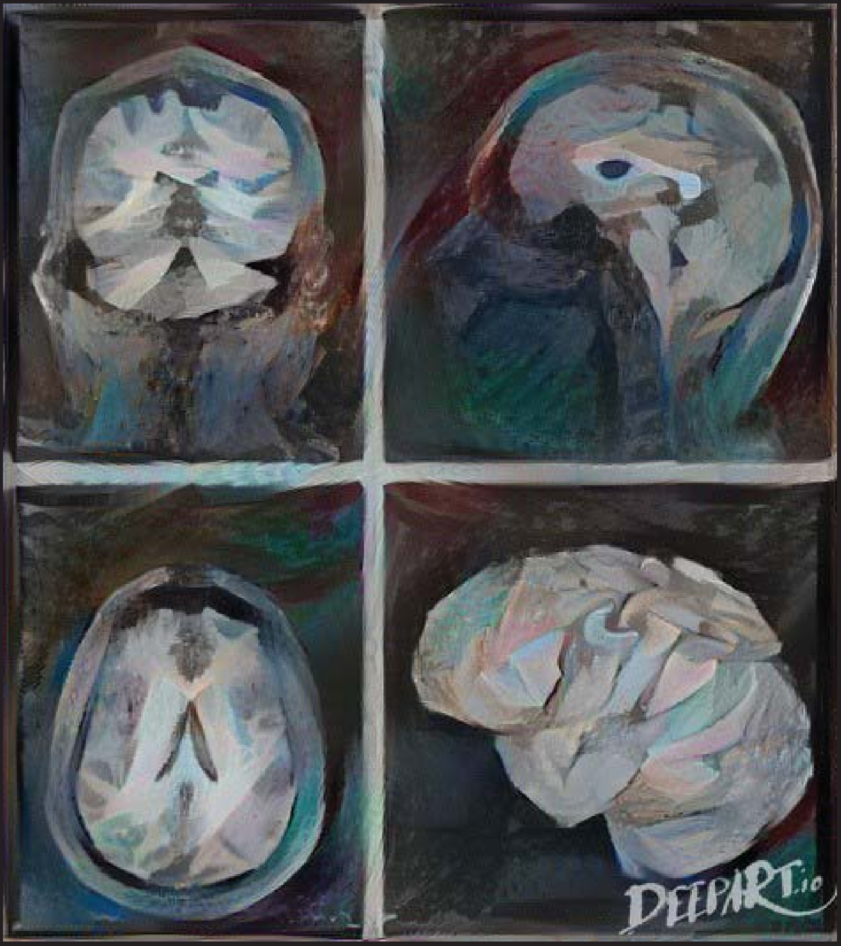
DEEP LEARNING FOR ATROPHY QUANTIFICATION IN BRAIN MAGNETIC RESONANCE IMAGING

Jose Bernal Moyano

Universitat de Girona

DOCTORAL THESIS



Deep learning for atrophy quantification in brain magnetic resonance imaging

Jose Bernal Moyano

2020

Universitat
de Girona

DOCTORAL THESIS

Deep learning for atrophy quantification in brain magnetic resonance imaging

Jose Bernal Moyano

18-08-2020

2020

DOCTORAL PROGRAM in TECHNOLOGY

Supervised by:
Dr Arnau Oliver
Dr Xavier Lladó

Thesis submitted to the University of Girona for the degree of Doctor

A mi familia por su apoyo incondicional.

Acknowledgments

This doctoral thesis reflects my accomplishments over the last three years...perhaps many more since this life plan started long ago. Evidently, I would not have gotten here should I have not had all those people who contributed, in a way or another, to the successful completion of this work. Foremost, I would like to express my immense gratitude to my supervisors, Dr Xavier Lladó and Dr Arnau Oliver, for the opportunity to work with them in such a fascinating topic, guiding me throughout this whole learning process, and their dedication and continuous support regardless of some of my decisions. This work would not have been possible without their help.

My sincere gratefulness goes to Dr Sergi Valverde who also mentored me during this doctoral thesis. His ideas, insights, and advice helped me to grow as a researcher.

I would like to thank my friends who helped me to achieve this goal with their advice, suggestions, endless discussions, friendship, and support during the past three years (or more). In particular, I thank Deisy, Claudia, Kaisar, and Oleksii. I am thankful for having you all by my side during this part of my life.

My deepest appreciation goes to my mother for her hard effort, my brother for being always by my side, my family for supporting and believing in me, Mauro for his company. I cannot wait to see what comes after this!

Research activities

Main research outcomes

The following research works are the pillars of the presented thesis:

- **Bernal, J.**, Valverde, S., Oliver, A., & Lladó, X. (2020). Deep learning for quantifying longitudinal cerebral atrophy in brain magnetic resonance imaging. Under preparation.
- **Bernal, J.**, Kushibar, K., Clèrigues, A., Oliver, A., & Lladó, X. (2020). Deep learning for medical imaging. In: Bacciu D., Lisboa P.J.G., Vellido A., eds. *Deep Learning in Biology and Medicine*. Singapore: World Scientific Publishing. Under review.
- **Bernal, J.**, Valverde, S., Kushibar, K., Oliver, A., & Lladó, X. (2019). Generating controlled atrophy change evaluation environments on brain MR using convolutional neural networks and segmentation priors. Under review in Neuroinformatics. Quality index: [JCR IF 5.127, Q1(9/106)]
- **Bernal, J.**, Kushibar, K., Cabezas, M., Valverde, S., Oliver, A., & Lladó, X. (2019). Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access*, 7, 89986-90002. Quality index: [JCR IF 4.098, Q1(23/155)]
- **Bernal, J.**, Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., & Lladó, X. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, 95, 64-81. Quality index: [JCR IF 3.574, Q1(5/26)]

Related journal publications

Other journal publications related to this PhD are as follows:

- Clèrigues, A., Valverde, S., **Bernal, J.**, Freixenet, J., Oliver, A., & Lladó, X. (2020). Acute and sub-acute stroke lesion segmentation from multimodal MRI. *Computer Methods and Programs in Biomedicine*, 194, 105521. Quality index: [JCR IF 3.424, Q1(6/26)]

- Kushibar, K., Valverde, S., González-Villà, S., **Bernal, J.**, Cabezas, M., Oliver, A., & Lladó, X. (2019). Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Scientific reports*, 9(1), 6742. Quality index: [JCR IF 4.011, Q1(15/69)]

- Clèrigues, A., Valverde, S., **Bernal, J.**, Freixenet, J., Oliver, A., & Lladó, X. (2019). Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Computers in Biology and Medicine*, 115, 103487. Quality index: [JCR IF 2.286, Q1(52/106)]

- Kushibar, K., Valverde, S., González-Villà, S., **Bernal, J.**, Cabezas, M., Oliver, A., & Lladó, X. (2018). Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Medical Image Analysis*, 48, 177-186. Quality index: [JCR IF 8.880, Q1(5/134)]

Conference participation

Works presented in national and major international conferences as oral or poster presentations:

- **Bernal, J.**, Kushibar, K., Salem, M., Clèrigues, A., Valverde, S., Cabezas, M., Salvi, J., Oliver, A. & Lladó, X. A hybrid multi-atlas and convolutional neural network based framework for six-month infant brain magnetic resonance image tissue segmentation. MICCAI Grand Challenge on infant brain MRI segmentation. MICCAI 2019. Shenzhen, China. 2019.

- Cabezas, M., Valverde, S., Clèrigues, A., Salem, M., Kushibar K., **Bernal, J.**, Oliver, A., Salvi, J., Lladó, X. Brain tumour segmentation and prediction via CNNs. MICCAI Grand Challenge on multimodal brain tumor segmentation challenge. MICCAI 2019. Shenzhen, China. 2019.

- Valverde, S., Cabezas, M., Salem, M., Kushibar, K., Clèrigues, A., **Bernal, J.**, Salvi, J., Oliver, A., Llaló, X.. Big Data, Intel · ligència Artificial, Machine Learning, etc. Què és què?. Jornades d'Esclerosis Múltiple del Mediterrani. Girona, Spain. 2019.

- **Bernal, J.**, Kushibar, K., Salem, M., Clèrigues, A., Valverde, S., Cabezas, M., Salvi, J., Oliver, A. & Lladó, X. A hybrid multi-atlas and convolutional neural network based framework for six-month infant brain magnetic resonance image tissue segmentation. MICCAI Grand Challenge on infant brain MRI segmentation. MICCAI 2019. Shenzhen, China. 2019.

- **Bernal, J.**, Salem, M., Kushibar, K., Clèrigues, A., Valverde, S., Cabezas, M., Gonzáles-Villà, S., Salvi, J., Oliver, A. & Lladó, X. MR brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors. MICCAI Grand Challenge on MR brain segmentation. MICCAI 2018. Granada, Spain. 2018.

- Clèrigues, A., Valverde, S., **Bernal, J.**, Kushibar, K., Cabezas, M., Oliver, A. & Lladó, X. Cascade of convolutional neural networks for acute stroke anatomy differentiation. MICCAI Grand Challenge on ischaemic stroke lesion segmentation in medical imaging. MICCAI 2018. Granada, Spain. 2018.

- Cabezas M., Valverde S., González-Villà S., Clèrigues, A., Salem, M., Kushibar, K., **Bernal, J.**, Oliver, A. & Lladó, X. Survival prediction using ensemble tumor segmentation and transfer learning. MICCAI Grand Challenge on multimodal brain tumor segmentation challenge 2018 in medical imaging. MICCAI 2018. Granada, Spain. 2018.

- Clerigues, A., Valverde, S., **Bernal, J.**, Pareto, D., Vilanova, J. C., Ramio-Torrenta, L., Rovira, A., Oliver, A. & Llado, X. (2018). A quantitative analysis of deep learning methods for multiple sclerosis white matter lesion segmentation. *Multiple Sclerosis*, 24, 637-638. Quality index: [JCR IF 5.280, Q1(22/197)]

- **Bernal, J.**, Kushibar, K., Valverde, S., Cabezas, M., Gonzáles-Villà, S., Salem, M., Salvi, J., Oliver, A. & Lladó, X.. Six-month infant brain tissue segmentation using three dimensional fully convolutional neural networks and pseudo-labelling. MICCAI Grand Challenge on infant brain MRI segmentation. MICCAI 2017. Quebec, Canada. 2017

- Valverde S., Cabezas M., **Bernal J.**, Kushibar, K., González-Villà, S., Salem, M., Salvi, J., Oliver, A. & Lladó X. White matter hyperintensities segmentation using a cascade of three convolutional neural networks. MICCAI Grand Challenge on White Matter Hyperintensities Segmentation. MICCAI 2017. Quebec, Canada. 2017.

- Lladó, X., Valverde, S., Cabezas, M., González-Villa, S., Salem, M., Kushibar, K., **Bernal, J.**, Freixenet, J., Salvi, J., Oliver, A. Neuroimatge de la Neurodegeneració: situació actual i futur. Jornades d'Esclerosis Múltiple del Mediterrani. Girona, Spain. 2017.

Teaching experience

- Medical image segmentation and applications course of the Erasmus Mundus Master in Medical Imaging and Applications (Sept 2017 - Jan 2018). Course organisers: Xavier Lladó and Robert Martí.

