Filho, A.G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F.J., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F., Barillot, C., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. 8, 13650.

Cortese, R., Battaglini, M., Sormani, M.P., Luchetti, L., Gentile, G., Inderyas, M., Alexandri, N., De Stefano, N., 2022. Reduction in grey matter atrophy in patients with relapsing multiple sclerosis following treatment with cladribine tablets. Eur. J. Neurol.

Di Filippo, M., Anderson, V.M., Altmann, D.R., Swanton, J.K., Plant, G.T., Thompson, A. J., Miller, D.H., 2010. Brain atrophy and lesion load measures over 1 year relate to clinical status after 6 years in patients with clinically isolated syndromes. J. Neurol. Neurosurg. Psychiatry 81, 204–208.

Ghione, E., Bergsland, N., Dwyer, M., Hagemeier, J., Jakimovski, D., Ramasamy, D., Hojnacki, D., Lizarraga, A., Kolb, C., Eckert, S., Weinstock-Guttman, B., Zivadinov, R., 2020. Disability improvement is associated with less brain atrophy development in multiple sclerosis. Am. J. Neuroradiol. 41, 1577–1583.

González-Villà, S., Valverde, S., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Alex Rovira, Oliver, A., Lladó, X., 2017. Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation. NeuroImage Clin. 15, 228–238.

Guha Roy, A., Conjeti, S., Navab, N., Wachinger, C., 2019. Quicknat: a fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage 186, 713–727.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30, 1617–1634.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, pp. 448–456.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Kuijf, H.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Biesbroek, J.M., Luna, M., Mahmood, Q., Mckinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S. H., Pezold, S., Puybareau, E., Bresser, J.D., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Heinen, R., Chen, C., Flier, W.V.D., Barkhof, F., Viergever, M.A., Biessels, G.J., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Trans. Med. Imaging 38, 2556–2568.

Lansley, J., Mataix-Cols, D., Grau, M., Radua, J., Sastre-Garriga, J., 2013. Localized grey matter atrophy in multiple sclerosis: a meta-analysis of voxel-based morphometry studies and associations with functional disability. Neurosci. Biobehav. Rev. 37, 819–830.

Magon, S., Gaetano, L., Chakravarty, M.M., Lerch, J.P., Naegelin, Y., Stippich, C., Kappos, L., Radue, E.-W., Sprenger, T., 2014. White matter lesion filling improves the accuracy of cortical thickness measurements in multiple sclerosis patients: a longitudinal study. BMC Neurosci. 15.

Manjón, J.V., Romero, J.E., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., Tourdias, T., Coupé, P., 2020. Blind mri brain lesion inpainting using deep learning. In: Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging 12417 LNCS, pp. 41–49.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the International Conference on Machine Learning, pp. 807–814.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. Neural Inf. Process. Syst.

Pérez-Miralles, F., Sastre-Garriga, J., Tintoré, M., Arrambide, G., Nos, C., Perkal, H., Río, J., Edo, M., Horga, A., Castilló, J., Auger, C., Huerga, E., Rovira, A., Montalban, X., 2013. Clinical impact of early brain atrophy in clinically isolated syndromes. Mult. Scler. J. 19, 1878–1886.

Prados, F., Cardoso, M.J., Kanber, B., Ciccarelli, O., Kapoor, R., Wheeler-Kingshott, C.A. G., Ourselin, S., 2016. A multi-timepoint modality-agnostic patch-based method for lesion filling in multiple sclerosis. NeuroImage 139, 376–384.

Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P.M., Glocker, B., 2018. Neuronet: Fast and robust reproduction of multiple brain image segmentation pipelines. arXiv preprint arXiv:1806.04224.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention, 9351, pp. 234–241.

Rovira, À., Wattjes, M.P., Tintoré, M., Tur, C., Yousry, T.A., Sormani, M.P., Stefano, C., Filippi, M., Auger, C., Rocca, M.A., Barkhof, F., Fazekas, F., Kappos, L., Polman, C., Miller, D., Montalban, X., 2015. Magnims consensus guidelines on the use of mri in multiple sclerosis—clinical implementation in the diagnostic process. Nat. Rev. Neurol. 8 (11), 471–482.

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

Sotirchos, E.S., Gonzalez-Caldito, N., Dewey, B.E., Fitzgerald, K.C., Glaister, J., Filippatou, A., Ogbuokiri, E., Feldman, S., Kwakyi, O., Risher, H., Crainiceanu, C., Pham, D.L., Zijl, P.C.V., Mowry, E.M., Reich, D.S., Prince, J.L., Calabresi, P.A., Saidha, S., 2020. Effect of disease-modifying therapies on subcortical gray matter atrophy in multiple sclerosis. Mult. Scler. 26, 312–321.

Souza, R., Lucena, O., Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rittner, L., Frayne, R., Lotufo, R., 2018. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. NeuroImage 170, 482–494.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tang, Z., Cabezas, M., Liu, D., Barnett, M., Barnett, W., Wang, C., 2021. Lg-net: lesion gate network for multiple sclerosis lesion inpainting. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 660–669.

Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion-filling approach to improve brain tissue volume measurements. NeuroImage Clin. 6, 86–92.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J.T., Barkhof, F., Benavente, O.R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12, 822–838.

Xiong, H., Wang, C., Barnett, M., Wang, C., 2020. Multiple sclerosis lesion filling using a non-lesion attention based convolutional network. In: Proceedings of the International Conference on Neural Information Processing 12532 LNCS, pp. 448–460.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Zhang, H., Bakshi, R., Bagnato, F., Oguz, I., 2020. Robust multiple sclerosis lesion inpainting with edge prior. Mach. Learn. Med. Imaging 120–129.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.

Chapter 6

Improving segmentation-based brain atrophy quantification with unsupervised deep learning using tissue similarity regularization

In this chapter, we present our approach for segmentation-based longitudinal atrophy quantification with an unsupervised deep learning approach using a novel tissue similarity regularization that penalizes volume differences between pairs of co-registered short interval scans. These scan pairs are typically used for evaluation of quantification error by assuming that an ideal method would measure zero change between them. In our approach, we use this same assumption to regularize the segmentation model during training and reduce the errors and biases that are learned from the reference method. Typically, skull-stripping and tissue segmentation are performed in separate steps, but here we perform them in an end-to-end fashion so that the training regularization can also reduce the intracranial cavity volume differences. While training is done with pairs of short interval scans, inference in the longitudinal case is performed independently for the baseline and follow-up scans.

Submitted to **Medical Image Analysis** (MedIA)
JCR CSAI IF 8.880, Q1(5/133)

Improving segmentation-based brain atrophy quantification with unsupervised deep learning using tissue similarity regularization

Albert Clèrigues[a,*], Sergi Valverde[b], Arnau Oliver[a], Xavier Lladó[a],
and for the Alzheimer's Disease Neuroimaging Initiative[**]

[a] *Institute of Computer Vision and Robotics, University of Girona, Spain*
[b] *Tensor Medical, Girona, Spain*

Abstract

Brain atrophy measurements derived from magnetic resonance imaging (MRI) are a promising marker for the diagnosis and prognosis of neurodegenerative pathologies such as Alzheimer's disease or multiple sclerosis. However, its use in individualized assessments is currently discouraged due to a series of technical and biological issues. In this work, we present an unsupervised deep learning pipeline for segmentation-based brain atrophy quantification that improves upon the reference method from which it learns. This goal is achieved through tissue similarity regularization that exploits the a priori knowledge that scans from the same subject made within a short interval must have similar tissue volumes. To train the presented pipeline, we use unlabeled pairs of T1-weighted MRI scans having a tissue similarity prior, and generate the target brain tissue segmentations in a fully unsupervised manner using the fsl_anat pipeline implemented in the FMRIB Software Library (FSL). Tissue similarity regularization is enforced during training through a weighted loss term that penalizes tissue volume differences between short-interval scan pairs from the same subject. In inference, the pipeline performs end-to-end skull stripping and brain tissue segmentation from a single T1-weighted MRI scan in its native space, i.e., without performing image interpolation. For longitudinal evaluation, each image is independently segmented first, and then measures of change are computed. We evaluate the presented pipeline in two different MR datasets, MIRIAD and ADNI1, which have longitudinal and short-interval imaging from healthy controls (HC) and Alzheimer's disease (AD) subjects. In short-interval scan pairs, tissue similarity regularization reduces the quantification error and improves the consistency of measured tissue volumes. In the longitudinal case, the proposed pipeline shows reduced variability of atrophy measures and higher effect sizes of differences in annualized rates between HC and AD subjects. Our pipeline obtains a Cohen's d effect size of $d = 1.89$ on the MIRIAD dataset, an increase from the reference pipeline used to train it ($d = 1.01$), and higher than that of SIENA ($d = 1.73$), a well-known state-of-the-art approach. In the ADNI1 dataset, the proposed pipeline improves its effect size ($d = 1.39$) with respect to the reference pipeline ($d = 0.80$) and surpasses SIENA ($d = 1.33$). The proposed data-driven deep learning regularization reduces the biases and systematic errors learned from the reference segmentation method, which is used to generate the training targets. Improving the accuracy and reliability of atrophy quantification methods is essential to unlock brain atrophy as a diagnostic and prognostic marker in neurodegenerative pathologies.

Keywords: magnetic resonance imaging, brain tissue segmentation, deep learning, brain atrophy quantification

1. Introduction

Global and regional brain atrophy quantification has been shown to be a relevant marker for prognosis of neurodegenerative pathologies, such as Alzheimer's disease (AD) (Pini et al., 2016) and multiple sclerosis (MS) (Rocca et al., 2017). Magnetic resonance imaging (MRI) allows for noninvasive quantitative measures of global and regional atrophy of the brain parenchyma. These measurements are typically obtained from longitudinal T1-weighted (T1-w) images, on which there is good contrast between the cerebrospinal fluid (CSF) and the distinct gray matter (GM) and white matter (WM) components that form the brain parenchyma. Methods for brain atrophy quantification are currently affected by a number of confounding factors related to image acquisition, technical issues and pathophysiological changes (Sastre-Garriga et al., 2020), reducing their reliability and applicability. Although MRI-derived measurements of atrophy have proven useful for clinical population studies analyzing disease progression

[*] Corresponding author. A. Clèrigues, Ed. P-IV, Campus Montilivi, University of Girona, 17003 Girona (Spain). e-mail: albert.clerigues@udg.edu. Phone: +34 683645681; Fax: +34 972 418976.

[**] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

or treatment effects, they are still not considered sufficiently accurate or reliable for their use in individualized assessments (Rovira et al., 2015).

In general, longitudinal brain atrophy quantification methods can be classified into either segmentation-based or registration-based techniques. In segmentation-based methods, a target set of structures or tissues is independently segmented in each of the longitudinal scans, and atrophy is quantified from differences in the measured volumes. In contrast, registration-based techniques derive measures of atrophy from the observed spatial deformation of structures or tissues between two longitudinal scans. Segmentation-based methods are typically regarded as less accurate and more variable than their registration-based counterparts, and their use has been discouraged for longitudinal studies (Sastre-Garriga et al., 2017). Although several segmentation-based methods for cross-sectional brain volumetry from T1-w MRI have been proposed in the recent literature, only SIENA-XL (Battaglini et al., 2018) has been purposefully built for longitudinal imaging. Registration-based methods are typically preferred for longitudinal change analysis since they have lower quantification error and better sensitivity to atrophy changes (Sastre-Garriga et al., 2017). SIENA (Smith et al., 2002) is a well-known and widely used registration-based atrophy quantification method based on the boundary shift integral (BSI) (Freeborough and Fox, 1997). Within SIENA, atrophy is measured between two linearly registered scans from the surface displacement of the interface between GM and WM, which was obtained from FAST (Zhang et al., 2001) tissue segmentations of each scan. Measures of longitudinal change can also be derived from the deformation fields obtained from nonlinear registration between baseline and follow-up. The work of Holland and Dale (2011) used the deformation field to approximate voxels as irregular hexahedrons and directly compute the fractional volume change of a certain region between timepoints. More recently, methods based on Jacobian integration of displacement fields have shown further improvements, such as larger effect sizes and lower quantification error (Nakamura et al., 2014; Smeets et al., 2016). These methods measure volume changes by integrating the determinant of the Jacobian of a nonlinear transformation between two longitudinal scans. The region for integration is typically obtained from a cross-sectional segmentation of tissues or structures in one of the scans. It is worth noting that even within registration-based methods, some form of cross-sectional segmentation of tissue or structures is still needed.

In recent years, deep learning techniques have achieved higher levels of accuracy and performance in brain MRI segmentation tasks for Alzheimer's disease (Yamanakkanavar et al., 2020). Several deep learning approaches have been recently proposed for cross-sectional brain tissue segmentation using a mix of automated and manually annotated data. QuickNAT (Guha Roy et al., 2019) is first trained on automated segmentations made with FreeSurfer (Fischl et al., 2002) and then fine-tuned on manually delineations of brain tissue. In contrast, FastSurfer (Henschel et al., 2020) and NeuroNet (Rajchl et al., 2018) are both trained solely on automated brain tissue segmentations made with FreeSurfer (Fischl et al., 2002) and FSL (Jenkinson et al., 2012), respectively, two of the most frequently used automated tools for neuroanatomical analysis. These approaches achieve greater consistency, reliability and shorter execution time than the reference methods on which they were trained. Moreover, both QuickNAT and FastSurfer also demonstrate improvements with respect to longitudinal brain atrophy quantification, having lower short interval error and higher sensitivity to atrophy changes. For pathological cases with brain lesions, Dorent et al. (2021) used several disjoint heterogeneous datasets with manual annotations to learn a joint brain tissue and lesion segmentation model. This approach is much more robust to the volumetric errors introduced by the presence of abnormal brain lesions and can also deal with a variable number of input modalities. While advances in methods for cross-sectional brain tissue segmentation can be used by atrophy quantification approaches to improve their longitudinal results, there is still no deep learning approach that is purposefully built toward improving brain tissue segmentation in the longitudinal case.

In this work, we present an unsupervised deep learning pipeline for segmentation-based brain atrophy quantification that uses tissue similarity regularization to improve upon the reference method used for training. The proposed regularization exploits a priori knowledge that pairs of scans from the same subject made within a short interval must have similar brain tissue volumes. The pipeline is trained using a set of short-interval scan pairs from which training targets are generated in a fully unsupervised manner using the fsl_anat pipeline provided in FSL. The reference tissue segmentations are obtained from fsl_anat in a similar fashion to SIENA-XL (Battaglini et al., 2018) by merging the resulting brain tissue segmentation of FAST (Zhang et al., 2001) and the deep gray matter structures of FIRST (Patenaude et al., 2011). Tissue similarity regularization is enforced during training through a weighted loss term that penalizes volume differences between similar scan pairs. In inference, the pipeline acts on a single T1-w scan in its native space and performs end-to-end skull stripping and brain tissue segmentation. For longitudinal evaluation, each image is independently segmented, and then change measures are computed. We performed a quantitative and qualitative evaluation of the improvements in brain atrophy quantification using two publicly accessible longitudinal MR datasets, MIRIAD and ADNI1. The presented pipeline improves upon the reference method used for training by having a lower quantification error, better intracranial cavity consistency and higher sensitivity to differences in brain atrophy rates between healthy controls and Alzheimer's disease (AD) patients.

2. Materials

2.1. MIRIAD dataset

The Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset (Malone et al., 2013) is a publicly accessible series of longitudinal T1 MRI scans of 46 mild–moderate Alzheimer's subjects and 23 healthy controls with an average age of 69.5±7.1 years old. The dataset consists of longitudinal scans taken at intervals of 2, 6, 14, 26, 38 and 52 weeks and 18 and 24 months from baseline, as well as rescan images at three of the timepoints, for both AD and controls. The rescan images were taken during three of the scanning sessions (0, 6 and 38 weeks) without repositioning of the subject. All scans were taken by the same radiographer on the same 1.5 T Signa MRI scanner (GE Medical systems, Milwaukee, WI, USA) with a voxel size of $0.9375 \times 1.5 \times 0.9375$ and total image dimensions of $256 \times 124 \times 256$. In our study, we consider both the rescan image pairs and the baseline to 2-weeks image pairs to have a tissue similarity prior that can be used for regularization. From the original dataset, some images were discarded due to poor scan quality or movement artifacts; details on which image pairs were used for training and evaluation can be found in the supplementary material.