Acronyms

$\theta_{\mathbf{flip}}$ Flip angle

ADNI Alzheimer's Disease Neuroimaging Initiative

AD Alzheimer's disease

ANTs Advanced Normalization Tools

ASD Average surface distance

BET Brain extraction tool

BEaST Brain extraction based on nonlocal segmentation technique

CGAN Conditional generative adversarial network

CNN Convolutional neural network

CSF Cerebrospinal fluid

CT Computed tomography

DF Deformation field

DSC Dice similarity coefficient

ELU Exponential linear units

FAST FMRIB's Automated Segmentation Tool

FCNN Fully convolutional neural network

FC Fully connected

FIRST

FLAIR Fluid attenuated inversion recovery

FLIRT FMRIB's Linear Image Registration Tool

FNIRT FMRIB's Non-linear Image Registration Tool

FSL FMRIB Software Library

GAN Generative adversarial network

GM Grey matter

GT Ground truth

IBSR Internet brain segmentation repository

LABEL Learning algorithm for brain extraction and labeling

MAE Median absolute error

MHD Modified Hausdorff distance

MICCAI2012 MICCAI multi-atlas labeling challenge

MICCAI Medical Image Computing and Computer-Assisted Intervention

MNI Montreal Neurological Institute
MRBrainS18 MR brain segmentation challenge 2018
MRI Magnetic resonance imaging
MSE Mean square error
MS Multiple sclerosis
MS Multiple sclerosis
NMR Nuclear magnetic resonance
OASIS Open Access Series of Imaging Studies
PBVC Percentage of brain volume change
PD Proton density
PET Positron-emission tomography
PReLU Parametric rectified linear units
RAVEL Removal of artificial voxel effect by linear regression
ROBEX Robust brain extraction
ROI Region of interest
RWMSE Region-wise mean square error
ReLU Rectified linear units
SIENA Structural image evaluation, using normalisation, of atrophy
SPM Statistical parametric mapping
SSIM Structural similarity coefficient
$\mathbf{T}_E$ Echo time
$\mathbf{T}_R$ Repetition time
WM White matter
iSeg2017 Six-month infant brain MRI segmentation challenge 2017
iSeg2019 Six-month infant brain MRI segmentation challenge 2019

Contents

List of Figures

List of Tables

Abstract

Cerebral atrophy is a neuroimaging feature of ageing and diverse brain pathologies that indicate of loss of neurons and their connections. Its quantification plays a fundamental role in neuroinformatics since it permits studying brain development, diagnosing brain diseases, assessing their progression, and determining the effectiveness of novel treatments to counteract these brain diseases. However, this is still an open and challenging problem in medical image analysis.

In this doctoral thesis, we question whether deep learning methods can be used for better estimating cerebral atrophy from magnetic resonance images at both cross-sectional and longitudinal levels. To fulfil this goal, we initially reviewed the literature on deep learning for brain medical image analysis to discover potential lines to explore. Our revision revealed that a direct comparison cannot be established between methods due to potential overfitting and there are no longitudinal atrophy quantification strategies using deep learning.

Overfitting to challenge data hinders comparing architectures. To overcome this issue, we built a framework for evaluating methods for brain tissue segmentation quantitatively using the same evaluation dataset, metrics, tasks, and pre- and post-processing. Our results suggest that deep learning can achieve state-of-the-art results in cross-sectional tissue segmentation and that certain design directives are experimentally better than others. Based on our analysis, we devised three proposals for three Challenges of the Conference in Medical Image Computing and Computer-Assisted Intervention between 2017 and 2019. In all three events, we achieved a compelling performance.

The lack of annotated longitudinal atrophy datasets prevents determining whether a certain method is accurate, and also training deep learning methods for detecting brain changes. To cope with this issue, we crafted a deep learning method for image synthesis allowing generating a plethora of scans for which the induced changes would be known. We provided our model with baseline scans and real follow-up segmentation maps and observed that our framework produced synthetically similar outputs, even when training and testing on different but harmonised domains. Moreover, our proposal induced changes that were detected by validated cross-sectional and longitudinal methods.

Finally, we presented an application of our data generator for training a deep learning method for atrophy quantification with many more samples. Our longitudinal atrophy quantification proposal involved data harmonisation, non-linear registration through deep learning, and estimation of brain edge displacement as a surrogate measure of brain atrophy through integrating the Jacobian of the resulting deformation fields. We tested our method on three cohorts of patients with Alzheimer's disease, dementia, and multiple sclerosis and compared its performance against relevant methods. We trained our method on a single domain and tested it on the same and other two domains. Our proposal worked well in these datasets and was particularly suitable for discerning between multiple sclerosis patients and control subjects.

This PhD thesis forms part of multiple projects carried out by our research group. To enable reproducibility and work continuation, we have released our development to the public. We believe that the proposed cross-sectional and longitudinal methods can be beneficial for the research and clinical community.

Resumen

La atrofia cerebral es una consecuencia de la pérdida de neuronas y sus conexiones debido al envejecimiento y a diversas patologías cerebrales. Su cuantificación en neuroimágenes juega un papel fundamental en la neuroinformática ya que permite estudiar el desarrollo del cerebro, diagnosticar enfermedades cerebrales, evaluar su progresión y determinar la eficacia de nuevos tratamientos para contrarrestarlas. Sin embargo, obtener una cuantificación precisa sigue siendo un problema y un reto abierto en el análisis de imágenes médicas.

En esta tesis doctoral, cuestionamos si los métodos de aprendizaje profundo (*Deep Learning*) pueden utilizarse para estimar mejor la atrofia cerebral a partir de imágenes de resonancia magnética tanto a nivel transversal como longitudinal. Para cumplir este objetivo, inicialmente revisamos el estado del arte en aprendizaje profundo aplicado al análisis de imágenes médicas cerebrales con el fin de descubrir potenciales líneas de investigación. Nuestra revisión reveló que no se podía establecer una comparación directa entre los métodos disponibles en la literatura debido al sobreajuste y a que no existían estrategias de cuantificación de la atrofia longitudinal que utilicen el aprendizaje profundo.

Para solucionar el problema del sobreajuste a los datos de entrenamiento propusimos y construimos un marco de referencia común para evaluar cuantitativamente métodos de segmentación de tejido cerebral utilizando el mismo conjunto de datos de evaluación, mismas métricas, mismas aplicaciones y mismo pre- y post-procesamientos. Los resultados obtenidos sugieren que el aprendizaje profundo puede lograr resultados de vanguardia en la segmentación transversal de tejidos y, también, que ciertas directivas de diseño son experimentalmente mejores que otras. Basándonos en nuestro análisis, ideamos tres propuestas para tres competencias internacionales de la *Medical Image Computing and Computer Assisted Intervention* (MICCAI, por sus siglas en inglés) entre 2017 y 2019. En los tres eventos, obtuvimos resultados competitivos.

Por otra parte, la falta de conjuntos de datos de atrofia longitudinal con anotaciones manuales impide realizar estudios cuantitativos, así como entrenar métodos de aprendizaje profundo para detectar cambios en el tejido cerebral. Para hacer frente a este problema, desarrollamos un método de aprendizaje profundo para la síntesis de imágenes que permite generar multiples imágenes de resonancia magnética cuyos cambios son inducidos y conocidos de antemano. Proporcionamos a nuestro modelo imágenes de referencia y mapas de segmentación de seguimientos reales, observando que producía resultados sintéticamente similares a los reales, incluso al entrenar y probar la generación en dominios diferentes pero armonizados. Además, nuestra propuesta permite generar cambios que pueden ser detectados por métodos transversales y longitudinales validados.

Finalmente, presentamos una aplicación directa de nuestro generador de datos para entrenar un método de aprendizaje profundo para la cuantificación de la atrofia cerebral en estudios longitudinales. Nuestra propuesta de cuantificación de la atrofia longitudinal se basa en la armonización de datos, el registro no lineal por medio del aprendizaje profundo y la estimación del desplazamiento de los contornos del cerebro como medida de atrofia cerebral, realizado por medio de la integración de los Jacobianos de los campos de deformación resultantes del registro. Este método se evaluó en tres cohortes de pacientes con Alzheimer, demencia y esclerosis múltiple, comparando su rendimiento con métodos relevantes del estado del arte. Los resultados obtenidos por nuestra propuesta en estos conjuntos de datos son prometedores, siendo particularmente útil para discernir entre los pacientes de esclerosis múltiple y los sujetos de control.

Esta tesis doctoral forma parte de los múltiples proyectos llevados a cabo por el grupo de investigación VICOROB. Para permitir la reproducibilidad y la continuación de la investigación realizada, nuestros desarrollos se han hecho públicos en el repositorio del grupo. Creemos que los métodos transversales y longitudinales propuestos pueden ser beneficiosos para la comunidad investigadora y clínica.

Resum

L'atròfia cerebral és una conseqüència de la pèrdua de neurones i de les seves connexions a causa de l'envelliment i a diverses patologies cerebrals. La seva quantificació en neuroimatge juga un paper fonamental en la neuroinformàtica, ja que permet estudiar el desenvolupament del cervell, diagnosticar malalties cerebrals, avaluar la seva progressió i determinar l'eficàcia de nous tractaments per a contrarestar-les. No obstant això, obtenir una quantificació acurada continua sent un problema i un repte obert en l'anàlisi d'imatges mèdiques.

En aquesta tesi doctoral qüestionem si els mètodes d'aprenentatge profund (*Deep Learning*) es poden utilitzar per estimar millor l'atròfia a partir d'imatges de ressonància magnètica, tant en estudis transversals com en estudis longitudinals. Per complir aquest objectiu, inicialment vam revisar l'estat de l'art sobre l'aprenentatge profund aplicat a l'anàlisi d'imatges mèdiques cerebrals, amb l'objectiu de descobrir potencials línies de recerca. La nostra revisió va revelar que no es podia establir una comparació directa entre els mètodes disponibles a la literatura degut al sobre-ajustament a les dades experimentals (*overfitting*) i que no existien estratègies de quantificació de l'atròfia longitudinal que utilitzessin l'aprenentatge profund.

Per solucionar el problema del sobre-ajustament a les dades d'entrenament, doncs, vam proposar i construir un marc de referència comú per tal de poder avaluar quantitativament els mètodes de segmentació de teixit cerebral, utilitzant el mateix conjunt de dades d'avaluació i les mateixes mètriques, aplicacions i pre- i post-processaments. Els resultats obtinguts suggereixen que l'aprenentatge profund pot aconseguir resultats d'avantguarda en la segmentació transversal de teixits i, també, que certes directives de disseny són experimentalment millors que d'altres. Basant-nos en aquestes anàlisis, vam idear tres propostes diferents per a tres competicions internacionals lligades a la Conferència *Medical Image Computing and Computer Assisted Intervention* (MICCAI, per les sigles en anglès) entre 2017 i 2019. En els tres esdeveniments, vam aconseguir obtenir resultats competitius.

D'altra banda, la manca de conjunts de dades d'atròfia longitudinal amb anotacions manuals impedeix realitzar-ne estudis quantitatius, i també entrenar mètodes d'aprenentatge profund per a la detecció de canvis en el teixit cerebral. Per fer front a aquest problema, vam elaborar un mètode d'aprenentatge profund per a la síntesi d'imatges que permet generar múltiples imatges de ressonància magnètica on els canvis són induïts i coneguts per endavant. Proporcionem al nostre model imatges de referència i mapes de segmentació d'estudis longitudinals reals, i obtenim resultats sintètics similars als reals, fins i tot a l'entrenar i provar la generació en dominis diferents però harmonitzats. A més, la nostra proposta permet generar canvis que poden ser detectats pels mètodes transversals i longitudinals validats.

Finalment, presentem una aplicació directe del nostre generador de dades sintètiques per a l'entrenament d'un mètode d'aprenentatge profund capaç de quantificar l'atròfia cerebral en anàlisis longitudinals. Aquesta proposta es basa en una harmonització de dades, un registre no lineal per mitjà de l'aprenentatge profund i la subseqüent estimació del desplaçament de les contorns del cervell com a mesura de l'atròfia cerebral, fet que es realitza mitjançant la integració dels Jacobians dels camps de deformació resultants del registre. Aquest mètode s'ha avaluat en cohorts de pacients amb Alzheimer, demència i esclerosi múltiple, comparant el seu rendiment amb mètodes rellevants de l'estat de l'art. Els resultats obtinguts per la nostra proposta en aquests conjunts de dades són prometedors, essent particularment útil per a discernir entre pacients d'esclerosi múltiple i subjectes de control.

Aquesta tesi doctoral forma part de múltiples projectes duts a terme pel grup de recerca ViCOROB. Per permetre la reproductibilitat i la continuació de la recerca feta, els nostres desenvolupaments s'han fet públics en el repositori del grup. Creiem que els mètodes transversals i longitudinals proposats poden ser beneficiosos per a la comunitat investigadora i clínica.

Chapter 1

Introduction

In this chapter, we introduce the reader to the research context, present the main thesis goals and specific steps to reach them, and summarise the main structure of the present thesis.

1.1 Cerebral atrophy

1.1.1 What is cerebral atrophy?

The human brain is an organ located inside the cranium that forms part of the central nervous system. The brain consists of two tissues: grey matter – neuronal cell bodies – and white matter – mainly myelinated axon tracts [1]. This organ is surrounded by cerebrospinal fluid, which provides it with mechanical protection [1] and helps it to drain toxins and waste [2]. A high-level scheme of a human brain is depicted in Fig. 1.1.

The loss of neurons and their connections and, therefore, a reduction in the volume of the grey and white matter, is referred to as **cerebral atrophy**. Because the total volume of the brain tissue, cerebrospinal fluid, and intracranial blood is fixed in adult brains, according to the Monro-Kellie hypothesis [3], cerebrospinal fluid volume increases as brain tissue volume declines as a mechanism for maintaining a normal intracranial pressure. Although brain tissue loss is thought to be a direct consequence of the normal ageing process [4], it is also a common neuroimaging feature of multiple disorders affecting the brain, as shown in Fig. 1.2.

Cerebral atrophy can be focal and rapid as a result of a head injury [5], radiotherapy [6], and stroke [7]; or diffuse and slow as a result of ageing, cerebral small vessel disease [8], schizophrenia [9,10], Alzheimer's disease [11], and multiple sclero-

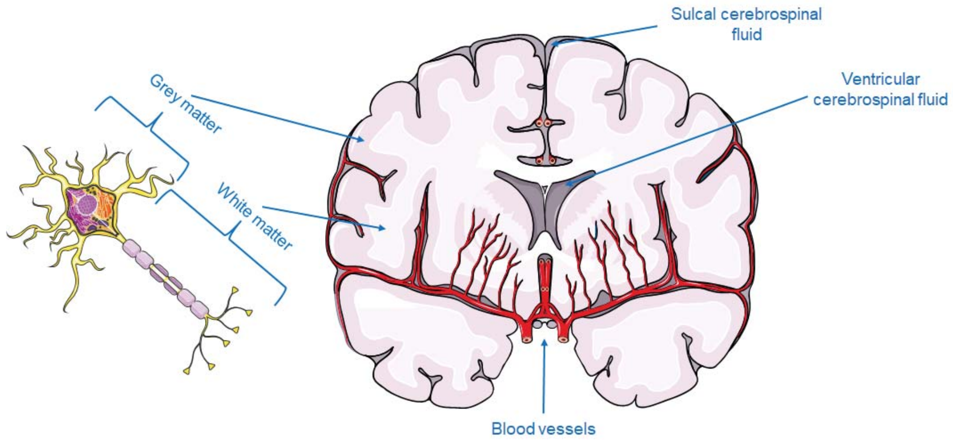


Figure 1.1: High-level scheme of a human brain. The brain is formed by the grey matter, which contains neuronal cell bodies, and white matter, which contains mainly myelinated axon tracts. The brain is surrounded by the cerebrospinal fluid.

sis [12,13], among other pathologies affecting brain tissues. Hence, providing medical doctors with accurate and precise cerebral atrophy measurements at cross-sectional (one time point) and longitudinal (variations over time) levels is fundamental for shedding light into their relationship with neurological diseases, monitoring their progression, and assessing treatment effectiveness [14–20].

1.1.2 Brain pathologies causing cerebral atrophy

Dementia

Dementia is an umbrella term in including cognitive decline, physical frailty, onset depression, and dependency that, ultimately, limits daily life. At a global scale, dementia is the 7th leading cause of death, approximately 50M people suffer from it, 10M new cases appear every year (a new case every three seconds), costs about 818 billion dollars (to carers mostly), affects both patients and their families, and, unsettlingly enough, these figures are expected to triple by 2050 [21]. The distribution of cases of dementia worldwide are depicted in Fig. 1.3. Multiple pathological processes and injuries compromising the optimal functioning of the brain result in dementia: endothelial dysfunction may lead to up to 45% of dementias [8], Alzheimer's disease contributes to 60-70% of the cases [21], and strokes double dementia risk [22]. Despite being a worldwide matter, little is known about its causes since much of it is clinically silent and late, and there is no clear way to treat it nowadays. However,

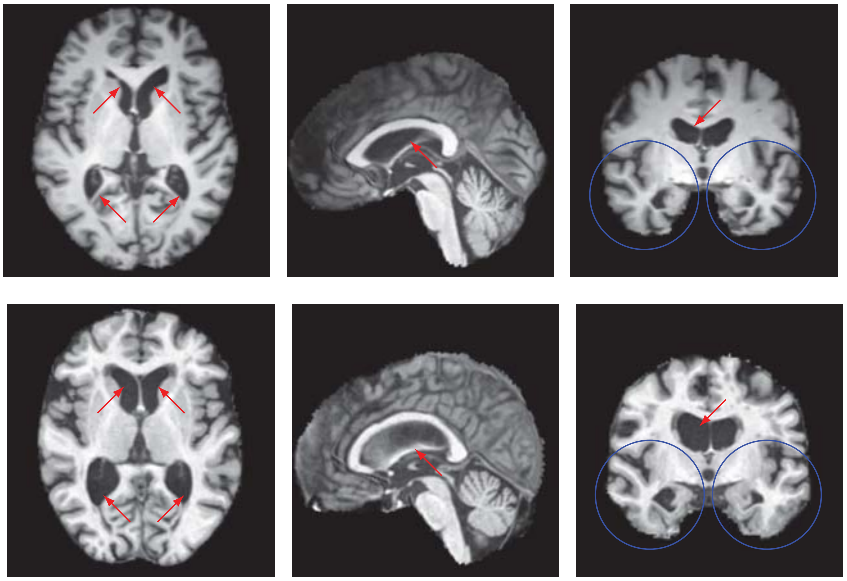

(a) Axial (b) Sagittal (c) Coronal

Figure 1.2: Brain atrophy in control (top) and Alzheimer's disease (bottom) subjects in their 70s. In the three orthogonal views, note the enlargement of the lateral ventricles and widening of sulci in the patient with Alzheimer's disease compared to that of the control subject. Red arrows point to the lateral ventricles in the three views. In coronal view, note the atrophy on the medial temporal lobes (blue circles). These scans were skull stripped and co-registered for enhancing visualisation.

cerebral atrophy is a feature of dementia [23–26] and its progress in middle-aged subjects may be associated with the future development of dementia [25]. Therefore, efforts for quantifying brain tissue loss and understanding the causes leading to such volume decline may lead to better characterisation and monitoring of dementia.

Multiple sclerosis

Multiple sclerosis is a common chronic immune-mediated neurological disease of the central nervous system [28, 29], characterised by the formation of lesions or plaques that damage the myelin sheaths present in nerve cells in the central nervous system – brain and spinal chord. This demyelineation process hampers axonal transmissions and clinically manifests in cognitive decline and physical disability [30, 31]. The distribution of cases of multiple sclerosis worldwide is depicted in Fig. 1.4. According to the Multiple Sclerosis International Federation, the number of people with MS worldwide are approximately 2.2M [32]. Genetic predisposition, biological sex, and geographical location are etiological factors of multiple sclerosis [33]. Although

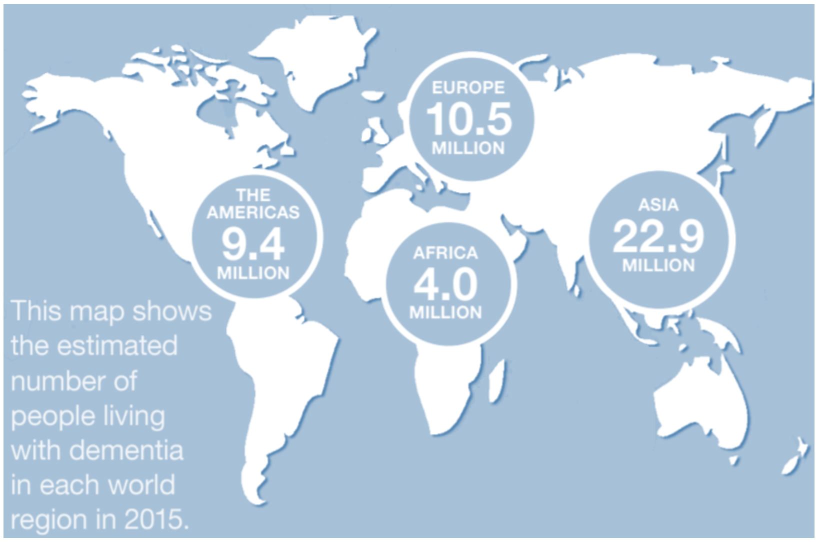

EUROPE
10.5
MILLION
THE AMERICAS
9.4
MILLION
AFRICA
4.0
MILLION
ASIA
22.9
MILLION
This map shows the estimated number of people living with dementia in each world region in 2015.

Figure 1.3: Figures of dementia worldwide according to the World Health Organisation [21, 27]. Approximately 50% of the cases of dementia worldwide are in Asia, 40% split between America and Europe, and less than 10% in Africa.

there are various treatment options, side effects and limited effectiveness result in poor treatment adherence [33]. Brain volume measurements are a key part in studies of this neurodegenerative disease as they appear in all patients, are associated with clinical risk factors, and predict disease evolution [34–37]. Therefore, accurate and reliable brain atrophy quantification methods may help to monitor disease progression and assess the effectiveness of new treatments for controlling it.