2.2. ADNI1 data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

In our work, we consider a subset of subjects originally included in the "ADNI1: Complete 1Yr 1.5T" standardized data collection and use similarly preprocessed scans with corrected gradient nonlinearity and B1 and N3 nonuniformity correction. In total, we consider 1063 scans from 105 AD patients and 145 healthy control subjects, having 250 pairs of baseline and 1 year of follow-up with 541 scan-rescan images taken at both timepoints. Within this cohort, scans were taken with varied voxel sizes ranging from $0.94 \times 0.94 \times 1.2$ to $1.3 \times 1.3 \times 1.2$. In total, 8 different scanners from 2 manufacturers (GE and Siemens) were used for image acquisition. In 45 of the 250 subjects, a different scanner model from the same manufacturer was reported for the 1-year follow-up scan. Details of the image pair IDs used for training and inference can be found in the supplementary material.

3. Methods

We present an unsupervised deep learning pipeline for segmentation-based brain volumetry that learns from tissue segmentations derived from fsl_anat while enforcing a tissue similarity regularization that improves longitudinal brain atrophy quantification. The proposed regularization exploits the assumption that two scans from the same subject taken within a short time interval should have similar brain tissue volumes. For training, we use a set of coregistered T1-w scan pairs having a tissue similarity prior and generate the segmentation targets in an unsupervised manner using fsl_anat. In inference, the pipeline performs end-to-end skull stripping and brain tissue segmentation from a single image in its native space, i.e., without image interpolation. For longitudinal evaluation, each image is independently segmented first, and then measures of change are computed. In the following sections, we describe in detail how to prepare the unsupervised training data and the architecture of our deep learning framework, along with the procedures for network training and image inference.

3.1. Training data preparation

The pipeline is trained from a set of T1-w scan pairs belonging to the same subject and acquired within a short interval, thus having a tissue similarity prior, from which we generate the reference brain tissue segmentations in a fully unsupervised manner, as shown in Figure 1. For this purpose, the fully automated fsl_anat anatomical image processing pipeline implemented in FSL is applied to each T1-w scan to perform skull stripping, as well as segmentation of brain tissue using FAST (Zhang et al., 2001) and deep gray matter structures using FIRST (Patenaude et al., 2011). This tissue segmentation procedure is very similar to that done by SIENA-XL (Battaglini et al., 2018), which also used the fsl_anat pipeline to generate the tissue and subcortical structure segmentation. More specifically, fsl_anat performs skull-stripping through a nonlinear registration to the MNI standard space, which is used to transform a dilated MNI brain mask back into the native space of the T1-w image. From this skull-stripped image, brain tissue probabilities are obtained using FAST (Zhang et al., 2001), which is run with the `--weakbias` option. Additionally, the deep gray matter of subcortical structures is segmented with the registration-based FIRST method (Patenaude et al., 2011). Similar to SIENA-XL (Battaglini et al., 2018), we merge the FIRST subcortical structure segmentation into the FAST tissue probabilities by setting them as pure gray matter, obtaining the final reference FAST + FIRST segmentation in the native space of each T1-w scan.

As part of the fsl_anat pipeline, we also obtain a transform to an MNI T1-w structural template with 2 mm resolution. We use the inverse of this transform to bring a 2 mm resolution MNI brain mask through nearest neighbor interpolation into the native T1-w scan space. This coarse brain mask is later used as a normalization mask to

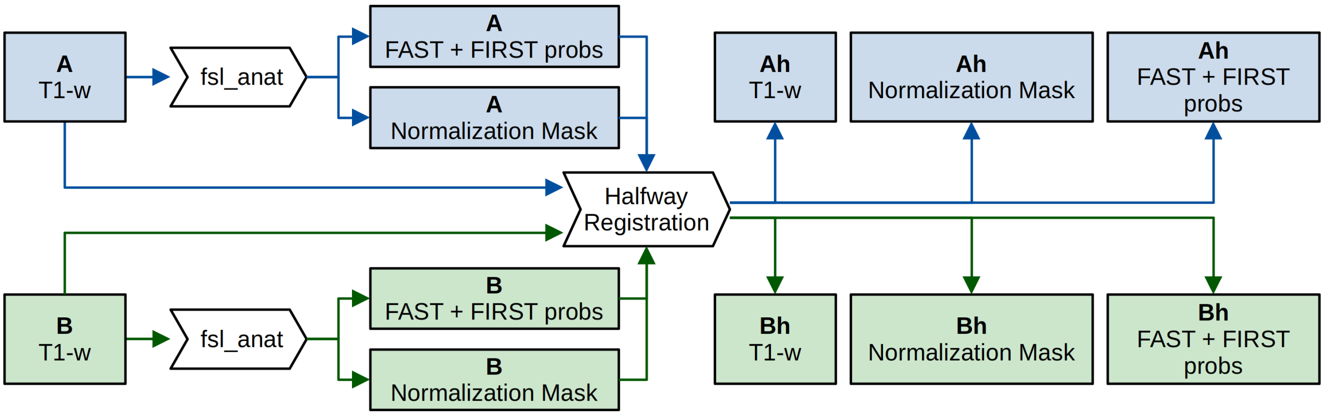


Figure 1: Training data preparation diagram for each pair of short-interval T1-w scans, A and B, which have tissue similarity prior. The T1-w scans are first processed using the fsl_anat anatomical image processing pipeline to obtain the reference FAST + FIRST brain tissue probabilities and input normalization mask for each of them. Then, the T1-w scans and their segmentations are spatially aligned by linear registration to a halfway space, Ah and Bh, between them.

constrain the computation of image statistics to the brain tissues for performing the input normalization of our deep learning system.

Finally, each pair of similar T1-w scans is spatially aligned to be able to exploit their tissue similarity prior during training. This goal is achieved by performing a linear registration to a halfway space between them using the mri_robust_register method (Reuter and Fischl, 2011) implemented in the FreeSurfer image analysis suite with default parameters, using cubic interpolation to transform the images. Then, the reference FAST + FIRST tissue probabilities are also transformed through linear interpolation into the halfway space, along with the normalization mask, which is transformed using nearest neighbor interpolation. Note that the halfway registered T1-w scans and tissue probabilities are exclusively used during training, while image inference for evaluation is performed in the native space of each scan without any type of interpolation.

3.2. Deep learning pipeline

We utilize a patch-based deep learning pipeline using a residual 3D architecture based on the U-Net (Ronneberger et al., 2015), which performs both skull stripping and brain tissue segmentation from a single T1-w scan. As input, the network receives a single 3D patch with spatial dimensions of $32 \times 32 \times 32$ and outputs a brain tissue probability distribution among four classes (background, CSF, GM and WM) for each input voxel. The selected patch size of $32 \times 32 \times 32$ provides sufficient context for accurate segmentation while balancing class representation and improving training stability through the use of a larger batch size. The network architecture, depicted in Figure 2, consists of a 3D U-Net model that uses residual convolution blocks and skip connections. All the convolutional layers use $3 \times 3 \times 3$ kernels and are always preceded, except for the input and output nodes, by a batch normalization (BN) layer (Ioffe and Szegedy, 2015) and a parametric rectified linear unit (PReLu) activation (Nair and Hinton, 2010). The parameter distribution is asymmetrical, with the residual blocks of the encoder part using two convolutional layers while a single one is used in the decoder. The network uses four different resolution levels, where the feature maps are downsampled by $2 \times 2 \times 2$ in each level of the encoder and upsampled back by the same factor in the decoder. Downsampling is performed by concatenating the result of a max pooling operation and strided convolution as proposed by Szegedy et al. (2016), while upsampling is performed using a transposed convolution that learns the upsampling operator for each feature map. The last layer outputs a four-channel patch with the same $32 \times 32 \times 32$ spatial size as the input and is activated with a softmax to obtain a probability distribution among the background and three considered tissue classes.

Before extracting patches for either training or inference, we normalize the T1-w image intensities to standardize the input range and reduce the influence of outliers. More specifically, the intensity range is winsorized within the 0.05% and 99.95% percentiles and then the minimum and maximum intensities are mapped to the [-1, 1] interval. To avoid influence from intensities not belonging to the brain tissues, image statistics are computed exclusively within a normalization mask, which is a coarse brain mask obtained through linear registration from a 2 mm resolution T1-w MNI template.

3.2.1. Training procedure

To train the proposed pipeline, we use the prepared halfway registered T1-w scans and their corresponding FAST + FIRST brain tissue probabilities derived from fsl_anat as the segmentation target. From these halfway registered scans, a patch set is generated with 100,000 pairs of samples, 85,000 for training and 15,000 for validation, extracted from the same spatial location of each halfway registered pair. The same number of patches is extracted from each of the available pairs with a deliberate sampling strategy to balance the representation of segmentation

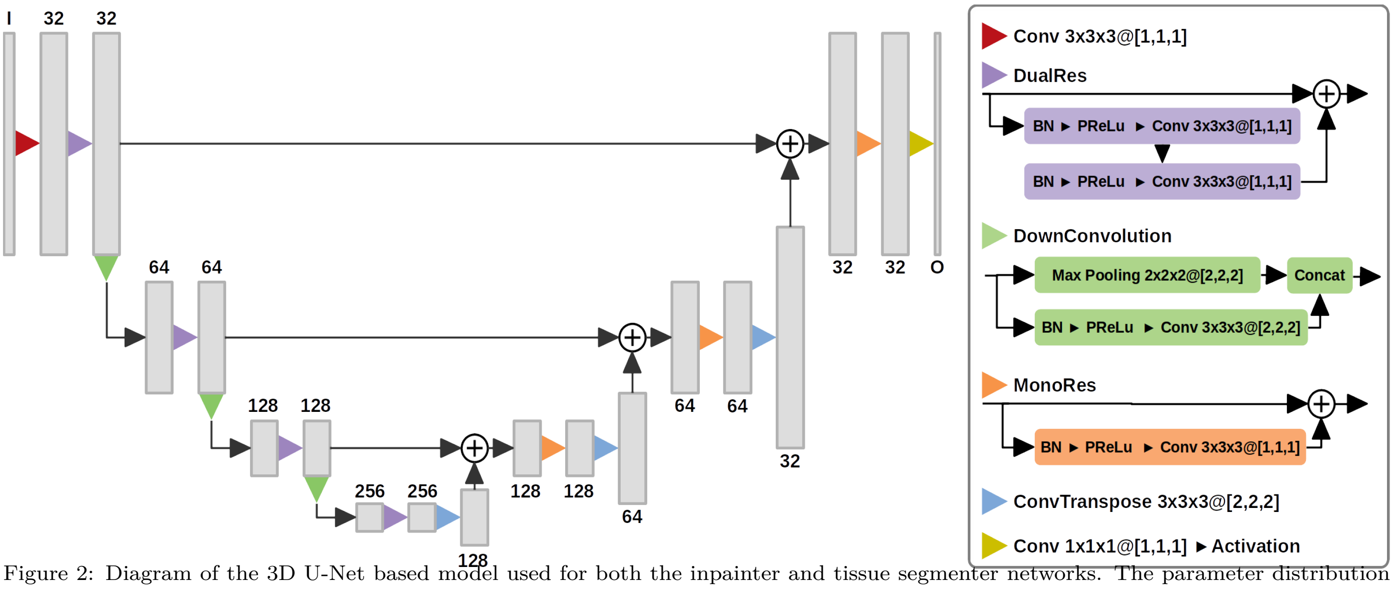

I
32
32
64
64
128
128
256
256
128
128
128
64
64
64
32
32
32
O
Conv 3x3x3@[1,1,1]
DualRes
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
DownConvolution
Max Pooling 2x2x2@[2,2,2]
Concat
BN ▸ PReLu ▸ Conv 3x3x3@[2,2,2]
MonoRes
BN ▸ PReLu ▸ Conv 3x3x3@[1,1,1]
ConvTranspose 3x3x3@[2,2,2]
Conv 1x1x1@[1,1,1] ▸ Activation

Figure 2: Diagram of the 3D U-Net based model used for both the inpainter and tissue segmenter networks. The parameter distribution is asymmetrical, with the residual blocks of the encoder using two convolutional blocks, while a single one is used in the decoder. In the convolutional layers (Conv), $\mathrm{K_x{\times}K_y{\times}K_z@[S_x,S_y,S_z]}$ indicates the kernel and stride dimensions on each axis. The gray boxes represent the feature maps with the number of channels indicated above or below it. The numbers of input and output feature maps are denoted by I and O, respectively.

classes. For this purpose, we use the FAST + FIRST segmentations derived from fsl_anat as a guide to extract 25% of patches centered on CSF, 25% on GM, 25% on WM, 20% on the head and 5% from the background. To obtain a rough approximation of the nonparenchyma voxels, we define the head class as any nontissue voxel with a T1-w intensity greater than the mean of the image and the background class as any nontissue voxel with an intensity less than the mean. Additionally, a random 3D offset of up to half the patch size is applied to each sampled patch to increase the representation of class boundaries.

Once the training patch set is built, the randomly initialized network weights are iteratively trained following the procedure depicted in Figure 3. Each training iteration consists of two separate forward passes through the network, obtaining a dense prediction for each halfway registered T1-w patch and a single backward pass that is used to update the network weights to minimize the loss function. In practice, each iteration is performed on a batch of 16 patch pairs, so that we first forward pass each of the 16 patch pairs and then perform a single backward pass from the average of their loss values. The network weights are updated through the Adadelta optimizer (Zeiler, 2012) with a learning rate of 0.05. To prevent overfitting, early stopping is performed when the loss on the validation set does not improve for 8 consecutive epochs.

As shown in Figure 3, the training loss function comprises the sum of three terms: two of them come from segmentation loss terms, one for each T1-w patch, and the third is a shared similarity loss term that enforces the tissue similarity regularization during training. The probabilistic version of the cross-entropy loss (PCE) is used as the segmentation loss, targeting the partial volume probabilities of the FAST + FIRST segmentation derived from fsl_anat. Using probabilities as targets, instead of categorical labels, we encourage approximating the partial volume probabilities instead of attempting to maximize the probability of the most likely tissue class. More specifically, given a predicted probability distribution of a patch P over C classes with dimensions $C \times X \times Y \times Z$ and a target probability distribution T of the same dimensions, the probabilistic cross-entropy segmentation loss term is defined as:

$$\mathcal{L}_{seg}(P,T) = \frac{1}{XYZ}\sum_{x,y,z}\sum_{c_i=0}^{C-1} - P(c_i,x,y,z)\cdot T(c_i,x,y,z)\cdot \cdot \ln\left(\sum_{c_j=0}^{C-1}\exp[P(c_j,x,y,z)]\right) \tag{1}$$

The similarity loss term is taken as the sum of the L1 norm between the CSF, GM and WM percentages of the two predicted patches. More specifically, given two output probability distributions, P_a and P_b, over C classes with dimensions $C \times X \times Y \times Z$, the similarity loss term is defined as:

$$\mathcal{L}_{sim}(P_a,P_b) = \sum_{c=1}^{C-1}\frac{100}{XYZ}\left|\sum_{x,y,z}P_a(c,x,y,z) - \sum_{x,y,z}P_b(c,x,y,z)\right| \tag{2}$$

Note that the two patches P_a and P_b used in the similarity loss term are forward passed separately so that the model cannot extract joint features between the short-interval scans to reduce the volume differences. In this way, the model is constrained to the use of cross-sectional features acting on a single patch to achieve this reduction. As a result, we obtain a model that performs inference on a single

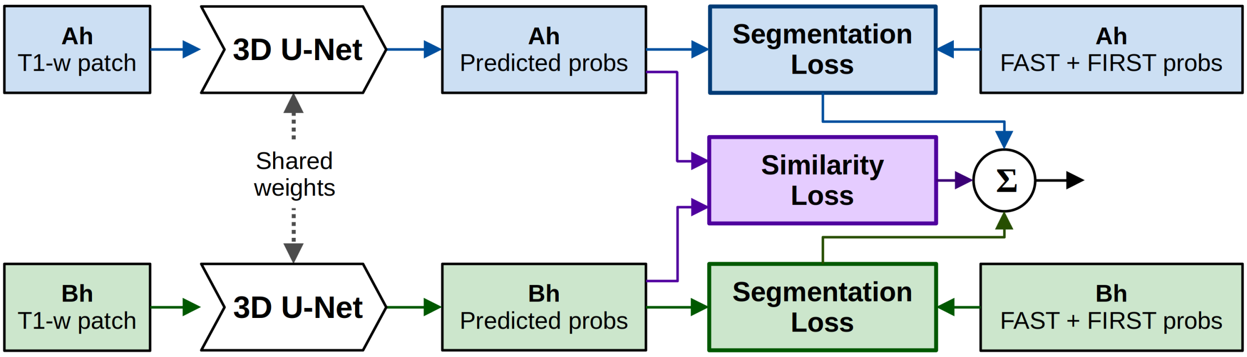


Figure 3: Training iteration diagram of the proposed pipeline. The input comprises two patches extracted from a pair of halfway registered T1-w scans having a tissue similarity prior. As a segmentation target, we use the FAST + FIRST brain tissue probabilities computed in the native space of each T1-w scan and transformed to the halfway space. The two T1-w patches are predicted in two independent forward passes through the network, and two separate patch predictions are obtained. Then, a single backward pass is performed that updates the network weights to minimize the two segmentation loss terms, as well as the shared similarity loss term, which enforces the tissue similarity regularization.

image at a time but does so with reduced quantification error thanks to the training regularization.