1.2 Brain image analysis in cerebral atrophy

1.2.1 Brain magnetic resonance imaging

Medical imaging comprises a wide range of medical techniques which allow visualising internal body structures. They are preferred over biopsies as imaging methods reduce collateral risks surgical procedures may involve. Some of these imaging techniques require the body to receive low doses of radiation (e.g. x-ray or computed tomography scans) which may not be prejudicial – in principle – but should be reduced as much as possible [39]. Unlike these technologies, magnetic resonance imaging (MRI) does not require exposing the body to ionic radiation. Instead, magnetic waves stimulate hydrogen atoms in molecules present in the body using the property of nuclear magnetic resonance (NMR). NMR is the phenomenon in which

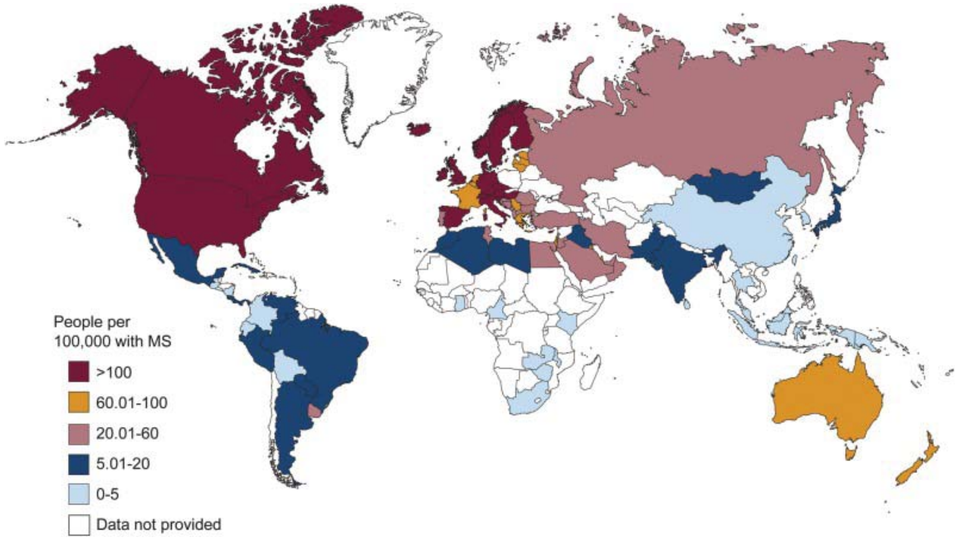


Figure 1.4: Figures of multiple sclerosis worldwide according to the Multiple Sclerosis (MS) International Federation [38]. Multiple sclerosis is more frequent in North America, Europe, and Oceania compared to Asia (except for Russia), South America, and Africa, where it is unusual.

magnetic nuclei in a magnetic field absorbs and re-emits electromagnetic radiation at a specific resonance frequency. The radiation energy depends on the magnetic field and atom properties.

MRI relies on the magnetic properties of hydrogen atoms present in water molecules in the body to produce images. Water molecules contain two hydrogen atoms with one electron and one proton. The protons spin around their axes in random directions. When the body is placed in a strong magnetic field, the axes of the protons align up in the direction of the field (longitudinal magnetisation). A radio-frequency pulse source is used to deflect the average magnetic vector of the protons, i.e. the average magnetic momentum vector turns θ_{flip} degrees away from the magnetic field vector (traverse magnetisation). At this point, the protons spin together in resonance with the radio-frequency. When the radio-frequency source is switched off, the magnetic vectors realign to the initial magnetic field in which the body is immersed. The time from which the average momentum goes from θ_{flip} degrees to 0 degrees is called the relaxation time. During this relaxation time, protons release energy in the form of radio frequency signals which is subsequently acquired by receiver coils to generate the MR image [40, 41]. The overall magnetisation process is repeated by successively applying radio-frequency pulse sequences with a delay between them and measuring the transverse relaxation with a recess between them, referred to as repetition time (T_R) and echo time (T_E), respectively.

The relaxation can be seen as two stages: T1 and T2 relaxations. As the radiofrequency is gradually eliminated, the transverse magnitude decays and the protons stop being in resonance. This process is known as spin-spin or T2 relaxation. Then, the protons move from a high energy state to a low energy one restoring the longitudinal magnetisation and, at the same time, liberating energy. This process is called spin-lattice or T1 relaxation. These two relaxation times vary from one tissue to another and, thus, are key elements of the imaging process [42].

Different acquisition sequences are determined by emphasising on T1 or T2 relaxations and varying the values for T_R and T_E. In T1-weighted (T1-w) acquisition sequences, differences in spin-lattice relaxation time are accentuated and both T_R and T_E are short ($T_R < 1000$ms, $T_E > 30$ms). T1-w acquired scans offer contrast between grey matter from white matter and are often used in brain tissue segmentation when only one modality is used (monospectral). In T2-weighted (T2-w) acquisition sequences, differences in spin-spin relaxation time are accentuated and both the T_R and T_E are long ($T_R > 2000$ms, $T_E > 80$ms). The transverse magnetisation still exists in fluids after a long T_E – fluids appear hyperintense –, but it does not in fatty tissues and, thus, the former appear hyperintense while the latter hypointense. T2-w scans permit distinguishing between normal-appearing tissues and regions of abnormal fluid content and, thus, area suitable for detecting brain lesions. More advanced acquisition methods are being used nowadays such as fluid attenuated inversion recovery (FLAIR) [43] in which fluids are suppressed from the image. This imaging modality has been widely used to classify periventricular hyperintense lesions, such as multiple sclerosis plaques [44]. An illustrative example of the appearance of T1-w, T2-w and FLAIR scans is shown in Fig. 1.5.

1.2.2 Visual clinical ratings and their pitfalls

Pioneer works in the field contemplated devising qualitative clinical ratings to discern between normal and abnormal brain atrophy [4,45–47]. The process consisted of visually inspecting anatomical landmarks, such as cerebrospinal fluid spaces [4, 45], frontal and parietal cortex [45], and medial temporal lobes [45–47], and grading their appearance based on prior anatomical knowledge or against reference templates [4, 48] using a discrete rating scale. For example, the Wahlund visual rating scale [4] consisted of identifying four standard axial slices from an incoming scan, as shown in Fig. 1.6, examining cerebrospinal fluid spaces in six regions (lateral ventricles, inter-hemispheric fissure anterior to the corpus callosum, left and right Sylvian fissures, occipital sulci, frontal sulci, and parietal sulci), and rating each of them based on their appearance compared to the normal case shown in Fig. 1.6 using a five-point scale (0 - normal size; 1 - normal, slightly enlarged; 2 - larger than normal; 2.5 - considerably larger than normal; 3 - extremely larger than normal). Note that

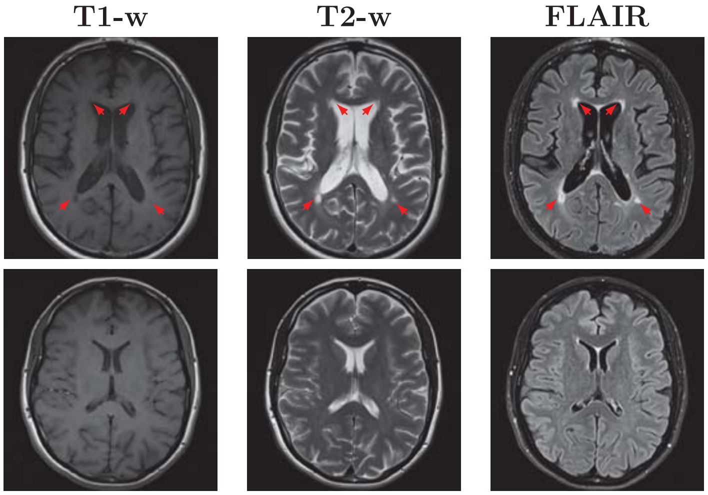


Figure 1.5: Magnetic resonance imaging sequences on a multiple sclerosis patient (top row) and a control subject (bottom row). The patient with multiple scleroris exhibits periventricular white matter hyperintensities (red arrows). Fluids and lesions appear hypointense in T1-w and hyperintense in T2-w and FLAIR.

the process requires serious training to select these references slices appropriately, recognise these anatomical regions unequivocally, and discern between "normal" and "abnormal".

Visual clinical ratings are relatively fast as raters make these assessments based on a few standard slices, a crucial factor in emergencies, and are resilient to data quality as a radiologist can holistically provide an estimate of the brain tissue loss even when reference points are not clearly visible. Nonetheless, their reliability depends on the expertise of the radiologists and their physical limitations: fatigued and inexperienced raters may exhibit lower diagnostic performance [49, 50]. Moreover, the development of these ratings target specific populations (patients with schizophrenia [45] or Alzheimer's disease [46]) and, hence, may not be useful to other samples and are limited by "flooring" or "ceiling" effects due to their discrete nature [51].

1.2.3 Computational approaches and their pitfalls

A plethora of computational approaches have been proposed throughout the years to quantify automatically brain volume at cross-sectional and longitudinal levels and overcome the aforementioned limitations of visual ratings [12].

Cross-sectional studies carry out brain volumetry at a specific time-point, i.e.

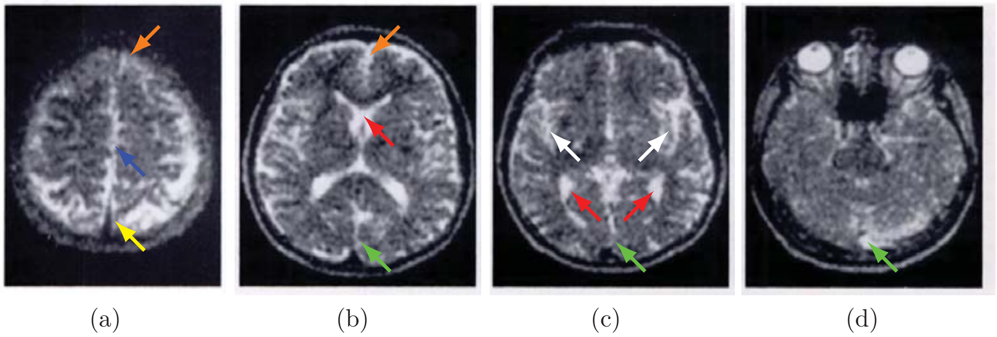


Figure 1.6: Four standard slices used for grading cerebral atrophy according to the Wahlund scale [4]. These four axial slices show cerebrospinal fluid spaces in six regions: inter-hemispheric fissure anterior to the corpus callosum (blue arrow), lateral ventricles (red arrows), Sylvian fissures (white arrows), occipital sulci (green arrows), frontal sulci (orange arrow), parietal sulci (yellow arrow). Images extracted from the original paper [4].

they do not incorporate temporal information. For that, intracranial regions are segmented into cerebrospinal fluid, grey matter, and white matter prior to estimating their volume. Developing traditional tissue segmentation methods required analysing and understanding the problem at hand and the images to deal with carefully, laying down assumptions about the data, and, finally, engineering an algorithm that would use such information to segment the regions of interest. For example, algorithms would assume brain tissues present "distinct" intensity profiles, are consistent among patients, and could be modelled through Gaussian mixtures [52] that would be equipped with spatial information in the form of neighbourhood constraints [53–55] or population-specific probabilistic atlases [56] to increase their robustness against intrinsic and extrinsic imaging factors. Cross-sectional brain volumetry can be conducted using validated segmentation tools, such as FAST [57], FIRST [58], SPM [56], and FreeSurfer [59], or whole-brain atrophy quantification algorithms, such as SIENAX [60]. Although they are still being used in clinical research due to their robustness and adaptability [61], preprocessing mistakes (e.g. poor skull stripping) [62,63], the presence of brain lesions (e.g. white matter hyperintensities or tumours) [64–66], the lack of contrast between tissues [67], intensity inhomogeneity [63], imaging differences (acquisition protocol, scanner vendor and version) [68], and the large differences between training and testing sets [69] compromised their performance.