In summary, given two predicted tissue probability distributions P_a and P_b and their corresponding target probability distributions T_a and T_b, respectively, the loss function is defined as:

$$\begin{aligned}\mathcal{L}(P_a, P_b, T_a, T_b) = \mathcal{L}_{seg}(P_a, T_a) + \mathcal{L}_{seg}(P_b, T_b) + \\ + w_{\text{sim}} \cdot \mathcal{L}_{sim}(P_a, P_b)\end{aligned} \tag{3}$$

where w_{sim} is a term that modulates the degree to which the model will be allowed to deviate from approximating the reference segmentation probabilities and instead focus on reducing tissue volume differences between similar patches. Setting $w_{\text{sim}} = 0.0$ would set the optimization target purely on approximating the target tissue probabilities as faithfully as possible, and any deviation from the target would be penalized by the segmentation loss terms. However, to avoid learning the biases and errors of the reference method, a level of disagreement is needed with respect to the target segmentations to allow room for improvement. By increasing the value of w_{sim}, we progressively shift the optimization target away from approximating the target probabilities and toward reducing the segmentation differences between short-interval scans. However, if w_{sim} is set too high, the learned segmentation model would be allowed to excessively ignore the FAST + FIRST segmentations to the point at which it might produce anatomically unfeasible results. For this reason, the preferred value for w_{sim} is the smallest one that provides sufficient improvement in brain atrophy quantification. The effect on segmentation accuracy and atrophy quantification of the proposed regularization is analyzed later in Section 5.1.

3.2.2. Image inference

Within the proposed pipeline, image inference performing end-to-end skull stripping and brain tissue segmentation is performed on a single T1-w scan in its native space, i.e., without image interpolation. First, input normalization is performed on the T1-w image as previously described within a normalization mask obtained by linear transformation of a brain mask from a 2 mm resolution MNI template. Then, highly overlapping patches of size 32 × 32 × 32 are extracted for inference at regular spatial steps of 10 × 10 × 10. This level of overlap helps to reduce block boundary artifacts and improve spatial coherence. Before patch extraction, the T1-w image is edge padded on all sides by 16 voxels, which is half the patch size, to ensure that every voxel in the image is predicted with a similar degree of overlap. The extracted patches are then forward passed through the trained segmentation model, obtaining dense tissue probability distributions for each patch. The use of overlapping patches results in several brain tissue probability distributions for each voxel of the input image. To achieve the final whole image segmentation, the overlapping predictions are averaged and normalized to produce a single brain tissue probability distribution for each input image voxel.

Additionally, the brain tissue probabilities are postprocessed to improve the accuracy in intracranial cavity segmentation. Since the proposed pipeline performs end-to-end skull stripping and brain tissue segmentation, there is no assumption made regarding which voxels should be pure tissue or pure background, leading to small background probabilities appearing inside the intracranial cavity and small probabilities of tissue appearing outside of the brain. To reduce these small errors from compounding onto large volume measurement errors, postprocessing is performed based on the assumption that the intracranial cavity will be the largest connected component in the output segmentation. In practice, we first define a *pure tissue* mask as $p(\text{CSF}) + p(\text{GM}) + p(\text{WM}) > 0.99$, which is processed using morphological operators by filling holes and then keeping only the largest connected component. Within the *pure tissue* mask, the background probability is set to zero, and the remaining tissue probabilities are normalized to ensure that they total one. Outside of the *pure tissue* mask, the background probability is set to one, and all tissue probabil-

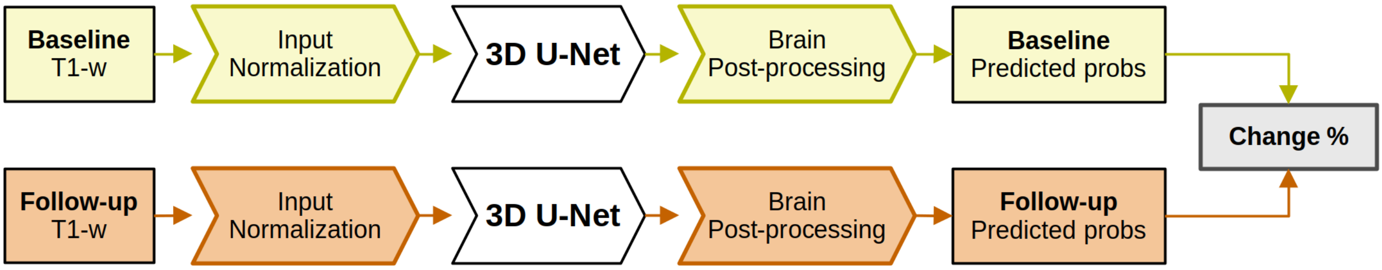


Figure 4: Longitudinal inference procedure. The baseline and follow-up images are independently segmented in their native space and change measures are computed from the predicted tissue probability distributions.

ities are set to zero. From these probabilistic segmentations, measures of volume for each tissue are obtained by taking the brain tissue probability distribution of each voxel as an estimation of its partial volume mixture. In this way, the volume of each tissue class is calculated by totaling its voxelwise probability across the whole image and then normalizing by the voxel size to obtain the volume in mm^3.

For longitudinal evaluation, inference is performed independently for the baseline and follow-up images, in their native space, and then measures of change are computed from the predicted tissue probability distributions, as shown in Figure 4.

3.3. Implementation details

The proposed method was implemented with Python using the Torch scientific computing framework (Paszke et al., 2017). All experiments were run on a GNU/Linux machine running the Ubuntu operating system, version 18.04, with 128 GB of RAM memory and an Intel ®Core $^{\text{TM}}$ i7-7800X CPU. The versions of the software packages used were 6.0.4 for FSL and 6.0.0 for FreeSurfer. The network training and inference were performed with an NVIDIA 1080 Ti GPU (NVIDIA Corp., United States) with 12 GB G5X memory. The proposed network architecture has 7 million trainable parameters and takes 3.6 GB of GPU memory during training and only 1.5 GB for inference. The time to perform inference for a whole image using the proposed pipeline within our system is between 2 and 3 minutes, depending on the image dimensions. The linear registration to obtain the normalization mask takes approximately 1 minute, while inference of an image using the GPU takes between 1 and 2 minutes.

4. Evaluation

To evaluate the proposed pipeline, we first studied the effect of tissue similarity regularization on brain volumetry by training the pipeline with increasing amounts of tissue regularization controlled by the w_{sim} parameter. More specifically, we trained seven instances of the pipeline for each dataset considering higher values for w_{sim}. From the results of this experiment, we set an optimal default value for w_{sim} and then performed a detailed quantitative analysis of the pipeline trained with the selected optimal weight.

The evaluation was performed on two publicly available MRI datasets with different characteristics, MIRIAD and ADNI1, which have short-interval and longitudinal imaging for healthy controls (HCs) and AD subjects. The MIRIAD dataset provides a small set of homogeneous MR scans taken with the same scanner and acquisition protocol, while the ADNI1 dataset has a larger number of subjects with more heterogeneous imaging acquired on different scanners and with varied voxel spacings. The proposed pipeline was independently trained on each dataset using the available short-interval scan pairs having a tissue similarity prior and preparing the data for training as described in Section 3.1. For the MIRIAD dataset, we used all 182 scan-rescan pairs, as well as 125 scan pairs made within 2 weeks, for a total of 307 training scan pairs. For the ADNI1 cohort, from the 541 available scan-rescan pairs, we considered a random subset of 298 pairs for training due to memory constraints on our system. Inference was then performed for the maximum interval scan pairs, i.e., the first and last available timepoints for each subject, which amounted to 69 longitudinal pairs for the MIRIAD dataset and 250 for ADNI1. We also performed inference on all of the short-interval scans, including those used for generating the training data, for evaluation and analysis purposes. While short-interval scans are registered to a halfway space for training, inference is performed in their native space without image interpolation.

We compared our results to the FAST + FIRST brain tissue probabilities derived from fsl_anat used as training targets and with SIENA, a well-known and widely used state-of-the-art brain atrophy quantification method also implemented in FSL. In practice, SIENA is run with the `-R` option in MIRIAD, which iterates the skull stripping several times to robustly estimate the brain center, and with the `-B` option in ADNI1, which removes the neck present on the images.

Measures of whole-brain atrophy in segmentation-based methods are typically based on the volume change of brain parenchyma between the baseline and follow-up segmentations, which can be computed either from raw or from normalized volumes. Additionally, since our pipeline not only segments the parenchyma but also its distinct gray and white matter components, we also provide individualized measures of change for these tissues. To account for different time intervals between longitudinal scans of different subjects, all reported measures of change are annualized. Relative change measures are not computed relative to the baseline or follow-up volumes, but instead we do so with respect to their average as follows:

$$\text{Change \%} = 100 \cdot \frac{2(V_{\text{follow-up}} - V_{\text{baseline}})}{V_{\text{follow-up}} + V_{\text{baseline}}} \tag{4}$$

where V can be any measure of volume derived from the probabilistic brain tissue segmentations. The percentage of brain volume change (PBVC) is obtained when V is set as the raw volume of the brain parenchyma. Similarly, setting V as the raw volume of the GM or WM provides the percentage of GM volume change (PGMVC) or percentage of WM volume change (PWMVC), respectively. However, measures of change based on raw unscaled volumes are affected by a number of technical and physiological confounding factors (Sastre-Garriga et al., 2017). A more robust measure of change can be obtained using tissue fractions, which are computed by normalizing the raw tissue volumes with respect to the intracranial volume (ICV), computed as the sum of all tissue volumes (CSF + GM + WM). In this way, the brain parenchymal fraction (BPF), gray matter fraction (GMF) and white matter fraction (WMF) are obtained by normalizing their respective raw volume measurements by the intracranial volume. Additionally, to study the longitudinal skull stripping consistency of the proposed pipeline, the ICV change is also measured by setting V as the raw intracranial volume. Although the ICV has been shown to decrease as a result of aging beyond adulthood (Royle et al., 2013), this change is expected to be close to zero within the time intervals between scans of the considered datasets.

These measures of volume change are also computed for the short-interval scan pairs to evaluate the quantification error. Between these images, an ideal atrophy quantification method should measure zero change between them; therefore, we consider any deviation from zero as quantification error.

In the absence of an atrophy ground truth, the measures of change by themselves are not indicative of the accuracy or quality of the brain tissue segmentations. However, the sensitivity of an atrophy quantification method to longitudinal changes can be assessed by quantifying the differences between two subject populations known to have different rates of change. In this way, atrophy quantification methods can be compared based on the assumption that better methods would detect larger and more pronounced differences between these two populations. In our case, we quantified differences between HC and AD subjects based on their annualized change measures. As in the work of Smith et al. (2007), a measure of discriminative power can be obtained from the t statistic of Welch's unequal variances test, which quantifies confidence in the existence of differences between both groups. A large t provides a high level of evidence that the observed differences between the two populations are statistically significant —in other words, a low probability that the observed differences could be due to chance. However, t does not reflect the strength or size of these differences; for instance, a large t could be obtained for a very small difference in the magnitude of annualized change, which would not necessarily be of any practical significance or clinical importance. For this purpose, measures of effect size are typically used to quantify the magnitude or strength of observed differences. More specifically, we use Cohen's d to measure the effect size, which is calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \tag{5}$$

where s, the pooled standard deviation, is defined for two independent populations as:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{6}$$

where n_1 and n_2 are the number of samples in each population, and s_1^2 and s_2^2 are the variances of each group, computed as:

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1,i} - \bar{x}_1)^2 \\ s_2^2 &= \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2,i} - \bar{x}_2)^2 \end{aligned} \tag{7}$$

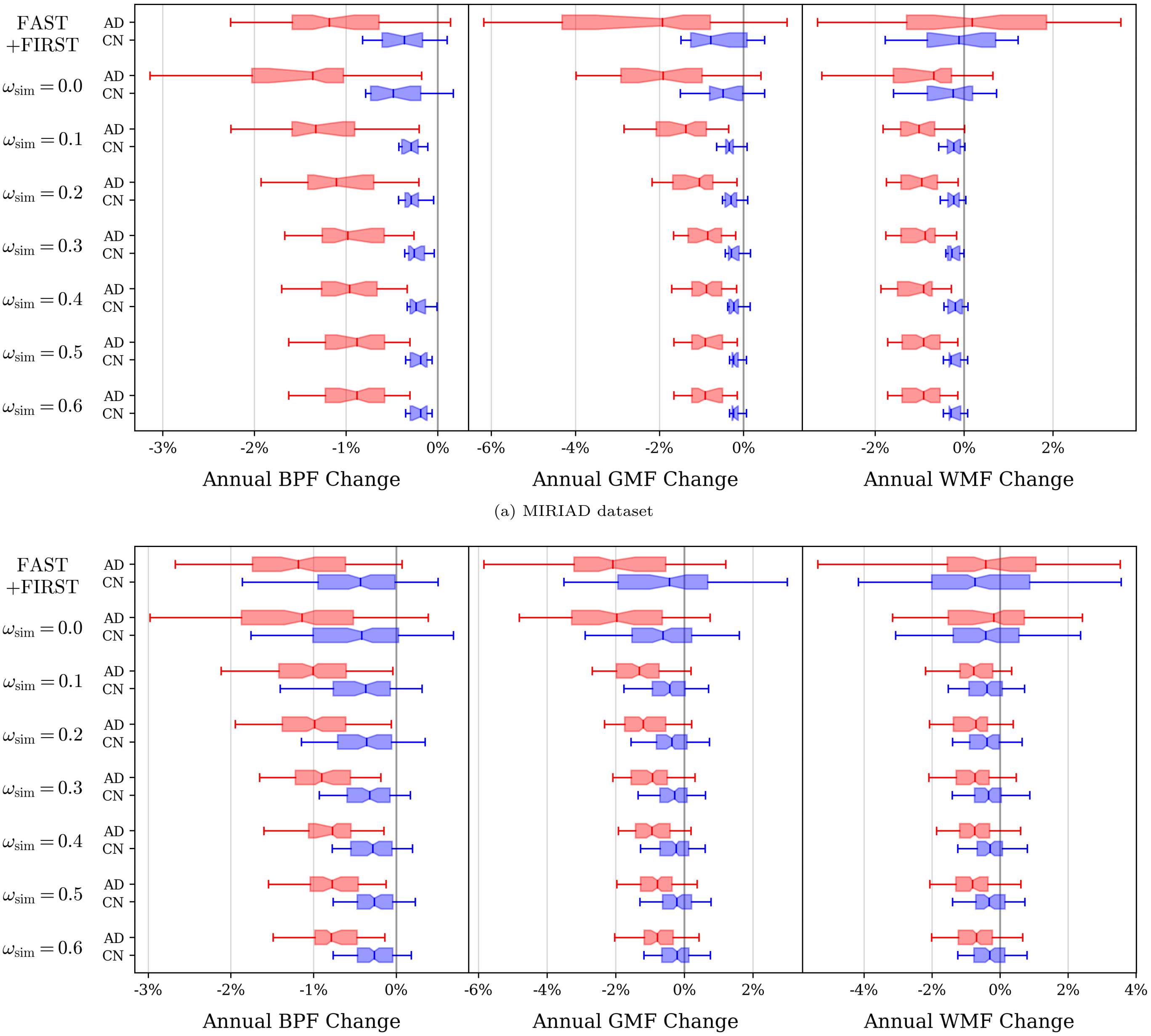
We also calculate the Dice similarity coefficient (DSC) between the segmentations of the proposed pipeline with respect to the FAST + FIRST reference derived from fsl_anat to quantify the extent to which the tissue similarity regularization shifts the segmentation away from the target. In practice, we calculate the DSC by first taking the argmax of the brain tissue probabilities to obtain a categorical multiclass segmentation.

5. Results and discussion

5.1. Similarity weight analysis

In this section, we study the effect of the tissue similarity regularization on the proposed pipeline by training several instances with increasing values for w_{sim}, from 0.0 to 0.6. To analyze the effect on brain atrophy quantification properties, we studied the response of annualized change measures as well as any improvement in the sensitivity to differences between healthy and AD subjects of these change measures. We also study how regularization affects the tissue segmentation model by calculating short interval error measures, as well as the DSC with respect to the reference FAST + FIRST segmentations used for training.

Figures 5a and 5b show boxplots of annualized change measures of BPF, GMF and WMF between the maximum interval pairs of the MIRIAD and ADNI1 datasets, respectively. Overall, increasing values of w_{sim} reduce the stan-



(a) MIRIAD dataset

(b) ADNI1 dataset

Figure 5: Annualized change measures between maximum interval pairs of healthy controls (HC) and Alzheimer's disease (AD) patients for the FAST + FIRST reference segmentations and the proposed pipeline with increasing similarity regularization weight. The boxes representing the interquartile range are notched within the confidence interval around the median, with the left and right whiskers set to the 5th and 95th percentiles, respectively.

dard deviation of all considered change measures in both datasets. The reduction in variability is more pronounced in the MIRIAD measures, especially for the healthy subjects, most likely due to the high similarity between images acquired with the same scanner and imaging protocol. In both datasets, higher values of w_{sim} increase the median BPF and GMF change, while the median WMF change is decreased.

The effect of increasing regularization on the discriminative power and effect size of change measures between healthy and AD subjects is summarized in Table 1. The results show that tissue similarity regularization improves the discrimination and effect size between groups in all change measures in both the MIRIAD and ADNI1 datasets. In the MIRIAD dataset, the proposed pipeline with $w_{\text{sim}} = 0.0$ already improves the sensitivity of BPF, GMF and WMF change compared to the reference FAST + FIRST segmentations. When the regularization is enforced, increasing the similarity weight value further improves the results until $w_{\text{sim}} = 0.4$, where the improvement reaches its peak, and beyond this point, higher values actually worsen the differences between groups. In the ADNI1 dataset, the proposed

Table 1: Discrimination and effect size of annualized change measures between healthy controls (HC) and Alzheimer's disease (AD) subjects for the maximum interval scan pairs. The discriminative power is measured with the t statistic from Welch's unequal variances test, while the effect size is measured using Cohen's d.

	MIRIAD						ADNI1					
	ΔBPF		ΔGMF		ΔWMF		ΔBPF		ΔGMF		ΔWMF	
Method	t	d	t	d	t	d	t	d	t	d	t	d
FAST + FIRST	4.66	1.01	3.35	0.71	-0.49	0.10	6.19	0.80	5.29	0.69	-0.56	0.07
$w_{\text{sim}} = 0.0$	6.29	1.31	5.10	1.12	2.05	0.44	5.38	0.72	5.97	0.78	0.47	0.06
$w_{\text{sim}} = 0.1$	9.07	1.76	7.93	1.59	7.71	1.52	8.28	1.11	7.64	1.01	4.72	0.63
$w_{\text{sim}} = 0.2$	9.05	1.73	7.14	1.38	7.54	1.47	8.63	1.16	8.27	1.09	5.58	0.74
$w_{\text{sim}} = 0.3$	9.47	1.82	7.65	1.56	8.72	1.68	9.84	1.34	8.43	1.11	7.35	0.98
$w_{\text{sim}} = 0.4$	9.85	1.89	8.05	1.63	9.31	1.85	10.14	1.39	9.13	1.23	7.88	1.06
$w_{\text{sim}} = 0.5$	9.76	1.89	8.43	1.65	8.29	1.64	10.21	1.40	9.14	1.21	7.92	1.07
$w_{\text{sim}} = 0.6$	9.22	1.83	7.17	1.50	7.99	1.61	10.37	1.43	9.46	1.27	8.03	1.08

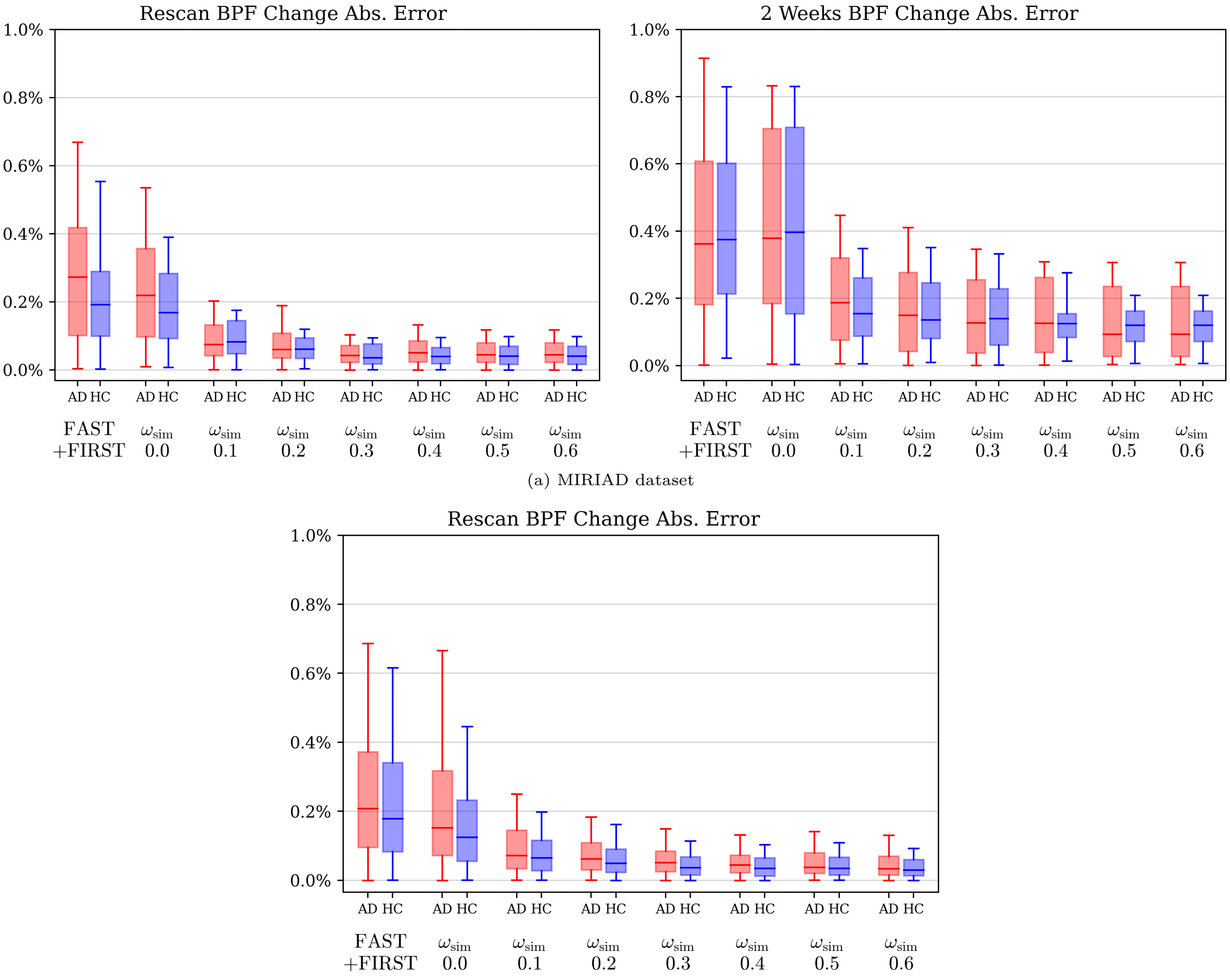
Table 2: DSC (mean ± std. dev. %) of the maximum interval pair segmentations between FAST + FIRST and the proposed pipeline with increasing tissue similarity regularization weights.

	MIRIAD			ADNI1		
Method	GM+WM	GM	WM	GM+WM	GM	WM
$w_{\text{sim}} = 0.0$	$96.9 \pm 1.2\%$	$88.6 \pm 3.0\%$	$92.4 \pm 2.1\%$	$97.7 \pm 0.8\%$	$90.5 \pm 1.9\%$	$94.3 \pm 1.1\%$
$w_{\text{sim}} = 0.1$	$97.1 \pm 0.7\%$	$88.5 \pm 2.8\%$	$91.7 \pm 2.1\%$	$97.6 \pm 0.8\%$	$90.3 \pm 2.0\%$	$94.1 \pm 1.2\%$
$w_{\text{sim}} = 0.2$	$96.9 \pm 0.7\%$	$87.6 \pm 2.9\%$	$90.8 \pm 2.2\%$	$97.5 \pm 0.9\%$	$89.3 \pm 2.2\%$	$93.1 \pm 1.5\%$
$w_{\text{sim}} = 0.3$	$96.9 \pm 0.8\%$	$87.4 \pm 2.9\%$	$90.9 \pm 2.2\%$	$97.3 \pm 0.9\%$	$88.4 \pm 2.3\%$	$92.5 \pm 1.5\%$
$w_{\text{sim}} = 0.4$	$96.8 \pm 0.8\%$	$87.2 \pm 3.1\%$	$90.3 \pm 2.4\%$	$97.2 \pm 0.9\%$	$88.2 \pm 2.3\%$	$92.4 \pm 1.5\%$
$w_{\text{sim}} = 0.5$	$96.8 \pm 0.8\%$	$86.7 \pm 3.1\%$	$90.0 \pm 2.3\%$	$97.0 \pm 1.0\%$	$87.2 \pm 2.5\%$	$91.8 \pm 1.6\%$
$w_{\text{sim}} = 0.6$	$96.4 \pm 0.9\%$	$85.9 \pm 3.2\%$	$89.6 \pm 2.3\%$	$97.0 \pm 0.9\%$	$87.7 \pm 2.4\%$	$92.2 \pm 1.5\%$

pipeline without regularization ($w_{\text{sim}} = 0.0$) improves both the GMF and WMF change sensitivity while having a worse effect on BPF change compared with the reference FAST + FIRST segmentations. When the regularization is enforced, the sensitivity in all three measures steadily improves for higher values of w_{sim}. In contrast to the MIRIAD results, the sensitivity of ADNI1 measures does not peak at $w_{\text{sim}} = 0.4$, but beyond this point, improvement gains decrease rapidly.

Table 2 shows DSC measures between the reference FAST + FIRST segmentations and the proposed pipeline with increasing w_{sim} values. The reported DSC results are calculated from the argmax classification of the probabilistic brain tissue segmentations and are given separately for parenchyma (GM+WM) as well as for its GM and WM components. As expected, higher amounts of regularization decrease the similarity with respect to the reference FAST + FIRST brain tissue segmentations. Moreover, it can also be observed that the DSC of GM and WM components that form the parenchyma decrease much more rapidly with increasing regularization than those of the parenchyma itself. This outcome suggests that the dissimilarity is due to a redistribution of probabilities between GM and WM classes. By increasing w_{sim}, the learning focus is progressively shifted away from approximating the FAST + FIRST probabilities, and a greater degree of deviation is allowed to reduce the segmentation differences between short-interval scans.

Figures 6a and 6b show the absolute BPF change error between short-interval scans of the two considered datasets. A reduction in short interval error is to be expected since the metrics are calculated on the same short-interval pairs used for training; however, the aim is to illustrate how and to what extent the segmentation is affected. Without regularization ($\omega_{\text{sim}} = 0.0$), the proposed pipeline exhibits levels of error similar to those of of the reference FAST + FIRST segmentations. Even when the smallest amount of regularization is enforced ($\omega_{\text{sim}} = 0.1$), the error is greatly reduced, with higher weights providing smaller improvements thereafter. Figure 6a also shows that, despite both the rescan and 2-week pairs of the MIRIAD dataset being used for training, the rescan pairs without repositioning show a much greater reduction in error than the 2-week pairs. This outcome suggests that the scan differences due to repositioning and/or small time intervals are larger than those of the rescan images. Surprisingly, even when directly using the 2-week pairs with repositioning to regularize the training, the effect size of differences between AD and HC groups is not much different than that of ADNI1, which only used rescan images without repositioning for



(a) MIRIAD dataset

(b) ADNI1 dataset

Figure 6: Short interval error

regularization. These results suggest that the proposed regularization does not especially benefit from images with repositioning to improve in the longitudinal case.

The experiments performed have shown that the proposed regularization can improve the sensitivity of atrophy measures to differences between healthy and AD subjects. However, these improvements are obtained at the cost of decreasing the segmentation similarity with respect to the reference FAST + FIRST segmentations. In our experiments, the sensitivity to differences between groups reached its maximum at $w_{\text{sim}} = 0.4$ for the MIRIAD dataset while reaching a point of diminishing returns at $w_{\text{sim}} = 0.4$ for the ADNI1 dataset. Thus, we decided on $w_{\text{sim}} = 0.4$ as an optimal default value for the proposed pipeline, providing the most improvement for the least deviation from the reference segmentations.

5.2. *Longitudinal atrophy quantification analysis*

In the previous section, we studied the effect of varying degrees of tissue similarity regularization on the presented deep learning pipeline for brain tissue segmentation in both single-site and multisite datasets. In this section, we now perform a detailed quantitative and qualitative evaluation of the presented pipeline trained with $w_{\text{sim}} = 0.4$, the empirically selected optimal regularization weight.

5.2.1. *Intracranial volume change*

The results for ICV measurements of the selected models trained with $w_{\text{sim}} = 0.4$ can be found in Table 3. On average, the absolute ICV change of the proposed pipeline is lower in both datasets than for the fsl_anat reference, suggesting a more consistent intracranial volume between longitudinal scans. In terms of ICV change, the brain masks from fsl_anat show a similarly negative rate in the MIRIAD dataset for both HC and AD subjects, while

Table 3: Annualized measures of ICV change (mean ± std. dev.) for the proposed and fsl_anat pipelines.

		MIRIAD		ADNI1	
		fsl_anat	Proposed ($w_{\text{sim}} = 0.4$)	fsl_anat	Proposed ($w_{\text{sim}} = 0.4$)
$\|\Delta\text{ICV}\|$	HC	0.38 ± 0.61 %	0.25 ± 0.32 %	0.80 ± 1.49 %	0.43 ± 0.69 %
	AD	0.67 ± 0.91 %	0.55 ± 0.52 %	0.83 ± 1.42 %	0.45 ± 0.30 %
ΔICV	HC	-0.23 ± 0.69 %	-0.06 ± 0.40 %	0.20 ± 1.68 %	-0.09 ± 0.81 %
	AD	-0.25 ± 1.11 %	-0.34 ± 0.68 %	0.30 ± 1.62 %	-0.26 ± 0.47 %

the ADNI1 dataset shows positive ICV changes for both groups, with a slightly higher average rate for the AD subjects. The proposed pipeline obtains a much more consistent ICV change between datasets and subject groups, having a small negative ICV change for the healthy subjects and a larger negative change for the AD subjects. The results suggest that the learned skull stripping of our pipeline is somehow affected by global atrophy since the longitudinal ICV change is negative and more pronounced for the AD group. This outcome would be caused by the way in which skull stripping is performed by fsl_anat, which nonlinearly registers a dilated brain mask to segment the brain parenchyma. In this way, instead of attempting to segment the entire intracranial cavity, fsl_anat essentially sets a fixed band around the parenchyma that does not encompass the entire intracranial cavity. Thus, in cases with greater amounts of atrophy, the fixed band around a more shrunken parenchyma means that there will be a larger amount of the intracranial cavity which will not be segmented by fsl_anat. Within the presented pipeline, tissue similarity regularization cannot reduce the learning of this fixed band bias, and the measured ICV is affected by the brain shrinkage observed on follow-up scans, which is higher for AD subjects than for healthy controls.