Longitudinal studies scrutinise changes between two – or more – scans, possibly acquired in different sessions [11]. Cerebral atrophy may be described through surrogate measurements given by brain parenchymal fraction [70] or brain tissue boundary

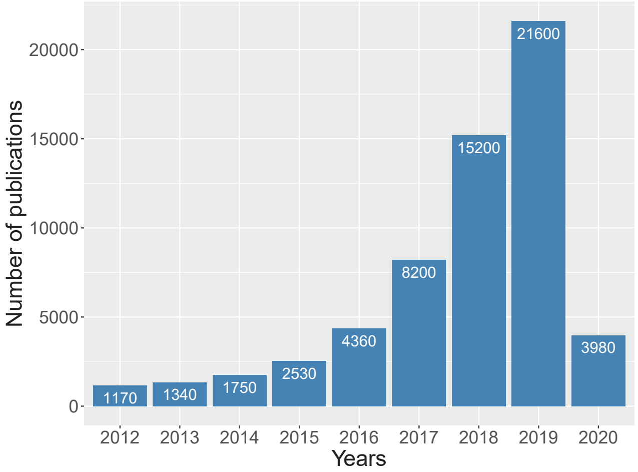


Figure 1.7: Number of peer reviewed publications per year according to Google Scholar using deep learning in medical imaging. Search keywords: “deep learning” and “medical imaging”. Search interval: between 2012 and 2020. (Queried: May 7th, 2020).

displacement [60, 71–73]. Longitudinal atrophy quantification can be conducted using tools, such as any cross-sectional tissue segmentation strategy, SIENA [57], and the Jacobian method [71]. In all cases, data harmonisation errors or presence of brain lesions may compromise the subsequent evaluation [62, 66, 74].

1.3 Motivation

Machine learning has become part of our daily basis: from intelligent systems recommending products and services [75] to complex natural language processors installed in smartphones capable of understanding questions and answering them accordingly [76]. In the medical domain, these intelligent systems permit supporting and easing medical decision making in sensible, intricate, and time-consuming tasks, primarily diagnostics [77], which not so long ago were unaddressable, such as automatic breast cancer screening [78], skin lesion classification [79], cardiac structure segmentation and diagnosis [80], segmentation and identification of retinal landmark and pathologies [81], histopathology image analysis [82], brain segmentation [83].

A branch of machine learning, referred to as deep learning, has become a hot topic in the last couple of years due to its astonishing performance in a myriad of computer vision applications [84–88]. Although early applications of deep learning in medical image analysis date back to the 1990s [89–91], the lack of sufficient and correctly labelled data, computational power limitations, and reduced interpretability discouraged researchers to continue developing deep learning techniques. With the arrival of graphic processing units [92], improvements in imaging, and efforts for collecting and processing vast amounts of data, this research area has rekindled and expanded considerably during the last couple of years, as depicted in Fig. 1.7.

Convolutional neural networks (CNN), an outstanding branch of deep learning applications to visual purposes, have become the state of the art in research medical imaging and have performed similarly to especially trained personnel in sensitive medical applications [79]. CNNs could be used to avoid defining ad-hoc spatial and intensity features explicitly as they learn mapping functions (input→output) based on the training data. In a nutshell, the process consists of optimising a set of convolutional kernels that permit extracting relevant hierachical features out of the input data and a set of units to mine the resulting characteristics to understand the content of the input and produce a response that matches the expected output accordingly. We hypothesise deep learning can improve brain tissue segmentation and atrophy quantification in cross-sectional and longitudinal studies, respectively. Both tasks are illustrated in Fig. 1.8.

1.4 Objectives

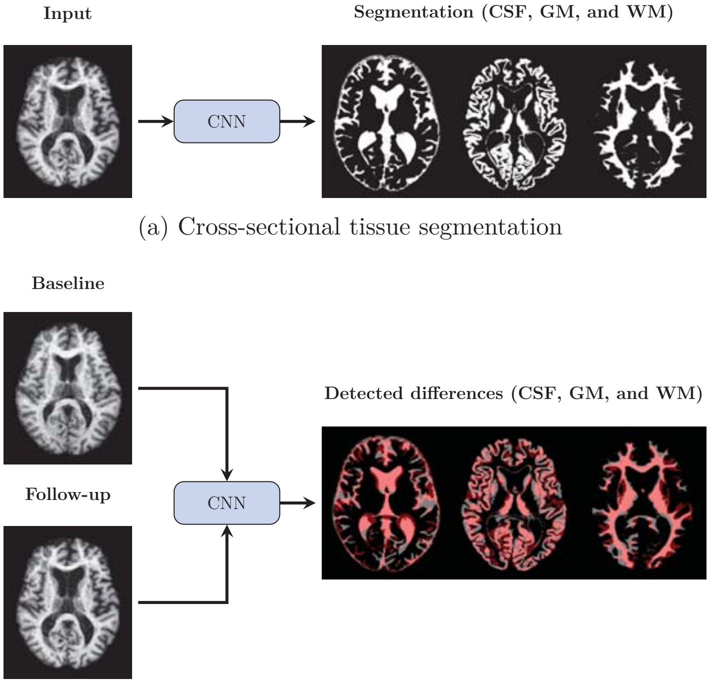
The main objective of this thesis is to:

develop deep learning methods for segmenting brain tissues and quantifying their temporal variations from magnetic resonance images.

We consider the following specific objectives to reach the aforementioned goal:

1. **Review the state of the art in deep learning for brain image analysis**. We review relevant literature on brain image analysis using deep learning to understand the needs from the medical point of view, study current trends and applications; analyse pipelines, strengths, weaknesses, and limitations; and general challenges that need to be addressed in the field.

2. **Compare quantitatively approximations relevant to cross-sectional brain MRI tissue segmentation in healthy and unhealthy subjects**. Based on our literature review, we implement and compare applicable brain MRI tissue segmentation algorithms quantitatively under the same evaluation framework to understand their practical strengths and weaknesses.


Input
Segmentation (CSF, GM, and WM)
CNN
Baseline
Follow-up
Detected differences (CSF, GM, and WM)
CNN

(a) Cross-sectional tissue segmentation

(b) Longitudinal atrophy quantification

Figure 1.8: Cross-sectional tissue segmentation and longitudinal atrophy quantification. The goal of tissue segmentation is to classify all voxels within the intracranial volume into cerebrospinal fluid, grey matter, and white matter for measuring brain volume afterwards. The goal of longitudinal atrophy quantification is to detect and quantify brain differences over time given a baseline and a follow-up scan. CNN: convolutional neural network. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter

3. **Propose a convolutional neural network for generating controlled atrophy change evaluation datasets**. Although there are a plethora of publicly available longitudinal brain MRI datasets, the lack of a ground truth prevents deep learning approaches from being used for quantifying temporal tissue changes. In light of that limitation, we propose a framework for generating longitudinal atrophy datasets, allowing evaluating the accuracy of atrophy quantification methods and training deep learning methods for performing such a task.

4. **Propose deep learning framework for quantifying longitudinal brain atrophy in healthy and unhealthy subjects**. To show one of the possible applications of our data augmentation network, we propose the first registration-based deep learning method for quantifying brain tissue changes

over time and assess its suitability versus validated tools in three cohorts of patients ongoing Alzheimer's disease, dementia, and multiple sclerosis.

1.5 Document structure

The rest of the thesis is structured as follows. Notions of tissue segmentation, atrophy quantification, and deep learning are presented in Chapter 2. The outcomes of our literature review on deep learning for medical imaging are described in Chapter 3. Details about the methodology and corresponding results and observations of our quantitative comparison of relevant methods for tissue segmentation in brain MRI are condensed in Chapter 4. Based on our bechmark, we proposed the processing approaches discussed in Chapter 4 for cross-sectional tissue segmentation that we submitted to various Grand Challenges of the International Conference on Medical Image Computer and Computer-Assisted Interventions. Our proposals for longitudinal atrophy generation and atrophy quantification are presented and evaluated in Chapter 5. We use our data generation framework as data augmentation strategy for training our deep learning based atrophy quantification proposal in Chapter 6. Final remarks of the overall work and future directions are outlined in Chapter 7.

Chapter 2

Theoretical background

This chapter contains the essentials of brain volumetry and deep learning. We published part of the theoretical background in the following paper:

Paper published in the **Artificial Intelligence in Medicine**
Volume: 95, Pages: 64-81, Published: April 2019
DOI: 10.1016/j.artmed.2018.08.008
JCR IF: 3.574 Q1(5/26)

2.1 Brain volumetry

2.1.1 Cross-sectional brain tissue segmentation

Brain tissue segmentation consists of classifying each voxel in an MRI acquisition into one of three regions: cerebrospinal fluid, grey matter, and white matter. Before the deep learning era, widespread tissue segmentation approaches were intensity based and assumed that intensities were a result of a mixture of these three regions of interest [93], as illustrated in Fig. 2.1, and, thus, the probability of a voxel X_i to have a certain intensity value x could be determined as follows

$$P(X_i = x) = \sum_{k=1}^{K} P(Z_i = k)P(X_i = x|Z_i = k), \tag{2.1}$$

where $K = 3$ the number of regions of interest and Z_i a latent variable indicating what region of interest X_i came from. In particular, intensity-based methods assumed intensities could be modelled through a mixture of Gaussian distributions [52], i.e.

$$P(X_i = x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x; \mu_k, \sigma_k), \tag{2.2}$$

where $P(Z_i = k) = \pi_k$ represents the weight of each component at the voxel X_i and μ_k and σ_k the mean and standard deviation of the k-th component. The segmentation process consisted of estimating the model latent variables Z_i using the expectation maximisation algorithm [94]. The segmentation performance of these types of methods varies depending on the image quality: images with low contrast, noise, and intensity inhomogeneity would produce unsatisfactory results as tissue distributions would overlap more.

Intensity-based tissue segmentation methods were later equipped with spatial information [94–97] to cope with the aforementioned issues. Four main strategies were distinguished in the literature: (i) impose local contextual constraints using Markov Random Fields [57], (ii) include penalty terms accounting for neighbourhood similarity in clustering objective functions [54], (iii) use Gibbs prior to model spatial characteristics of the brain [55], and (iv) introduce spatial information using probabilistic atlases [56]. Of note, some of these methods, like FAST [57] and SPM [56], are still being used in medical centres due to their robustness and adaptability [61].