5.2.2. Short interval error

Within the performed evaluation, measures of short interval error are biased since our pipeline is explicitly and directly trained to reduce differences between the same short-interval scan pairs that would be used for evaluation. However, there is a set of 243 rescan image pairs from ADNI1, 139 pairs from healthy control subjects and 104 from AD subjects that were not seen during training and from which meaningful measures of short interval error can be obtained. Table 4 shows the short interval error for raw and normalized measures of change in these ADNI1 rescan pairs not used during training. Compared with the reference FAST + FIRST results and those of SIENA, the short interval error of our pipeline is significantly lower in all measures ($p < 10^{-6}$) and with much lower variability. Moreover, while the reference FAST + FIRST segmentations have much greater error for individual GM and WM tissue than for the parenchyma, in the proposed pipeline, the error is much more similar between the parenchyma and its GM and WM components. These results show that tissue similarity regularization not only reduces the quantification error of the pipeline but also increases the consistency of GM and WM volumes between short-interval scans.

5.2.3. Annualized atrophy rates

Table 5 shows the annualized rates of ΔBPF, ΔGMF, ΔWMF, PBVC, PGMVC and PWMVC of all maximum interval pairs for the reference FAST + FIRST segmentations and the proposed pipeline trained with $w_{\text{sim}} = 0.4$. In general, our pipeline shows much less variability in all measures of change than the FAST + FIRST reference segmentations. Compared with the results of FAST + FIRST, the annualized PBVC in the MIRIAD dataset is slightly reduced, especially for the HC subjects, while it is slightly increased in the ADNI1 dataset for both subject groups. In terms of BPF changes, the average rate of the proposed pipeline is reduced in both datasets compared to the reference FAST + FIRST segmentations. This finding would be mostly explained by the generally smaller ICV obtained by our pipeline for the follow-up scans, especially for the AD subjects, which slightly biases the follow-up tissue fractions toward larger values and reduces the apparent atrophy rate.

It can also be observed that the WMF change, as measured from the FAST + FIRST segmentations suggests that healthy controls have greater WM atrophy than AD subjects. In contrast, our pipeline shows greater WM atrophy for the AD subjects than for the healthy controls, which makes more intuitive sense in the context of a generalized brain atrophy process. As seen in Figure 5, the amount of regularization is directly related to the lowering of the median WMF change, suggesting that the segmentation of WM is directly improved by tissue similarity regularization.

For comparison, we also calculated the annualized atrophy rates with SIENA (Smith et al., 2002). Our results in the MIRIAD dataset (HC: -0.26 ± 0.43%; AD: -1.31 ± 0.86%) showed reduced average rates for both groups, with a pronounced reduction of variability for the AD group, compared to those of SIENA (HC: -0.53 ± 0.45%; AD: -2.43 ± 1.34%). In the ADNI1 dataset, the annualized PBVC of our pipeline (HC: -0.41 ± 0.92%; AD: -1.14 ± 0.76%) also shows lower average rates when compared to SIENA (HC: -0.61 ± 0.75%; AD: -1.85 ± 1.13%).

Table 4: Short interval error (mean ± std. dev. (median)) for raw and normalized change measures on the 243 rescan pairs from ADNI1 that were not seen during training of the presented pipeline. In the case of SIENA, which uses the BSI method, PBVC is the only provided measure of atrophy.

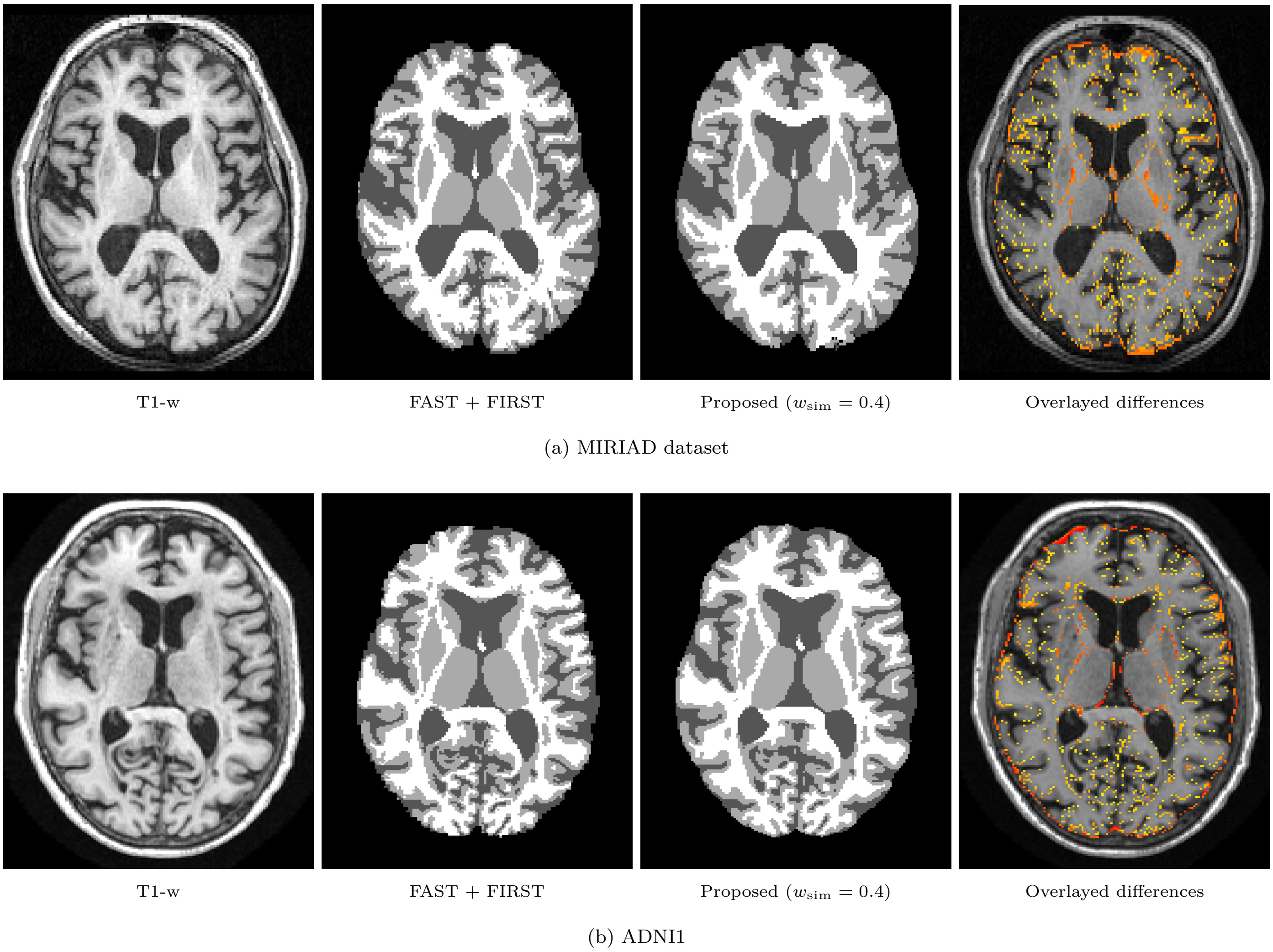
		FAST + FIRST	Proposed ($w_{\text{sim}} = 0.4$)	SIENA
PBVC	HC	$0.43 \pm 0.88\%$ (0.21%)	$0.16 \pm 0.71\%$ (0.06%)	$0.33 \pm 0.39\%$ (0.20%)
	AD	$0.43 \pm 0.66\%$ (0.23%)	$0.12 \pm 0.20\%$ (0.07%)	$0.38 \pm 0.53\%$ (0.19%)
PGMVC	HC	$1.17 \pm 1.38\%$ (0.71%)	$0.17 \pm 0.69\%$ (0.06%)	N/A
	AD	$1.35 \pm 1.34\%$ (0.92%)	$0.14 \pm 0.19\%$ (0.09%)	N/A
PWMVC	HC	$1.35 \pm 1.59\%$ (0.82%)	$0.21 \pm 0.77\%$ (0.08%)	N/A
	AD	$1.32 \pm 1.43\%$ (0.79%)	$0.14 \pm 0.22\%$ (0.08%)	N/A
ΔBPF	HC	$0.31 \pm 0.46\%$ (0.19%)	$0.06 \pm 0.12\%$ (0.03%)	N/A
	AD	$0.40 \pm 0.61\%$ (0.21%)	$0.09 \pm 0.16\%$ (0.05%)	N/A
ΔGMF	HC	$1.09 \pm 1.29\%$ (0.63%)	$0.08 \pm 0.21\%$ (0.04%)	N/A
	AD	$1.19 \pm 1.25\%$ (0.80%)	$0.11 \pm 0.17\%$ (0.06%)	N/A
ΔWMF	HC	$1.26 \pm 1.38\%$ (0.75%)	$0.12 \pm 0.19\%$ (0.06%)	N/A
	AD	$1.44 \pm 1.48\%$ (0.84%)	$0.12 \pm 0.17\%$ (0.07%)	N/A

Table 5: Annualized measures of atrophy (mean ± std. dev.) from maximum interval scan pairs.

		MIRIAD		ADNI1	
		FAST + FIRST	Proposed ($w_{\text{sim}} = 0.4$)	FAST + FIRST	Proposed ($w_{\text{sim}} = 0.4$)
ΔBPF	HC	-0.35 ± 0.51 %	-0.20 ± 0.15 %	-0.52 ± 0.84 %	-0.32 ± 0.31 %
	AD	-1.13 ± 0.87 %	-0.96 ± 0.48 %	-1.23 ± 0.92 %	-0.88 ± 0.50 %
ΔGMF	HC	-0.46 ± 1.50 %	-0.19 ± 0.22 %	-0.51 ± 2.12 %	-0.28 ± 0.42 %
	AD	-2.21 ± 2.83 %	-0.90 ± 0.51 %	-2.07 ± 2.41 %	-0.88 ± 0.57 %
ΔWMF	HC	-0.21 ± 1.25 %	-0.20 ± 0.21 %	-0.54 ± 2.41 %	-0.35 ± 0.42 %
	AD	0.02 ± 2.61 %	-1.05 ± 0.53 %	-0.36 ± 2.53 %	-0.87 ± 0.58 %
PBVC	HC	-0.58 ± 0.74 %	-0.26 ± 0.43 %	-0.32 ± 1.46 %	-0.41 ± 0.92 %
	AD	-1.38 ± 1.35 %	-1.31 ± 0.86 %	-0.93 ± 1.45 %	-1.14 ± 0.76 %
PGMVC	HC	-0.69 ± 1.39 %	-0.26 ± 0.44 %	-0.31 ± 2.27 %	-0.37 ± 0.87 %
	AD	-2.47 ± 3.15 %	-1.24 ± 0.81 %	-1.77 ± 2.50 %	-1.14 ± 0.78 %
PWMVC	HC	-0.44 ± 1.59 %	-0.27 ± 0.48 %	-0.34 ± 2.83 %	-0.44 ± 1.04 %
	AD	-0.24 ± 2.64 %	-1.39 ± 0.97 %	-0.06 ± 2.95 %	-1.14 ± 0.84 %

5.2.4. Sensitivity to differences between groups

As shown in Table 1, the sensitivity to differences between groups of ΔBPF is improved by tissue similarity regularization. In the MIRIAD dataset, the results of the proposed pipeline with $w_{\text{sim}} = 0.4$ $(t = 9.85; d = 1.89)$ improve with respect to the FAST + FIRST reference $(t = 4.66; d = 1.01)$ and are also better than those of SIENA $(t = 8.99; d = 1.73)$. In the ADNI1 dataset, the proposed pipeline results $(t = 10.14; d = 1.39)$ are also improved with respect to the FAST + FIRST reference $(t = 6.19; d = 0.80)$ and are better than those of SIENA $(t = 9.78; d = 1.33)$. The sensitivity of the ΔGMF and ΔWMF measures is also improved with respect to the reference FAST + FIRST segmentations. In the MIRIAD dataset, the sensitivity of ΔGMF as measured with our pipeline $(t = 8.05; d = 1.63)$ is higher than that of the reference used for training $(t = 3.35; d = 0.71)$. The ΔWMF sensitivity of our pipeline $(t = 9.31; d = 1.85)$ is much higher than the reference $(t = -0.49; d = 0.10)$. In the ADNI1 dataset, the ΔGMF sensitivity of our pipeline $(t = 9.13; d = 1.23)$ improves with respect to the reference $(t = 5.29; d = 0.69)$. Similarly, the ΔWMF sensitivity of our pipeline $(t = 7.88; d = 1.06)$ is improved with respect to the FAST + FIRST reference $(t = -0.56; d = 0.07)$.



(a) MIRIAD dataset

(b) ADNI1

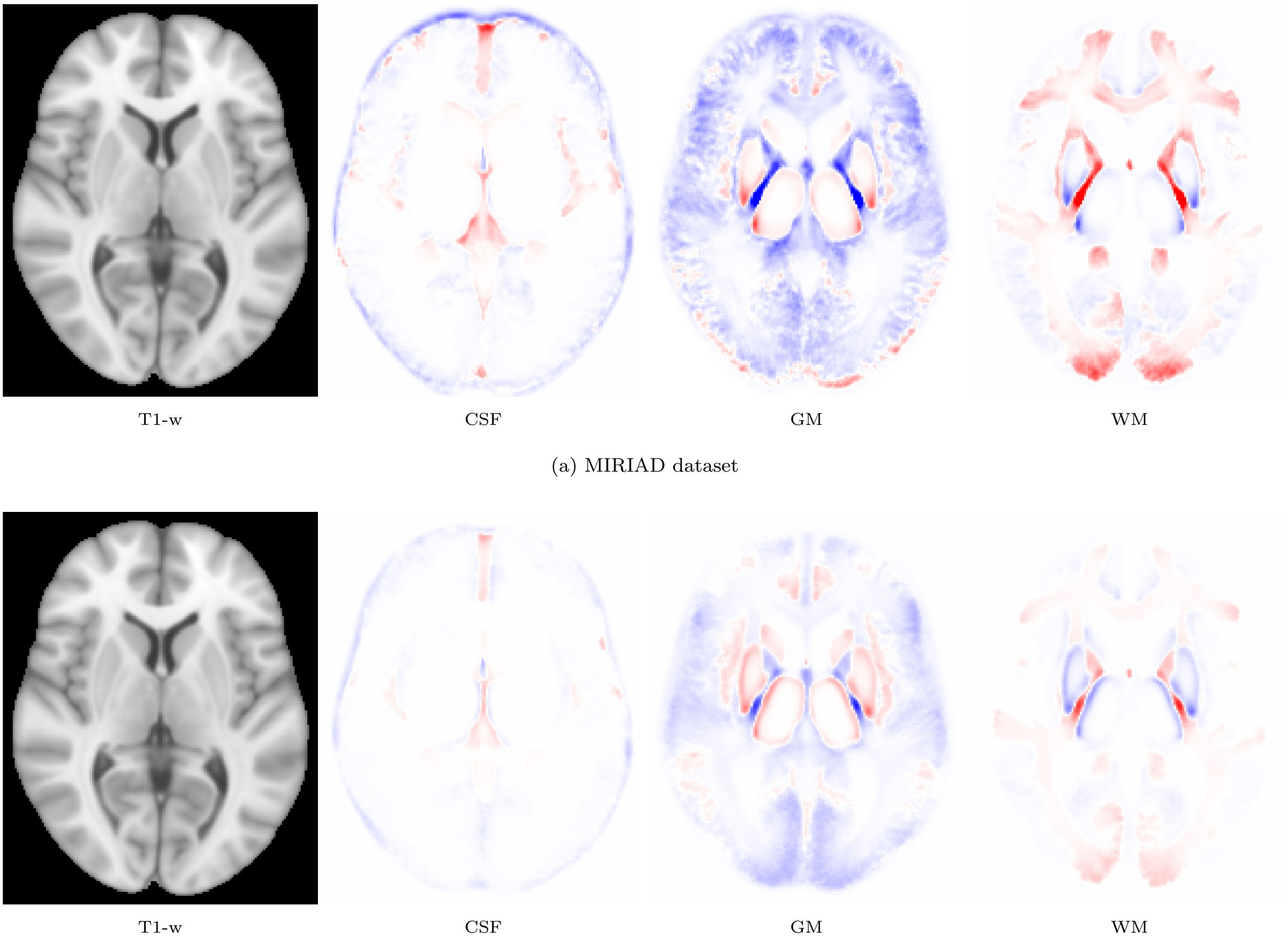
Figure 7: Comparison of argmax segmentation results between FAST + FIRST and the proposed pipeline for a representative case of each dataset. The last column shows the absolute probability differences of voxels changing their most likely tissue class overlaid with a yellow to red colormap, where yellow corresponds to a difference greater than 0.0 and red to a difference of 1.0 in the voxelwise sum of absolute probability differences. Differences for both datasets are mainly located in the cortex, in the interfaces between subcortical structures and in the outer brain border.

5.2.5. Tissue segmentation

It can also be seen that, overall, the effect size of all measures is larger for the MIRIAD dataset than for ADNI1, most likely due to the more consistent imaging parameters that introduce a lower level of confounding factors.

To study the effect of regularized deep learning on the resulting brain tissue segmentations, we perform a qualitative evaluation comparing the reference FAST + FIRST segmentations to those of the presented pipeline trained with $w_{\text{sim}} = 0.4$. Figure 7 shows the tissue segmentation results of FAST + FIRST and our pipeline for two representative cases of MIRIAD and ADNI1. In both datasets, the segmentation of our pipeline presents some differences with respect to the reference, having generally smoother edges between tissues and less noise. The largest segmentation differences are located in the outer brain interface, where our pipeline tends to segment a larger area as brain, and in the borders of subcortical structures, which depending on the case are either enlarged or shrunken. Smaller segmentation differences are also observed in the interfaces between tissues throughout the cortex, where our pipeline tends to segment less WM and more GM than the reference FAST + FIRST segmentation.

Figure 8 shows the median differences between the probabilistic segmentations of FAST + FIRST and the proposed pipeline across all of the cases from each dataset. In practice, we subtract the FAST + FIRST probabilities of each tissue from those of our pipeline, which are then transformed to the MNI space, where we obtain the voxelwise median across all available cases for each dataset. The differences for both datasets show a very similar behavior, with MIRIAD displaying stronger differences, most likely due to its more homogeneous single-center images. In terms of CSF, the blue color around the outer brain border indicates a tendency for our pipeline to segment



(a) MIRIAD dataset

(b) ADNI1

Figure 8: Median probability differences between the probabilistic segmentations of the proposed ($w_{\text{sim}} = 0.4$) pipeline with respect to the reference FAST + FIRST segmentations. For this purpose, the tissue probability maps of FAST + FIRST and the presented pipeline from each case are subtracted and then transformed to the MNI space for joint analysis across the whole dataset. The differences are displayed per tissue in a red to white to blue colormap, where red corresponds to a median difference of -0.25 or less, white to 0.0 and blue to an increase in median probability of 0.25 or higher.

more CSF in that region compared to the reference method. Conversely, the red color in the midline region, ventricle borders and temporal lobes suggests that our pipeline segments less CSF in these regions when compared to FAST + FIRST. Median GM differences display a generalized blue color throughout the cortex, while the WM differences take on a red color, indicating that the presented pipeline segments less WM and more GM in those regions than FAST + FIRST. Another area showing large differences consists of the subcortical structures; the red color in their inner borders suggests that our pipeline tends to reduce their size compared to FAST + FIRST. However, this behavior is reversed when examining the outer borders, where more GM is segmented in favor of reducing the WM. In some cases, such as that illustrated in Figure 6a, different subcortical structures with a thin WM interface between them are merged by our pipeline. Most likely, thin WM interfaces between subcortical structures are a large source of variability within short-interval scans, and the tissue similarity regularization ends up segmenting it as mostly GM to avoid increasing the segmentation differences.

5.3. Limitations

This study has some limitations related to the evaluation of atrophy measures and the clinical applicability of the presented pipeline. Within this work, we have not been able to evaluate the quality or accuracy of either the brain tissue segmentation or the measured atrophy rates. For this purpose, we would need a dataset similar to those considered in this work with sufficiently accurate manual delineations of brain tissue. Despite this limitation, we have evaluated our pipeline on several metrics typically used in the literature to assess longitudinal atrophy quantification and have shown that it improves over extensively validated state-of-the-art methods. In this sense, the comparison with fsl_anat is nuanced since our data-driven pipeline has been previously trained and optimized for the evaluation dataset, whereas fsl_anat was not. However, the main goal of these comparisons is only to quantify the relative improvement of our pipeline, which is trained on these same fsl_anat outputs. Similarly, the comparison with SIENA is also nuanced in the same way since it was not trained or optimized beforehand for the evaluation dataset.

Another limitation is that the performance of deep learning methods could be degraded when applied to images that differ in excess from those seen during training, i.e., from a different image domain, such as one acquired with a different MR scanner, acquisition protocol or different voxel spacing. In this case, training the proposed pipeline from scratch would only require a set of unlabeled short-interval scan pairs from the target domain. Alternatively, the number of training images could be reduced by using a model pretrained on public data along transfer learning or one-shot domain adaptation techniques for deep learning methods (Valverde et al., 2021).

6. Conclusion

In this work, we have presented a novel unsupervised deep learning pipeline for segmentation-based brain atrophy quantification that uses tissue similarity regularization to improve upon the reference segmentation method from which it is trained. We have analyzed the tissue similarity regularization effect and empirically selected $w_{\text{sim}} = 0.4$ as an optimal default value for the similarity weight loss term, which performs well across single-site and multisite datasets.

In general, the presented pipeline improves upon atrophy evaluation metrics and produces smoother and less noisy segmentations than the reference method used for training. The regularization introduces differences in the segmentation of GM/WM in the cortex, the outer brain interface and borders of subcortical structures compared with the reference method. Our evaluation results on short-interval scan pairs show that the proposed regularization lowers the quantification error and improves the overall tissue segmentation consistency, especially for the gray and white matter components. In this sense, our pipeline shows lower and more similar levels of error between the parenchyma and its distinct GM and WM components, whereas the reference method had much larger errors for GM and WM than for the parenchyma. In the longitudinal case, we observed lower variability in atrophy rates and greater sensitivity to differences between healthy controls and AD subjects. Furthermore, while the reference method measured higher levels of WM atrophy for healthy controls than for the AD group, which does not make intuitive sense within a generalized atrophy process, the proposed regularization in our pipeline reverses this tendency and shows more coherent WM atrophy rates between the HC and AD groups.

The presented pipeline is based on the idea that regularized deep learning can exploit data priors to reduce the biases and systematic errors learned from a reference segmentation method. We have shown that the proposed regularization, which aims at reducing short-interval scan differences, can directly improve brain atrophy quantification in the longitudinal case. To the best of our knowledge, this study is the first application of deep learning techniques specifically aimed at improving longitudinal brain atrophy quantification. Data-driven approaches have the potential to surpass their classical counterparts and unlock brain atrophy as a useful diagnostic and prognostic marker for neurodegenerative pathologies.

Acknowledgements

Albert Clèrigues holds an FPI grant from the Ministerio de Ciencia e Innovación with reference number PRE2018-083507. This work has been supported by DPI2020-114769RB-I00 from the Ministerio de Ciencia, Innovación y Universidades. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research. This work has been also supported by ICREA Academia program.

Data used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the

Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Battaglini, M., Jenkinson, M., Stefano, N.D., 2018. Siena-xl for improving the assessment of gray and white matter volume changes on brain mri. Human Brain Mapping 39, 1063.

Dorent, R., Booth, T., Li, W., Sudre, C.H., Kafiabadi, S., Cardoso, J., Ourselin, S., Vercauteren, T., 2021. Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. Medical Image Analysis 67, 101862.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat mri. IEEE Transactions on Medical Imaging 16, 623–629.

Guha Roy, A., Conjeti, S., Navab, N., Wachinger, C., 2019. Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage 186, 713–727.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012.

Holland, D., Dale, A.M., 2011. Nonlinear registration of longitudinal images and measurement of change in regions of interest. Medical Image Analysis 15, 489–497. Special section on IPMI 2009.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning , 448–456.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. NeuroImage 62, 782–790.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M., 2013. Miriad—public release of a multiple time point alzheimer's mr imaging dataset. NeuroImage 70, 33–36.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. International Conference on Machine Learning , 807–814.

Nakamura, K., Guizard, N., Fonov, V.S., Narayanan, S., Collins, D.L., Arnold, D.L., 2014. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. NeuroImage: Clinical 4, 10–17.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. Neural Information Processing Systems .

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., Frisoni, G.B., 2016. Brain atrophy in alzheimer's disease and aging. Ageing Research Reviews 30, 25–48.

Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P.M., Glocker, B., 2018. Neuronet: Fast and robust reproduction of multiple brain image segmentation pipelines. arXiv preprint arXiv:1806.04224 .

Reuter, M., Fischl, B., 2011. Avoiding asymmetry-induced bias in longitudinal image processing. NeuroImage 57, 19–21.

Rocca, M.A., Battaglini, M., Benedict, R.H., Stefano, N.D., Geurts, J.J., Henry, R.G., Horsfield, M.A., Jenkinson, M., Pagani, E., Filippi, M., 2017. Brain mri atrophy quantification in ms. Neurology 88, 403–413.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention 9351, 234–241.

Rovira, A., Wattjes, M.P., Tintoré, M., Tur, C., Yousry, T.A., Sormani, M.P., Stefano, N.D., Filippi, M., Auger, C., Rocca, M.A., Barkhof, F., Fazekas, F., Kappos, L., Polman, C., Miller, D., Montalban, X., 2015. Magnims consensus guidelines on the use of mri in multiple sclerosis—clinical implementation in the diagnostic process. Nature Reviews Neurology 2015 11:8 11, 471–482.

Royle, N.A., Hernández, M.C., Maniega, S.M., Arabisala, B.S., Bastin, M.E., Deary, I.J., Wardlaw, J.M., 2013. Influence of thickening of the inner skull table on intracranial volume measurement in older people. Magnetic resonance imaging 31, 918–922.

Sastre-Garriga, J., Pareto, D., Battaglini, M., Rocca, M.A., Ciccarelli, O., Enzinger, C., Wuerfel, J., Sormani, M.P., Barkhof, F., Yousry, T.A., Stefano, N.D., Tintoré, M., Filippi, M., Gasperini, C., Kappos, L., Río, J., Frederiksen, J., Palace, J., Vrenken, H., Montalban, X., Àlex Rovira, 2020. Magnims consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. Nature Reviews Neurology 2020 16:3 16, 171–182.

Sastre-Garriga, J., Pareto, D., Àlex Rovira, 2017. Brain atrophy in multiple sclerosis: Clinical relevance and technical aspects. Neuroimaging Clinics 27, 289–300.

Smeets, D., Ribbens, A., Sima, D.M., Cambron, M., Horakova, D., Jain, S., Maertens, A., Vlierberghe, E.V., Terzopoulos, V., Binst, A.M.V., Vaneckova, M., Krasensky, J., Uher, T., Seidl, Z., Keyser, J.D., Nagels, G., Mey, J.D., Havrdova, E., Hecke, W.V., 2016. Reliable measurements of brain atrophy in individual patients with multiple sclerosis. Brain and Behavior 6, e00518.

Smith, S.M., Rao, A., Stefano, N.D., Jenkinson, M., Schott, J.M., Matthews, P.M., Fox, N.C., 2007. Longitudinal and cross-sectional analysis of atrophy in alzheimer's disease: Cross-validation of bsi, siena and sienax. NeuroImage 36, 1200–1206.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., Stefano, N.D., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17, 479–489.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2818–2826.

Valverde, J.M., Imani, V., Abdollahzadeh, A., Feo, R.D., Prakash, M., Ciszek, R., Tohka, J., 2021. Transfer learning in magnetic resonance brain imaging: A systematic review. Journal of Imaging 2021, Vol. 7, Page 66 7, 66.

Yamanakkanavar, N., Choi, J.Y., Lee, B., 2020. Mri segmentation and classification of human brain using deep learning for diagnosis of alzheimer's disease: A survey. Sensors (Switzerland) 20, 1–31.

Zeiler, M.D., 2012. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701 .

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging 20, 45–57.

Chapter 7

Main results and discussion

In this thesis, we have explored a range of clinically relevant brain imaging tasks for the prognosis of neurological pathologies such as stroke, multiple sclerosis or Alzheimer's disease. More specifically, we have presented works for ischemic stroke lesion segmentation, reducing the effect of WM in brain tissue segmentation and for longitudinal atrophy quantification. Dealing with several image markers and their challenges has provided a deeper understanding of the issues and possible approaches for their combined use in predictive models. In this chapter, we outline the main results and findings obtained in this thesis and discuss their significance towards completion of the main objectives.

7.1 Brain lesion segmentation

7.1.1 Class imbalance handling

Within deep learning approaches, a deliberate strategy to balance the representation of lesioned and healthy tissue in the training set has been essential towards achieving good segmentation performance. In our patch-based approaches, this has been addressed with the proposed patch sampling strategy, which is based on a mixture of recently presented works in brain lesion segmentation [26, 51]. We proposed a balanced sampling of patches centered in healthy and lesion classes as done by Kamnitsas et al. [26]. However, systematically sampling patches centered on lesion voxels will induce a neighborhood bias by which the lesion is always seen centered in the patch and surrounded by a uniform band of healthy tissue [51]. As proposed by Guerrero et al. [51], we add a random 3D offset of up to half of the patch size to the lesion sampled patches. This increases the representation of lesion boundaries as well as the variability of healthy neighborhoods in which the lesion is seen during training. However, even with the use of balanced patch sampling strategies, which equalize the number of patches with and without lesions, the training set will still tend to be highly imbalanced at a voxel level due to the generally higher proportion of healthy voxels in each patch. To avoid this, Guerrero et al. [51] proposed exclusively sampling the lesion class, however, overrepresenting the lesion class implies underrepresenting the healthy class and actually increasing the error rate in healthy regions. Since these areas represent the majority of the segmentation, even a tiny increase in error rates of the healthy class will translate in higher overall segmentation error.

To further reduce the effect of class imbalance without underrepresenting any class or region of the brain, we also proposed the use of class balancing loss functions, which work by modulating the contribution of each class to the final loss term. More specifically, we have used two functions based on different principles to handle class imbalance, the Focal loss [52] and the Generalized Dice Loss (GDL) [53]. The Focal loss, used in Chapter 3, is a difficulty weighted version of the crossentropy loss which is dynamically scaled to avoid learning from examples that are already correctly segmented with high confidence. When the network specializes more in one class, its classification confidence will be higher and thus, the focal loss will scale down its contribution to favor the other class. In contrast, the GDL, used in Chapter 4, is a differentiable loss function based on the Dice similarity coefficient (DSC). The class balancing mechanism is provided by the underlying concept of segmentation overlap as well as by a class weight which normalizes the class contributions based on their relative representation in the training set. In general, the use of balancing loss functions raises the positive predictive value (PPV) and DSC metrics by reducing the number of false positives. However, the class balancing mechanisms of loss functions might come at the cost of worsened stability and convergence properties. More specifically, the GDL is quite unstable and slow in training, which is why its typically used in conjunction with a crossentropy loss term. The sum of both provides the class balancing of the GDL along with the good convergence properties of the crossentropy loss. Furthermore, we have observed in our experiments that some class balancing losses are better suited to different tasks. Although we tested both the focal loss and GDL in both of the proposed approaches, the best results were obtained with the Focal loss in our 2D CT approach and with the GDL and crossentropy in our 3D MRI method.