Cross-sectional brain volume and atrophy measurements could be easily obtained after tissue segmentation by normalising the total volume of brain tissues or cerebrospinal fluid regions by the intracranial volume [60, 70, 98].

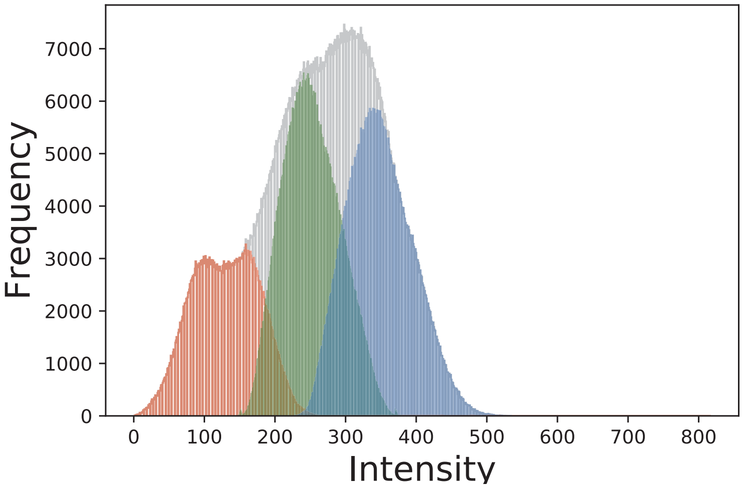


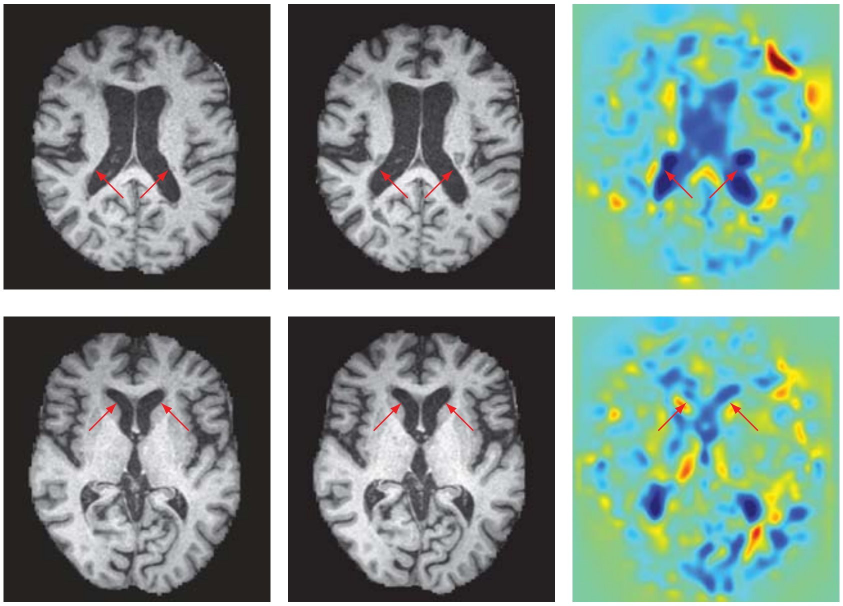
Figure 2.1: Histogram of intensities as a mixture of components. The multimodal distribution in grey represents the histogram of all tissues. The red, green, and blue distributions correspond to those of cerebrospinal fluid, grey matter, and white matter, respectively.

2.1.2 Longitudinal cerebral atrophy quantification

In longitudinal studies of cerebral atrophy, the assessments consists of examining brain tissue variations over time based on two input MRI scans, a baseline and a follow-up acquisition.

The examination of brain atrophy can be carried out taking advantage of the cross-sectional approach by segmenting brain tissues on both scans, computing the corresponding percentages of cerebrospinal fluid in the intracranial volume, and calculating the relative volume change [70, 98]. Although this approach does not necessarily involve registering input scans and is relatively simple to compute, its application might be limited by the accuracy of skull stripping and tissue segmentation steps [99]. Moreover, the method dispenses with spatial information and, hence, it does not provide any information about potential atrophied regions.

Alternatively, longitudinal cerebral atrophy can be studied by analysing brain tissue boundary displacement over time [60, 71–73]. A popular segmentation-based approach named SIENA (Structural Image Evaluation, using Normalization, of Atrophy) tool in the FSL package evaluates brain tissue displacement by co-registering baseline and follow-up scans, segmenting tissues in both, and calculating brain edge displacement following the normal of the brain boundary [57]. Such an approach provides local atrophy information since local brain boundary displacements indicate local atrophy. However, the accuracy of this approximation is conditioned by



(a) Baseline (b) Follow-up (c) Jacobian

Figure 2.2: Jacobian determinant as surrogate measure of atrophy. From left to right, axial view of baseline and follow-up scans and their corresponding Jacobian determinant. In the Jacobian determinant images, blue and red indicate contraction and expansion, in that order. The red arrows point to obvious changes between input scans, such as new lesions (top row) and ventricle widening (bottom row).

the accuracy of the registration, skull stripping, and tissue segmentation. Subtle inaccuracies in each one of these steps may result in significant errors in the final brain volume change values [74].

Recently, a registration-based method referred to as the Jacobian integration method has been proposed to cope with these issues [71]. The approach uses the information contained in deformation vectors obtained through non-linear registration as a surrogate measure of cerebral atrophy, as depicted in Fig. 2.2. The Jacobian determinant of the deformation vectors indicates the magnitude and direction of the change: the farther the magnitude from one, the higher the variation between baseline and follow-up; a positive direction indicates expansion while a negative direction contraction. Consequently, the integral of Jacobian determinants over the region of interest expresses the total brain change over time (longitudinal atrophy). Like in SIENA, the Jacobian method encodes the local displacement of the brain boundary and its also limited by the accuracy of the linear and non-linear registrations, skull stripping and the segmentation. However, unlike SIENA, segmentation is only done to the follow-up scan, reducing potential computation errors.

2.2 Deep learning

For years, conventional machine-learning techniques were built using automatic learning techniques and well-engineered algorithms as explained in the previous sections. The approach consisted in taking the raw data, describing its content with low-dimensional feature vectors – using specific prior knowledge of the addressed scenario – and inputting the vectors to a trainable classifier. While the classifier was certainly useful for other purposes, the ad-hoc features were not necessarily. Indeed, the overall accuracy of the method would depend on how appropriately designed were the heuristics [100].

Representation learning appears as an alternative to this drawback: discover automatically detection- and classification-suitable representations from the input data. One of the first successful attempts using this strategy took place in 1989 when LeCun *et al.* [101] presented a 5-layer fully-adaptive architecture for addressing handwritten digit recognition. Despite its accuracy results (1% error rate and 9% reject rate from a dataset of around 1000 samples), the authors were able to apply neural networks on a real world task. From thereon, several strategies considering much deeper and complex – yet trainable – networks have been successfully implemented not only on computer vision tasks, such as image segmentation and understanding; but also on natural language processing and sentiment analysis [102].

One of the most widely adopted approaches of deep neural networks is the convolutional neural networks (CNN) in which are able to process array-like data [102]. From a high-level perspective, the idea behind CNN is to identify compositional hierarchy features which objects from the real world exhibit: low level features (e.g. edges) form patterns and these specific patterns form high level ones (e.g. shapes, textures). Further information of the building blocks of CNN is provided in following sections. Fig. 2.3 depicts a CNN and its principal modules.

2.2.1 Convolution layer

CNNs are networks that share parameters across space. The convolution layer is one of its essential building blocks. From a general point of view, the convolutions are stack one after the other and, hence, a convolutional pyramid is created. This pyramid representation allows to take the spatial information given in the input layer and turn it into a semantic representation. Each convolutional layer contains a set of filters which are learnt during training. Each one of these kernels is slid over the entire input image to extract local dependencies and produce a feature map. This feature map, also referred as activation map, varies in complexity according to the depth at which it is calculated: shallow layers extract simple features whereas deep ones represent more complex and high-level structures [103]. The number of

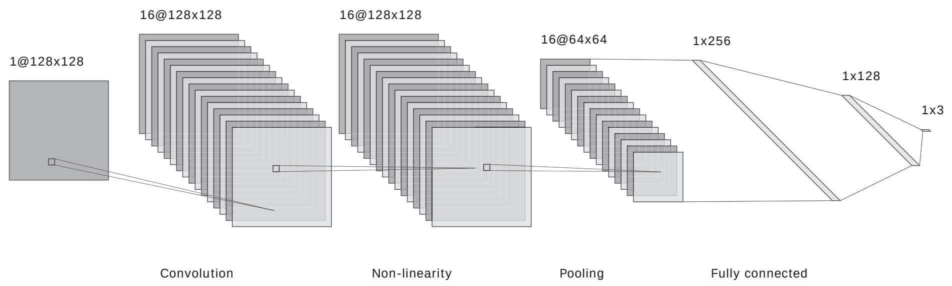


Figure 2.3: Building blocks of convolutional neural networks: convolutional, non-linear, pooling, and fully connected layers. The convolutional layers extract local dependencies and produce feature maps. Non-linearity layers ensure that the representation in the input space is mapped to a sparse one. Pooling layers summarise the information extracted by convolutional layers. Fully connected layers mine the feature maps extracted by the feature extractor part of the network (convolutional, non-linear, and pooling layers).

kernels in each layer (or depth), and their dimensions are design decisions.

2.2.2 Non-linearity layer

The above convolutional layer is usually followed by non-linearity operations. Non-linearity is achieved using a specific family of functions called activation functions. These activation functions ensure that the representation in the input space is mapped to a sparse one, hence achieving a certain invariance to data variability and a computationally efficient representation [104]. The former situation refers to the fact that sparse representations are more resilient to slight modifications than dense ones.

In the past, sigmoid and hyperbolic tangent functions were commonly used for this purpose. However, for large-scale image recognition, novel activation functions are being continuously proposed. We categorise them into three broad families.

Rectified linear units and variants

Rectified linear units (ReLU) can be expressed in general as follows

$$f\left(z_{lk}^{xy}\right) = \begin{cases} \max\left(z_{lk}^{xy},\, 0\right) & \text{if } z_{lk}^{xy} > 0, \\ \min\left(\alpha \cdot z_{lk}^{xy},\, 0\right) & \text{if } z_{lk}^{xy} \leq 0, \end{cases} \tag{2.3}$$

where z_{lk}^{xy} is the input value at position (x, y) on the k^{th} feature map at the l^{th} layer and α is the slope of the negative linear function. There are five special cases depending on α. First, if $\alpha = 0$, the expression results in the so-called ReLU [84, 105] which is one of the most commonly used activation functions [102]. Despite being computationally efficient, this activation scheme presents problems concerning gradient discontinuity [86, 106]. Second, if α is a small constant, the variant is referred as Leaky ReLU (LReLU) [107]. This approximation enables to cope with the problem of zero gradient. Third, if α is tuned up in the training process along with other parameters using back-propagation, the approach is referred to as Parametric ReLU (PReLU) [106]. Fourth, Xu *et al.* [108] sampled α from a uniform distribution for each input sample. Such an approach is called Randomised ReLU (RReLU). Fifth, Jin *et al.* [109] proposed an activation function, called S-shaped ReLU (SReLU), which considers a piecewise function composed of three linear functions, i.e.

$$f\left(z_{lk}^{xy}\right) = \begin{cases} t_r + a_r \cdot \left(z_{lk}^{xy} - t_r\right) & \text{if } z_{lk}^{xy} \geq t_r, \\ z_{lk}^{xy} & \text{if } t_r > z_{lk}^{xy} > t_l, \\ t_l + a_l \cdot \left(z_{lk}^{xy} - t_l\right) & \text{if } z_{lk}^{xy} \leq t_l, \end{cases} \tag{2.4}$$

where a_r, a_l, t_r and t_l are learnable parameters. Although there are insights on theoretical advantages and disadvantages of each one, their suitability is commonly assessed experimentally [108].