7.1.2 Symmetry features in stroke

The use of features based on the symmetry between brain hemispheres for ischemic stroke lesion segmentation has consistently improved our results in both CT and MRI imaging as well as within 2D and 3D approaches. The use of symmetry features is based upon the observation that ischemic stroke typically affects only one hemisphere and that comparing with the opposing healthy side can help better outline the tissue with abnormal appearance on the affected side. In our implementation, we generate symmetrically augmented modalities through mid-sagittal flipping and linear registration, which robustly aligns opposing hemispheres between the original and augmented modality. Then, we perform early fusion by concatenating the original and symmetrically augmented modalities along the channel dimension. In this way, the convolutional kernels of our network are able to extract spatial features jointly considering the intensities of left and right hemispheres. Moreover, early fusion allows the use of a standard single-path U-Net, which avoids multi-path architectures that would raise the amount of training parameters and increase the risk of over-fitting.

The main limitation of our patch-based implementation of symmetry features is the reduced receptive field due to the use of small patches. This constrains the spatial extent from which the model can extract symmetry features useful for segmentation. In stroke, due to the vascular nature of the pathology, information about the perfusion of distant regions can be informative for segmentation. However, we have empirically tested that the use of larger patches decreases segmentation performance due to worsened class imbalance. Although reduced spatial context is a big drawback of using small patches, class imbalance handling is currently the main limitation. Another possible drawback of the current symmetric modality augmentation procedure is that it uses image interpolation to align the opposing hemispheres. Although it is only applied to the augmented modalities, and not the original ones which are the ones actually being segmented, the interpolation distorts the image intensities and might limit the quality and accuracy of symmetry features.

7.1.3 Post-processing of segmentation results

Post-processing of segmentation results has been an essential technique to correct the biases of deep learning models and further improve lesion segmentation results. Perhaps the most basic technique is that of variable thresholding. Typically, segmentation probabilities are binarized by taking the class with largest probability, which would equate to using a threshold of 0.5 in a binary classification task such as brain lesion segmentation. However, this makes the implicit assumption that the confidence of the deep learning model is perfectly calibrated around the 0.5 threshold. It might be the case that a model is under or overconfident, which would mean that the best segmentation results would be obtained with a lower or higher threshold respectively. The confidence of a model can be influenced by a different number of factors, such as the training set class distribution, loss function, input normalization, etc. Lowering the threshold would increase the model sensitivity, reducing the number of false negatives but also increasing the amount of false positives. The opposite is also true if the threshold is increased, leading to a reduction of false positives but also an increase in the amount of false negatives. In our works dealing with brain lesion segmentation, we have selected the optimal threshold by empirical tests on the validation data. For instance, in Chapter 4, the output of our under-confident model was compensated by selecting a lower threshold of 0.2, which raised the sensitivity but also the number of false positive lesions. To complement the variable thresholding, we have also utilized morphological post-processing to leverage the a priori knowledge about the morphology characteristics of target lesions. In the case of ischemic stroke, the core and penumbra are typically large in size, so we discarded lesions under an empirically chosen minimum volume size, which led to an overall improvement of the segmentation metrics by reducing the number of false positive lesions. However, minimum lesion size filtering must be carefully considered in the clinical setting, as lesions under the selected threshold will be discarded during inference.

Another technique which we explored, inspired by the work of Nair et al. [54], is segmentation filtering based on Monte Carlo (MC) uncertainty estimation. In our implementation, we added prediction dropout layers in the lower levels of the U-Net model and performed the patch-wise average of three predictions made with a 10% dropout rate during inference. The use of uncertainty filtering lowered the HD and improved lesion volume estimation without affecting the DSC, which suggests a better segmentation of the lesion borders. However, the use of this technique increased the inference time three-fold, which is why it was only applied within the 2D approach presented in Chapter 3, that had lower computational requirements than the 3D approach presented in Chapter 4.

7.2 Brain tissue segmentation

In this thesis, we have also developed a deep learning framework for brain tissue segmentation which has been the basis for the WM lesion effect analysis done in Chapter 5 and the longitudinal atrophy quantification method presented in Chapter 6. The framework consists on a 3D patch-based unsupervised deep learning approach for segmentation-based volumetry of brain tissue. To avoid the need for manual ground truth segmentations, we generate the training targets using fully automated classical methods, which are regarded as more accurate and reproducible than those manually made by experts [55]. In recent literature, similar deep learning approaches for brain tissue segmentation trained on the outputs of classical methods have shown improved properties. For instance, NeuroNet [56] obtained similar DSC to manual segmentations than the reference methods used for training, but having a shorter execution time and with less failed cases. FastSurfer [57] also reports similar DSC to manual segmentations than FreeSurfer, the method used to generate the training targets. Moreover, results of FastSurfer on longitudinal data show reduced short interval error and increased sensitivity to group differences between healthy and dementia patients. In contrast with previous works focusing on categorical segmentation, our framework focuses on providing accurate brain volumetry in pathological cases. Thus, we incorporate a series of specific techniques to improve partial volume estimation accuracy and to deal with pathological images more robustly:

- **Input normalization.** The proposed minmax input normalization based on intensity percentiles within the brain parenchyma provides several benefits for brain volumetry. Since the normalization parameters are computed only from intensities of the parenchyma, the intensity range of brain tissue is always within the same interval regardless of the amount of face or neck present in the images. In addition, the use of percentiles allows to reduce the influence of outlier or abnormal intensities, such as WM lesions. Furthermore, percentiles are also more robust to changes in relative tissue volumes, such as those caused by atrophy, since the minimum or maximum intensity of an image is typically not affected by these changes.

- **Probabilistic loss.** In our experiments, we observed that the categorical crossentropy loss, which uses class labels as targets, resulted in very sharp and over-confident segmentation probabilities, which differed from the softer probability maps used for training. Since our goal in inference is not to maximize the probability of the most likely tissue class, but rather, to approximate the partial volume probabilities, we propose the use a soft version of the crossentropy loss using probabilistic targets.

- **Small patches.** The use of small patches provides several advantages for brain tissue segmentation. First, small patches allow the use of bigger batch sizes, which stabilize the training procedure. Second, it allows for the use of patch sampling strategies to balance the representation of tissue and structures in the training set. Although class imbalance is not a big issue in tissue segmentation, there is still a difference between the relative volumes and representation of tissue classes. Finally, as we will later discuss, small patches also limit the spatial extent to which a lesion can introduce errors in the tissue segmentation.

Despite the use of much smaller patches, our framework obtains a DSC to the reference method of 94.6 ± 2.5%, a level comparable to the ones reported by similar state-of-the-art approaches, such as NeuroNet (93.1 ± 2.4%) or FastSurfer (avg. 89%). This suggests that a large spatial context is not important for brain tissue segmentation and that small patches can be used without any performance penalties. The competitive results also show that the proposed input normalization and probabilistic loss function do not harm segmentation performance in any way, while providing the previously discussed benefits. Compared to traditional statistical methods, a big advantage of deep learning approaches is that they can adapt their architecture and training procedure to incorporate additional elements or data-driven techniques to improve or amplify their capabilities. The works in this thesis take advantage of this fact and propose modified architectures and data-driven techniques to extend and improve upon the reference method used for training.

7.2.1 Effect of WM lesions

In Chapter 5, we have shown that the ways in which WM lesions affect patch-based deep learning approaches are different than those of classical statistical methods. In the reference tissue segmentation method used for training our approach, WM lesions interfered with the characterization of tissue intensity distributions and shifted the partial volume estimation probabilities across the whole image. In contrast, the error from WM lesions is introduced in our deep learning framework at the global level through changes in the input normalization parameters and at the local level through changes in the segmentation probabilities of patches where the lesion appears. The proposed percentile-based minmax input normalization on the parenchyma reduces the effect of the WM lesion intensities in the normalization parameters and avoids shifting the intensities of the whole image on most cases. Since lesions only affect a small percentage of white matter, the upper intensity percentile is rarely affected and the normalization parameters remain fairly constant. Moreover, WM lesion intensities in the T1-w images used for volumetry are typically darker than the WM itself and do not influence the upper intensity percentiles. At the local level, the use of small patches limits the spatial extent of patches affected by the WM lesion without compromising segmentation performance. Without any kind of lesion filling or inpainting, our deep learning framework already showed much lower volume error due to WM lesions (GM: 0.07 ± 0.09%; WM:

0.10 ± 0.11%) than the reference approach used for training (GM: 0.89 ± 1.14%; WM: 1.22 ± 1.58%).

To reduce the effect of WM lesions in the measured volumes, the gold standard technique is to perform WM lesion inpainting prior to the tissue segmentation. Recently, several deep learning methods have been proposed that perform much more realistic and accurate WM lesion inpainting that further reduce the error on upcoming segmentation tasks. However, since the ultimate goal of the lesion inpainting is to provide more accurate brain volumetry, we proposed a deep learning method in Chapter 5 that is trained to perform both inpainting and tissue segmentation in and end-to-end fashion using a multi-task double U-Net architecture. In our approach, we introduce artificial lesions to a healthy image and then train our model to inpaint the lesions and approximate the originally healthy tissue probabilities. By jointly optimizing both tasks, the inpainting is also trained to improve the downstream segmentation task through the gradients coming from the segmentation loss. This data-driven technique allows to directly minimize the WM lesion effect in the tissue segmentation without going through the intermediate step of accurately concealing its intensities. The error from WM lesions on the tissue volumes measured with our approach is almost negligible (GM: 0.01 ± 0.03%; WM: 0.02 ± 0.04%) thanks to the use of small patches, the proposed input normalization and the end-to-end multi-task optimization. The main limitation of our method is that its not trained on pathological images with actual WM lesions, since it needs to be trained on tissue segmentations from healthy scans to which we later add artificial lesions. WM lesions might induce inflammatory edema or mass effect deformation on the surrounding healthy tissue and change its appearance in a way that the network might not be trained to deal with.

7.2.2 Longitudinal atrophy quantification

In Chapter 6, we presented a novel unsupervised deep learning method for segmentation-based longitudinal brain atrophy quantification in which we learn from the outputs of an automated reference method while regularizing the training using data priors to avoid learning its errors and biases. Typically, short interval scan pairs are used for evaluating the quantification error of brain atrophy approaches based on the assumption that an ideal method would measure zero volume change between them. In our approach, we use the zero change assumption to propose a tissue similarity regularization which penalizes volume differences between pairs of patches from short interval scan pairs during training. A key element of our method is that it segments only one patch at a time, which means that the network cannot extract features jointly considering the intensities from both short interval scans to reduce the volume differences. Thus, the regularization can only rely on features extracted from a single scan to reduce the influence of intensity variations, such as those from noise or movement artifacts, and provide more consistent tissue volumes between the two short interval scans. For training, we co-register the two short interval scans into a halfway space to obtain the voxel-wise spatial alignment needed to apply the tissue similarity regularization. However, since the network is trained to segment only one patch at a time, inference can be directly performed on a single scan at a time and without having to perform any kind of image registration or interpolation.

By penalizing volume differences between short interval scans during training, we obtain a tissue segmentation and partial volume estimation which is more robust to noise and image artifacts. The improvements imprinted on the model during training translate directly to the inference of longitudinal imaging and provide a huge improvement on brain atrophy evaluation metrics. The most clinically relevant is the increase in sensitivity to differences in atrophy rates between healthy and AD patients, where the Cohen's d effect size of the reference method is improved from d=1.01 to d=1.89 in the MIRIAD dataset and from d=0.8 to d=1.39 in the ADNI1 dataset. It is worth noting that the only segmentations that the network has ever seen are those from the reference method, and that the observed improvements are achieved solely by the proposed tissue similarity regularization applied during training. At a qualitative level, our approach produces a smoother and less noisy segmentation that seems to correlate better with the expected brain morphology. Compared with the reference method, the improvement is mainly achieved by differences in the segmentation of subcortical structure borders, the cortex and outer brain border.

We also observed other relevant improvements of our approach with respect to the segmentation of the distinct GM and WM components that form the brain parenchyma. In our experiments, the short interval error of the reference method was three times as high for the GM and WM than for the parenchyma. This suggests that the reference method is very inconsistent in estimating the internal mixture of GM and WM components of the parenchyma and measures differing volumes from each of the short interval scans. In contrast, our approach has more similar levels of errors for the GM and WM components than for the parenchyma, which suggests a much more consistent volume of the internal GM/WM components from the parenchyma between short interval scans. We also observed a remarkable WM atrophy rate correction effect due to the tissue similarity regularization. In both of the considered datasets, the reference method measured slightly higher levels of WM atrophy for the healthy controls (HC) than for the AD patients, which does not make intuitive sense within a generalized brain atrophy process. In contrast, we show that our proposed regularized approach reverts this trend from the reference segmentations and produces much more coherent rates of WM atrophy which are much higher for the AD patients than for the HC controls. These two results suggest that the regularization directly improves the segmentation and partial volume estimation of the GM and WM tissues.

The main limitation of our approach is that we do not really control where or how the regularization is affecting the tissue segmentation model. For instance, we observed that, in some cases, nearby subcortical structures were merged by our approach, which removed the WM interface between them. Most likely, these thin WM channels were a large source of volume differences between short interval scans and the regularization ended compromising the segmentation accuracy to reduce volume variability.

Chapter 8

Conclusions

In this chapter, we summarize the work accomplished and outline the main contributions of this PhD thesis. We also discuss on future research that could extend upon the novel ideas that have been explored in this thesis.

8.1 Contributions

The work done in this PhD has explored techniques and methods to extract neuroimage markers relevant for prognosis and outcome prediction of brain pathologies. In the following paragraphs, we provide the main conclusions and contributions of this thesis.

- We have proposed two patch-based deep learning methods for stroke lesion segmentation from MRI and CT imaging. In both of these approaches, we have achieved state-of-the-art results by proposing novel ways to mitigate the class imbalance, exploit brain symmetry features and post-process the segmentation results. In our patch-based framework, we tackled the class imbalance issue with a balanced training patch sampling along with class balancing functions that improved the overall lesion segmentation accuracy. We also proposed a way to utilize brain symmetry features for stroke lesion segmentation within U-Net based architectures by using early fusion of symmetrically augmented modalities. Within our implementation, we can exploit distant symmetry features while using small patches and also use a standard single-path U-Net architecture that avoids increasing the number of trainable parameters. To adjust the model confidence, variable thresholding combined with morphological post-processing allows for improved sensitivity without a significant increase of false positives. Furthermore, the use of uncertainty based filtering has refined lesion borders and reduced the number of false positives.

  The proposed 2D patch-based deep learning method for stroke core lesion segmentation from CT perfusion is an improved version of the approach we submitted to the ISLES 2018 challenge and that was among the onsite challenge finalists. The approach presented in this thesis is an updated version that is currently ranked 2nd in the ongoing testing leaderboard of the ISLES 2018 challenge online platform and was published at *Computers in Biology and Medicine* [JCR MCB IF 3.434, Q1(8/59)].

The proposed 3D patch-based deep learning approach for stroke lesion segmentation from multimodal MRI is currently ranked 1st in both of the ongoing testing leaderboards for the SISS and SPES sub-tasks from the ISLES 2015 challenge online platform and was published at *Computer Methods and Programs in Biomedicine* [JCR CSTM IF 3.424, Q1(15/104)].

- We have analyzed the effect of WM lesions on a patch-based deep learning method for brain tissue segmentation. In the reference method from which we trained, the lesion had a global effect and shifted the partial volume estimation for the whole image, producing large errors in the measured volumes. Within our patch-based deep learning framework, we observed that the lesion has both a global and local effect. The global effect is introduced through changes on the input normalization parameters which shift the input intensities of the whole image. The local effect is introduced through changes in the segmentation probabilities of patches where the WM lesion appears. To mitigate these, we have proposed the use of a minmax normalization based on intensity percentiles to reduce the global effect and the use of small patches to limit the spatial extent of the local effect. Our deep learning approach shows a much smaller influence from WM lesions on the measured volumes than the reference method used for training.

- In addition, we have proposed a novel data-driven technique to minimize the effect of WM lesions on deep learning based brain volumetry. The use of a separate preprocessing step doing lesion filling or inpainting is currently the gold standard technique to reduce the error introduced by WM lesions. However, recently proposed state-of-the-art deep learning techniques optimize inpainting through appearance or intensity based metrics which do not consider the downstream segmentation task. In our approach, we train the inpainting and tissue segmentation tasks in an end-to-end fashion. By jointly optimizing both tasks, we obtain an inpainting model that is also trained to aid in the segmentation task through the gradients coming from the segmentation loss term. This data-driven technique can adapt to the WM intensity profiles of the target images and to the morphology and locations of WM lesions of the training set. The error from WM lesions is reduced to almost negligible levels by combining the use of small patches, the proposed input normalization and the end-to-end multi-task training. This work is currently under a second revision at *Computerized Medical Imaging and Graphics* [JCR RNMMI IF 4.790, Q1(27/314)]

- We have proposed a novel unsupervised segmentation-based deep learning method for longitudinal brain atrophy quantification. In our patch-based approach, we learn from a reference tissue segmentation method while using data priors to regularize the training and avoid learning its errors and biases. For this, we proposed a tissue similarity regularization which penalizes volume differences between short interval scans during training. The training data is generated in an unsupervised manner by using a fully automated tissue segmentation pipeline on unlabeled pairs of co-registered short interval scans. In inference, the model segments one image at a time in its native space, i.e. without any kind of image interpolation. For longitudinal evaluation, the baseline and followup images are independently segmented and then measures of change are derived from the tissue volumes. The improvements in tissue segmentation and partial volume estimation imprinted during training thanks to the regularization translate to the longitudinal case and improve brain atrophy quantification metrics. Despite being exclusively trained on the reference method segmentations, we achieve much lower short interval error and higher sensitivity to differences between healthy and AD patients thanks to the tissue similarity regularization. The proposed approach achieves state-of-the-art results and is the first unsupervised deep learning method purposefully built for longitudinal brain atrophy quantification. This work has been submitted to *Medical Image Analysis* [JCR CSAI IF 8.880, Q1(5/133)].

During this PhD thesis, various collaborations have taken place with other researchers of the VICOROB group. In particular, a deep learning approach for hematoma segmentation in hemorrhagic stroke [58] which stemmed from the project carried out with Dra. Yolanda Silva from Hospital Dr. Josep Trueta. Other relevant collaborations include a study on the use of synthetic images for MS lesion segmentation [59] and a contribution to a book chapter on deep learning applications for medical image analysis [60]. Moreover, contributions to on-site international MICCAI challenges were done for stroke lesion segmentation in 2018 and 2022. In the ISLES 2018 challenge, the proposed method was among the finalists and was presented on-site in the workshop.

8.2 Future work

Throughout the realization of this PhD thesis, some limitations and aspects of our research were not fully developed or reaching out of scope and are left as future work.

For instance, the symmetric modality augmentation proposed to exploit brain symmetry features in stroke could also be considered for the segmentation of brain lesions from other pathologies. In the literature, an assumption is made that symmetry features are only useful on pathologies which mainly affect one of the brain hemispheres, such as stroke. However, we believe that deep learning methods can still take advantage from brain symmetry features even if the opposing hemisphere contains pathological intensities, since these kind of approaches can learn to ignore these during training.

Our deep learning framework for brain tissue segmentation has shown much less influence from WM lesions than classical state-of-the-art approaches. Moreover, the proposed data-driven technique to further reduce the effect of WM lesions on brain tissue segmentation achieves an almost negligible error on the measured volumes. However, both of these studies have been performed with artificial lesions made from registered binary lesion masks and with intensities sampled from a normal intensity distribution. To obtain more meaningful and clinically relevant results, a more sophisticated and realistic way to generate artificial WM lesions should be developed. For this, recent generative deep learning techniques could be used such as conditional variational autoencoders (VAEs) or generative adversarial networks (GANs). Moreover, inaccurate lesion masks could be used during training to increase robustness to noisy ground truth labels and to the pathological effects of lesions on surrounding tissue, such as edema or mass effect.

In the presented longitudinal brain atrophy quantification method, we applied the tissue similarity regularization on the typically available short interval scans which are used to evaluate the atrophy quantification error. These are always acquired back-to-back with the same scanning parameters and without repositioning of the patient. We believe that the tissue similarity regularization could also be used to improve the robustness to changes of scanner, acquisition protocol or patient repositioning. For this, the regularization would be applied between multi-domain short interval scans, i.e. made using different scanners, with diverse acquisition parameters and with inaccurate repositioning of the patient. Multi-domain short interval scans performed in this way would still hold the tissue similarity prior and could in theory be used to train the model to measure more consistent tissue volumes across a diverse range of scanning conditions. However, to the best of our knowledge, there are no readily available datasets having multi-domain short interval scans made in this way. A possible way to overcome the lack of this kind of data would be to explore the use of artificially generated or interpolated data. However, since brain atrophy quantification relies on very fine and accurate analysis of intensities, the usefulness of synthesized or generated data should be carefully validated in real images.

Furthermore, we believe that the proposed data-driven technique to minimize the effect of WM lesions could be combined within our longitudinal brain atrophy quantification method to provide accurate atrophy measures even on pathological cases with WM lesions. In this case, the multi-task double U-Net architecture would be used along with the tissue similarity regularization acting on pairs of short interval scans from healthy subjects with artificially added WM lesions. However, training and evaluation of such a method would require an extensive and diverse set of imaging data. For training, this method would require unlabeled pairs of short interval scans from healthy subjects and also from pathological ones having manually delineated WM lesion masks. Evaluation would then be performed on short interval and longitudinal imaging of both healthy and pathological subjects, with manually labeled WM lesion masks, from a similar image domain as the training one.

Ultimately, the purpose of these image markers is to be used within prognostic and functional outcome prediction tasks. We believe that future research should focus on development of models trained and optimized to directly provide the final predictive marker from a combination of these imaging markers along with relevant clinical and patient data. Within these models, much higher robustness and accuracy could be obtained through the use of end-to-end multi-task optimization which would be explicitly trained to extract several image markers as intermediate features within the predictive model. In this way, the extraction of different imaging markers would be coupled and directly optimized towards improvement of the final prediction task. Moreover, higher levels of interpretability could be achieved since the intermediate image markers can be obtained as an additional output of the model and reviewed by physicians to provide insights on what the model is taking into account to make its prediction.

Bibliography

[1] J. Redon et al. “Stroke mortality and trends from 1990 to 2006 in 39 countries from Europe and Central Asia: implications for control of high blood pressure”. In: *European Heart Journal* 32.11 (2011), pp. 1424–1431.

[2] Bernd F. Tomandl et al. “Comprehensive Imaging of Ischemic Stroke with Multisection CT”. In: *RadioGraphics* 23.3 (2003), pp. 565–592.

[3] Sunil A. Sheth et al. “Time to endovascular reperfusion and degree of disability in acute stroke”. In: *Annals of Neurology* 78.4 (2015), pp. 584–593.

[4] W. Serles et al. “Endovascular stroke therapy in Austria: a nationwide 1-year experience”. In: *European Journal of Neurology* 23.5 (2016), pp. 906–911.

[5] Rakeshsingh K Singh et al. “Acute Ischemic Stroke Treatment Using Mechanical Thrombectomy: A Study of 137 Patients.” In: *Annals of Indian Academy of Neurology* 20.3 (2017), pp. 211–216.

[6] Jo Lane et al. “Multiple sclerosis incidence: A systematic review of change over time by geographical region”. In: *Multiple Sclerosis and Related Disorders* 63 (2022), p. 103932.

[7] Alan J Thompson et al. “Multiple sclerosis”. In: *The Lancet* 391.10130 (2018), pp. 1622–1636.

[8] Claudio Gasperini et al. “Unraveling treatment response in multiple sclerosis: a clinical and MRI challenge”. In: *Neurology* 92.4 (2019), pp. 180–192.

[9] Zeinab Breijyeh and Rafik Karaman. “Comprehensive Review on Alzheimer’s Disease: Causes and Treatment”. In: *Molecules* 25.24 (2020).

[10] Samaneh A. Mofrad et al. “Cognitive and MRI trajectories for prediction of Alzheimer’s disease”. In: *Scientific Reports 2021 11:1* 11 (1 Jan. 2021), pp. 1–10.

[11] Rahul S. Desikan et al. “Automated MRI measures predict progression to Alzheimer’s disease”. In: *Neurobiology of Aging* 31 (8 Aug. 2010), pp. 1364–1374.

[12] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444.

[13] Y. LeCun et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1.4 (1989), pp. 541–551.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[15] Di Lin et al. "Neural networks for computer-aided diagnosis in medicine: A review". In: *Neurocomputing* 216 (2016), pp. 700–708.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI).* 2015, pp. 234–241.

[17] C Oppenheim et al. "Is there an apparent diffusion coefficient threshold in predicting tissue viability in hyperacute stroke?" In: *Stroke* 32.11 (2001), pp. 2486–91.

[18] Juan F Arenillas et al. "Prediction of early neurological deterioration using diffusion- and perfusion-weighted imaging in hyperacute middle cerebral artery ischemic stroke." In: *Stroke* 33.9 (2002), pp. 2197–203.

[19] S. B. Coutts et al. "Reliability of Assessing Percentage of Diffusion-Perfusion Mismatch". In: *Stroke* 34.7 (2003), pp. 1681–1683.

[20] Carly S. Rivers et al. "Acute Ischemic Stroke Lesion Measurement on Diffusion-weighted Imaging–Important Considerations in Designing Acute Stroke Trials With Magnetic Resonance Imaging". In: *Journal of Stroke and Cerebrovascular Diseases* 16.2 (2007), pp. 64–70.

[21] Oskar Maier et al. "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". In: *Medical Image Analysis* 35 (2017), pp. 250–269.

[22] Chaolu Feng, Dazhe Zhao, and Min Huang. "Segmentation of Stroke Lesions in Multi-spectral MR Images Using Bias Correction Embedded FCM and Three Phase Level Set". In: *Proceedings of ISLES (SISS and SPES) challenge.* 2015.

[23] Halla-Leena Halme, Antti Korvenoja, and Eero Salli. "ISLES (SISS) challenge 2015: Segmentation of stroke lesions using spatial normalization, Random Forest classification and contextual clustering". In: *Proceedings of ISLES (SISS) challenge* (2015), pp. 31–34.

[24] Oskar Maier, Matthias Wilms, and Heinz Handels. "Random forests for acute stroke penumbra estimation". In: *Proceedings of ISLES (SPES) challenge.* 2015.

[25] Richard Mckinley et al. "Segmenting the ischemic penumbra: a spatial Random Forest approach with automatic threshold finding". In: *Proceedings of ISLES (SPES) challenge.* 2015.

[26] Konstantinos Kamnitsas et al. "Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI". In: *Ischemic stroke lesion segmentation* 13 (2015), p. 46.

[27] Matthew Bobinski et al. "The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease". In: *Neuroscience* 95.3 (1999), pp. 721–725.

[28] Jennifer L Whitwell et al. “Neuroimaging correlates of pathologically defined subtypes of Alzheimer’s disease: a case-control study”. In: *The Lancet Neurology* 11.10 (2012), pp. 868–877.

[29] N. De Stefano et al. “Evidence of early cortical atrophy in MS”. In: *Neurology* 60 (7 Apr. 2003), pp. 1157–1162.

[30] Katrin Morgen et al. “Evidence for a direct association between cortical atrophy and cognitive impairment in relapsing–remitting MS”. In: *NeuroImage* 30 (3 Apr. 2006), pp. 891–898.

[31] Richard A. Rudick et al. “Gray matter atrophy correlates with MS disability progression measured with MSFC but not EDSS”. In: *Journal of the Neurological Sciences* 282 (1-2 July 2009), pp. 106–111.

[32] Maria A. Rocca et al. “Brain MRI atrophy quantification in MS”. In: *Neurology* 88 (4 Jan. 2017), pp. 403–413.

[33] Jaume Sastre-Garriga, Deborah Pareto, and Àlex Rovira. “Brain Atrophy in Multiple Sclerosis: Clinical Relevance and Technical Aspects”. In: *Neuroimaging Clinics* 27 (2 May 2017), pp. 289–300.

[34] Hákon Gudbjartsson and Samuel Patz. “The Rician distribution of noisy MRI data”. In: *Magnetic resonance in medicine* 34.6 (1995), pp. 910–914.

[35] Zografos Caramanos et al. “Gradient distortions in MRI: Characterizing and correcting for their effects on SIENA-generated measures of brain volume change”. In: *NeuroImage* 49.2 (2010), pp. 1601–1611.

[36] Kunio Nakamura et al. “Diurnal fluctuations in brain volume: statistical analyses of MRI from large populations”. In: *Neuroimage* 118 (2015), pp. 126–132.

[37] Daniel-Paolo Streitbürger et al. “Investigating structural brain changes of dehydration using voxel-based morphometry”. In: (2012).

[38] Emma R. Mulder et al. “Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST”. In: *NeuroImage* 92 (2014), pp. 169–181.

[39] Marco Battaglini, Mark Jenkinson, and Nicola De Stefano. “SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI”. In: *Human Brain Mapping* 39 (3 Mar. 2018), p. 1063.

[40] Stephen M. Smith et al. “Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis”. In: *NeuroImage* 17 (1 Sept. 2002), pp. 479–489.

[41] Peter A. Freeborough and Nick C. Fox. “The boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat MRI”. In: *IEEE Transactions on Medical Imaging* 16 (5 1997), pp. 623–629.

[42] Yongyue Zhang, Michael Brady, and Stephen Smith. "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm". In: *IEEE Transactions on Medical Imaging* 20 (1 Jan. 2001), pp. 45–57.

[43] Kunio Nakamura et al. "Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis". In: *NeuroImage: Clinical* 4 (Jan. 2014), pp. 10–17.

[44] Marco Battaglini, Mark Jenkinson, and Nicola De Stefano. "Evaluating and reducing the impact of white matter lesions on brain volume measurements". In: *Human Brain Mapping* 33 (9 Sept. 2012), pp. 2062–2071.

[45] Michael Kistler et al. "The virtual skeleton database: an open access repository for biomedical research and collaboration." In: *Journal of medical Internet research* 15.11 (2013), e245.

[46] Roberto Souza et al. "An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement". In: *NeuroImage* 170 (Apr. 2018), pp. 482–494.

[47] Olivier Commowick et al. "Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure". In: *Scientific Reports* 8 (1 Dec. 2018), p. 13650.

[48] Aaron Carass et al. "Longitudinal multiple sclerosis lesion segmentation data resource". In: *Data in Brief* 12 (June 2017), pp. 346–350.

[49] Hugo J. Kuijf et al. "Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge". In: *IEEE Transactions on Medical Imaging* 38 (11 Nov. 2019), pp. 2556–2568.

[50] Ian B. Malone et al. "MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset". In: *NeuroImage* 70 (Apr. 2013), pp. 33–36.

[51] R. Guerrero et al. "White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks". In: *NeuroImage: Clinical* 17 (2018), pp. 918–934.

[52] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2980–2988.

[53] Carole H. Sudre et al. "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (2017), pp. 240–248.

[54] Tanya Nair et al. "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation". In: *Medical image analysis* 59 (2020), p. 101557.

[55] Ewelina Marciniewicz et al. “The role of MR volumetry in brain atrophy assessment in multiple sclerosis: A review of the literature.” In: *Advances in Clinical and Experimental Medicine: Official Organ Wroclaw Medical University* 28.7 (2019), pp. 989–999.

[56] Martin Rajchl et al. “NeuroNet: Fast and Robust Reproduction of Multiple Brain Image Segmentation Pipelines”. In: *arXiv preprint arXiv:1806.04224* (2018).

[57] Leonie Henschel et al. “FastSurfer - A fast and accurate deep learning based neuroimaging pipeline”. In: *NeuroImage* 219 (Oct. 2020), p. 117012.

[58] Valeriia Abramova et al. “Hemorrhagic stroke lesion segmentation using a 3D U-Net with squeeze-and-excitation blocks”. In: *Computerized Medical Imaging and Graphics* 90 (2021), p. 101908.

[59] Liliana Valencia et al. “Evaluating the use of synthetic T1-w images in new T2 lesion detection in multiple sclerosis”. In: *Frontiers in Neuroscience* 16 (2022).

[60] Davide Bacciu, Paulo JG Lisboa, and Alfredo Vellido. *Deep Learning In Biology And Medicine*. World Scientific, 2022.