Maxout and variants

Maxout activation functions [110] were proposed to improve the optimisation and overall performance of dropout networks. The approach consists of computing the maximum across K feature maps, i.e.

$$f\left(z_{lk}^{xy}\right) = \max_{k \in [1,K]} \left(z_{lk}^{xy}\right). \tag{2.5}$$

A major drawback of this technique is that the number of trainable weights in each layer increases by a factor of K. A workaround to this situation was proposed by Springenberg *et al.* [111] in which a probabilistic sampling procedure was considered to compute the output feature maps. This activation function called probout empirically matched or improved the performance of maxout.

Exponential Linear Units (ELU) and variants:

Exponential Linear Units (ELU) [112] are similar to ReLU variants as they employ an identity for positive inputs. However, unlike them, ELUs provide saturated output for negative inputs. The saturation in the negative regions was found beneficial for expediting learning and improving the performance of very deep CNNs. They are defined as

$$f\left(z_{lk}^{xy}\right) = \max\left(z_{lk}^{xy},\, 0\right) + \alpha \cdot \min\left(\exp(z_{lk}^{xy}) - 1,\, 0\right). \tag{2.6}$$

Trottier *et al.* [113] defined parameters controlling different aspects of the ELU function and proposed learning them with gradient descent during training. This parametric ELU (PELU) further improved the speed and performance of training deep networks. Using off-the-shelf ResNet [85], PELU performed better than ELU and ReLU in image classification tasks on MNIST, CIFAR-10/100 and ImageNet datasets [113].

2.2.3 Pooling layer

Convolutional modules typically consists of three steps. First, the layer performs several convolutions to produce feature maps. Second, non-linear activation functions are used on the resulting maps. Third, the output is modified by the pooling layer before reaching the next convolutional layer. The idea of a pooling function is to summarise the information extracted in different non-overlapping neighbourhoods – usually – to reduce the number of parameters in the following layers, (ii) control over-fitting, and achieve slight translation invariance [114]. Among several pooling options, max pooling [115] is the most common approach due to its empirically demonstrated performance [116]. In the work of Springenberg *et al.* [117], convolutional layers with increase stride were used instead of max pooling operations without compromising the overall performance of the network.

2.2.4 Fully connected layer

Unlike the convolutional layer, the fully connected (FC) layer has full connection to all the units in the previous layer. Essentially, the main task of the FC layer is to mine the incoming features to extract information about the content of the input image. The process usually consists in flattening the feature maps coming from convolutional layers, to achieve a one dimension feature vector representation, and, then, inputting it into the FC layer. The output of this layer could either be network's response or used by another FC layer (consecutive FC layers can be stacked together).

Implementing FC layers usually require a large number of parameters – compared to other layers – since each neuron is fully connected to all elements in the previous layer. Additionally, networks using FC layers produce a single output per input and only accept fixed-size inputs. The former situation is computationally inconvenient if the CNN is used for segmentation and not classification. The latter issue implies that either input images are scaled to fit the requirements of the network or the network is re-factorised to be able to process the new data. FC layers can be converted to convolutional layers of 1×1 kernels [118] to solve this problem. In this way, the model keeps the fully connected functionality while accepting arbitrary input size image and making dense predictions.

2.2.5 Regularisation

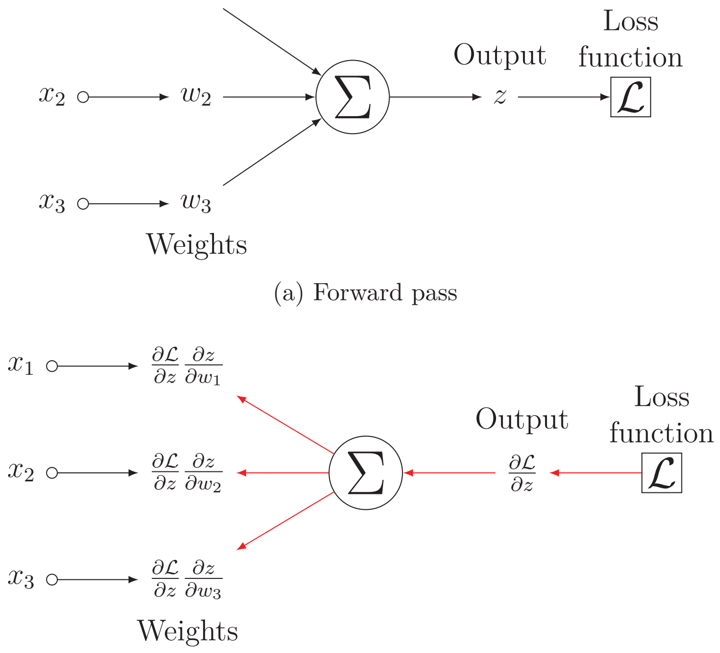
We observed the appealing performance of deep CNN methods in different domains, although they use enormous numbers of parameters. Unless trained on a large, labelled training datasets, proper regularisation should be employed to mitigate over-fitting. There are several regularisation methods widely used in the community, such as L_1 or L_2 regularisation approaches encouraging sparsity and small weight magnitude; early stopping [114] forcing the training to stop when there is a sign of over-fitting; batch normalisation [88] in which each batch is preprocessed to achieve mean equal to zero and standard deviation equal to one; and dropout [119] in which some feature map units are skipped. This last approach being the dominant since it is computational inexpensive and prevents co-adaptation among feature map units.

2.2.6 Output layer

CNNs are well-known for their ability to extract discriminative features using learned weights in each layer. The learning process is reinforced by employing appropriate loss functions. Loss functions are designed to encourage intra-class similarity and inter-class separability. In image classification tasks, most CNNs employ softmax loss, which is a combination of the softmax function and cross-entropy loss, mainly because of its simplicity and the probabilistic interpretation of softmax classifiers.

2.2.7 Optimisation

A CNN learns a mapping function between input training cases and corresponding expected responses. This function is learnt by adapting each of its trainable parameter values through backpropagation [120]. In principle, one could think of randomly perturbing one of the weights and examine whether such a variation led to improvement or not. However, it is practically inefficient since the number of parameters in CNNs is high. Alternatively, the backpropagation algorithm adjusts automatically the weight of these parameters by focusing on the error with respect to the desired output and not on the desired state of each network element: the


x_1
w_1
x_2
w_2
x_3
w_3
$\sum$
Output
z
Loss function
$\mathcal{L}$
Weights
x_1
$\frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_1}$
x_2
$\frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_2}$
x_3
$\frac{\partial \mathcal{L}}{\partial z}\frac{\partial z}{\partial w_3}$
$\sum$
Output
$\frac{\partial \mathcal{L}}{\partial z}$
Loss function
$\mathcal{L}$
Weights

(a) Forward pass

(b) Backward pass

Figure 2.4: Optimising a neural network using backpropagation. In (a), the network transforms the inputs x_i into the output z. The loss function $\mathcal{L}$ computes the error between the output and the expected result. In (b), the error is backpropagated to update weights accordingly. The weight adjustment consists of calculating the partial derivatives of the cost function with respect to each of the output and hidden nodes and, then, modifying the weights using the gradient descent algorithm [121].

higher the distance between the expected output and the obtained one, the more the weights have to be adjusted. The weight adjustment consists of calculating the partial derivatives of the cost function with respect to each of the output and hidden nodes and, then, modifying the weights using the gradient descent algorithm [121], as illustrated in Fig. 2.4.

The parameter update depends on the loss function, its value at a certain point, and the learning rate. First, the loss function can be evaluated for a single training case, a small subset, or the whole training set, referred in the literature as stochastic, mini-batch and batch gradient descent, respectively [122]. Updating values using the whole training set can be computationally expensive and, hence, mini-batch gradient descent is commonly considered in the community, leading to smoother parameter updating and more stable convergence than the stochastic approximation. Second, loss functions are typically non-convex and, thus, they may contain several local minima and saddle points in which gradient descent methods could get trapped easily [123–126]. The concepts of momentum and the Nesterov accelerated gradient descent [127, 128] were introduced to avoid oscillation around the local optima and fasten the optimisation process by accounting for previous gradients: if they have been on the same direction as the current one, speed up (increase) the update; and slow it down otherwise. Third, a major concern of momentum-based optimisation methods is to select an adequate learning rate, a parameter determining how substantial a change in the update should be made. This parameter is commonly set globally to be equal for all settings. Much work has been carried out on tuning the global learning rate adaptively based on the gradient of each parameter. For example, the learning rate of each parameter could be scaled according to the all past gradients [129], by a certain extent of past gradients [130, 131], or by jointly optimising momentum and learning rate [132].

2.3 Summary

In this chapter, we discussed the theoretical background of topics covered in this thesis: brain volumetry and deep learning. The goal of brain volumetry is to quantify brain tissue volume at cross-sectional levels and scrutinising diffuse and focal volume variations or boundary displacement at a longitudinal levels. Prior to the deep learning era, these methods required studying the problem at hand beforehand, selecting a set of representative features that help to discern between brain regions, and engineering a classifier that could use that information to produce accurate measurements. Nowadays, deep learning learns a suitable mapping function between input and output directly from the training data by iteratively adapting its trainable weights.

In the next chapter, we analyse different deep learning strategies that have been proposed in the literature for processing brain magnetic resonance images for various tasks in the medical domain, their processing pipelines, and potential advantages and disadvantages.

Chapter 3

Deep learning for brain image analysis

Deep learning has attracted the attention of researchers in the last few years due to their impressive performance on a plethora of computer vision tasks, medical image analysis is no exception. In this chapter, we discuss general deep learning strategies that have been considered in medical imaging; widespread preprocessing, processing, and postprocessing schemes; their targets and applications; and discuss key challenges to address in the future for easing their applicability for clinical practice. We published the outcome of this work in a journal paper and, recently, submitted an updated version as part of a book chapter. Details as follows:

Paper published in the **Artificial Intelligence in Medicine**
Volume: 95, Pages: 64-81, Published: April 2019
DOI: 10.1016/j.artmed.2018.08.008
JCR IF: 3.574 Q1(5/26)

Invited book chapter submitted to **World Scientific Publishing**
Book title: Deep learning for medical imaging
Editors: Bacciu D., Lisboa P.J.G., and Vellido A.
Expected publication year: 2021

Chapter 4

Benchmarking brain tissue segmentation methods

In this chapter, we compare patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging quantitatively to understand experimental strengths and weaknesses. Additionally, we propose cross-sectional methods for segmenting brain tissues in babies and adults. We published part of our work in the following paper:

Paper published in **IEEE Access** Volume: 7, Pages: 89986-90002, Published: July 2019
DOI: 10.1109/ACCESS.2019.2926697
JCR IF 4.098, Q1(23/155)

Moreover, based on the benchmark findings, we devised proposals to participate in three Grand Challenges of the Medical Image Computing and Computer-Assisted Intervention Conference

- Six-month old infant brain MRI tissue segmentation 2017
- MR brain segmentation 2018
- Six-month old infant brain MRI tissue segmentation 2019

4.1 Introduction

Several public brain MR datasets are available to the community, especially those organised by Medical Image Computing and Computer-Assisted Intervention (MICCAI) society[1], actively encouraging research and publications in the field. Each one of these evaluation frameworks has been proposed to quantitatively compare segmentation algorithms under the same directives: common training and testing data sets and evaluation metrics. Although they have indeed carried out their mission successfully, the algorithms are generally tweaked to perform the best. Hence, it is possible that the top-performing algorithm on a specific dataset does not achieve excellent scores on another one using the same pipeline (i.e. pre-processing, data preparation, and post-processing). Moreover, a direct comparison of architectures cannot be set up as each pipeline varies. Thus, hindering understanding the underlying properties of the different networks.

In this chapter, we compare quantitatively 4×2 fully convolutional neural networks (FCNN) architectures for tissue segmentation on brain MRI. We assess them more fairly by fixing training and test sets, processing pipeline (e.g. skull stripping, data normalisation, and reconstruction), training and optimisation schemes (e.g. epochs, early stopping policy, loss function, learning rate, optimiser, hardware), and performance evaluation metrics. The considered networks, comprising 2D and 3D implementations, are inspired in four recent works [103, 148, 152, 159]. The models are tested on three well-known datasets of infant and adult brain scans, with different spatial resolution, voxel spacing, and image modalities. In this work, we (i) compare different FCNN strategies for tissue segmentation; (ii) quantitatively analyse the effect of network dimensionality (2D or 3D) and the impact of fusing information from single or multiple modalities; (iii) study the influence of patch size on the segmentation performance; and (iv) investigate the effects of extracting patches with a certain degree of overlap as a sampling strategy in both training and testing. We made the repository available to the public as to provide a ready-to-use framework for exploring various state-of-the-art methods, valuable for newcomers to the topic. As all architectures are part of a standard pipeline, a direct comparison can be established, allowing us to understand the advantages and disadvantages of one architecture over another.

[1] http://www.miccai.org/

4.2 Methodology

4.2.1 Fully convolutional neural networks for brain MRI segmentation tasks

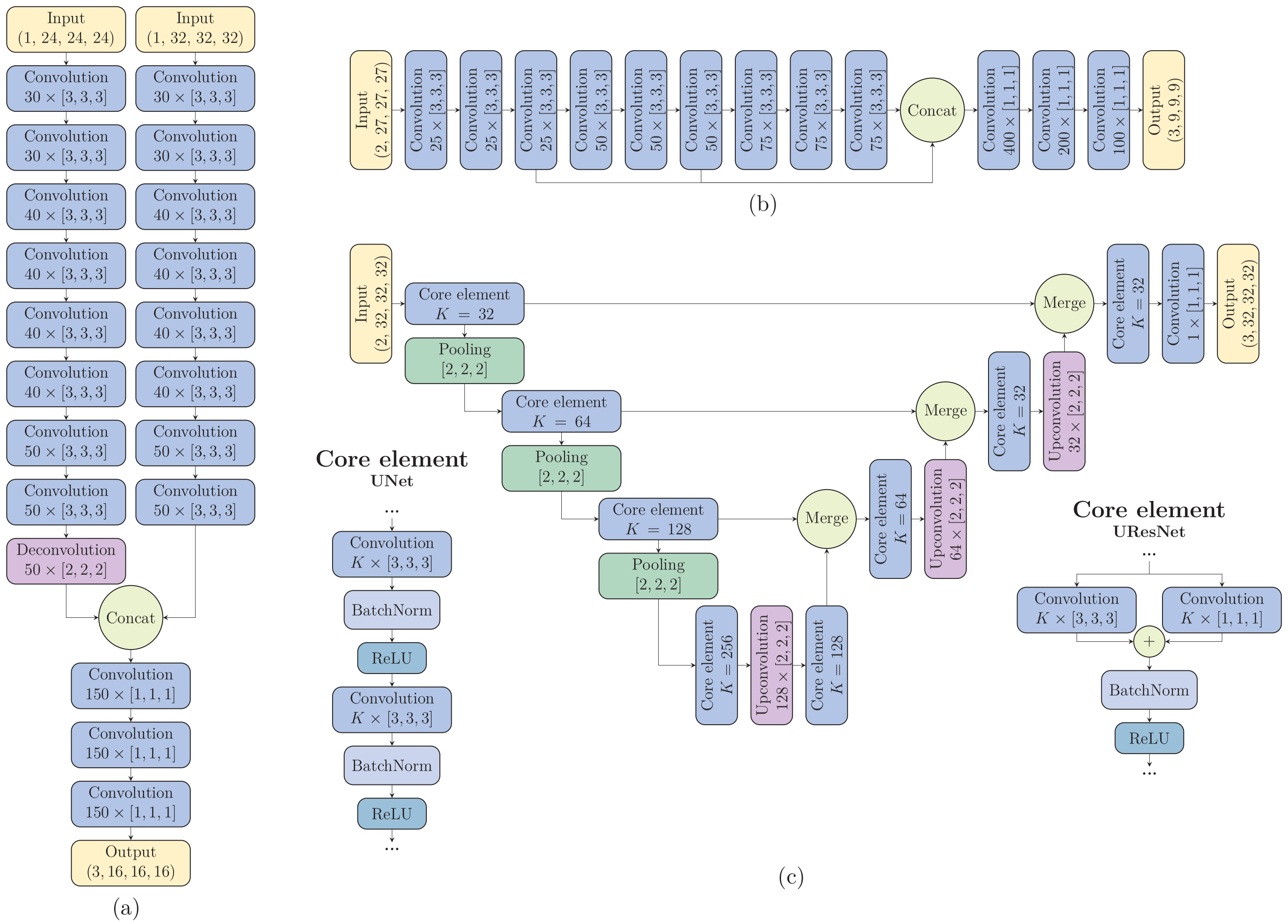
From the papers using fully convolutional neural networks indexed in Table 3.1, we built four multi-path architectures inspired by the works of Kamnitsas *et al.* [152], Dolz *et al.* [103], Çiçek *et al.* [148], and Guerrero *et al.* [159] (i.e. two convolution-only and two u-shaped architectures). The networks were implemented in 2D and 3D to investigate the effect of the network dimensionality on tissue segmentation. All these architectures were implemented from scratch following the architectural details given in the original work and are publicly available at our research website[2]. Although we made slight architectural changes, we retained the core idea of the original proposals. More details of the networks are given in the following sections.

Networks incorporating multi-resolution information

Kamnitsas *et al.* [152], proposed a two-path 3D FCNN for brain lesion segmentation. This approach achieved top performance on two public benchmarks, BRATS 2015 and ISLES 2015. By processing information of the targeted area from two different scales simultaneously, the network incorporated local and larger contextual information, providing a more accurate response [171]. A high-level scheme of the architecture is depicted in Fig. 4.1a. Initially, two independent feature extractor modules extracted maps from patches from normal and downscaled versions of an input volume. Each module consisted of eight $3 \times 3 \times 3$ convolutional layers using between 30 and 50 kernels. Afterwards, two intermediate $1 \times 1 \times 1$ convolutional layers with 150 kernels fused and mined resulting features maps. Finally, a classification layer (another $1 \times 1 \times 1$ convolutional layer) produced the segmentation prediction using a softmax activation.

Dolz *et al.* [103] presented a multi-resolution 3D FCNN architecture for subcortical structure segmentation. A general illustration of the architecture is shown in Fig. 4.1b. The network consisted of 13 convolutional layers: nine $3 \times 3 \times 3$, and four $1 \times 1 \times 1$. Each one of these layers was immediately followed by a Parametric Rectified Linear Unit (PReLU) layer, except for the output layer which activation was softmax. Multi-resolution information was integrated into this architecture by concatenating feature maps from shallower layers to the ones resulting from the last $3 \times 3 \times 3$ convolutional layer. As explained by Hariharan *et al.* [320], these kinds of connections grant networks to learn semantic – coming from deeper layers – as well as fine-grained localisation information – coming from superficial layers.

[2] http://github.com/NIC-VICOROB/tissue_segmentation_comparison



(a)

(b)

(c)

Figure 4.1: Architecture of benchmarked networks. Our implementations are inspired by the works of (a) Kamnitsas *et al.* [152], (b) Dolz *et al.* [103], (c) Çiçek *et al.* [148], and Guerrero *et al.* [159]. We show 3D versions only. The four-element tuples indicate number of channels K and patch size in x, y and z, in that order; triples in brackets indicate kernel size. Merging strategies corresponds to concatenation and addition for the UNet and UResNet, respectively.

U-shaped networks

In the u-shaped network construction scheme, feature maps from higher resolution layers are commonly merged to the ones on deconvolved maps to keep localisation information. Merging has been addressed in the literature through concatenation [148, 150] and addition [159, 166, 321]. In this chapter, we consider networks using both approaches. A general scheme of our implementations inspired in both works is displayed in Fig. 4.1c.

Çiçek *et al.* [148] proposed a 3D u-shaped FCNN, known as 3D u-net. The network is formed by four convolution-pooling layers and four deconvolution-convolution layers. The number of kernels ranged from 32 in its bottommost layers to 256 in its topmost ones. In this design, maps from higher resolutions were concatenated to upsampled maps. Each convolution was immediately followed by a Rectified Linear Unit (ReLU) activation function.

Guerrero *et al.* [159] designed a 2D u-shaped residual architecture for lesion segmentation, referred as u-ResNet. The building block of this network was the residual module which (i) added feature maps produced by 3×3- and 1×1-kernel convolution layers, (ii) normalised resulting features using batchnorm, and, finally, (iii) used a ReLU activation. The network consisted of three residual modules with 32, 64 and 128 kernels, each one followed by a 2×2 max pooling operation. Then, a single residual module with 256 kernels was applied. Afterwards, successive deconvolution-and-residual-module pairs were employed to enlarge the networks' output size. The number of filters went from 256 to 32 in the layer before the prediction one. Maps from higher resolutions were merged with deconvolved maps through addition.

From here on, our implementations of [103, 148, 152, 159] are denoted by DM, KK, UN and URN, respectively.

4.2.2 Aspects to evaluate

We analyse (i) overlapping patch extraction in training and testing, (ii) single and multi-modality architectures, (iii) patch size, and (iv) 2D and 3D strategies. Details on these four evaluation cornerstones are discussed in the following sections.

Overlapping sampling in training and testing

A drawback of networks performing dense-inference is that, under similar conditions, the number of parameters increases. This issue implies that more samples should be used during training to obtain acceptable results. A common approach consists of augmenting the input data through transformations – e.g. translation, rotation,

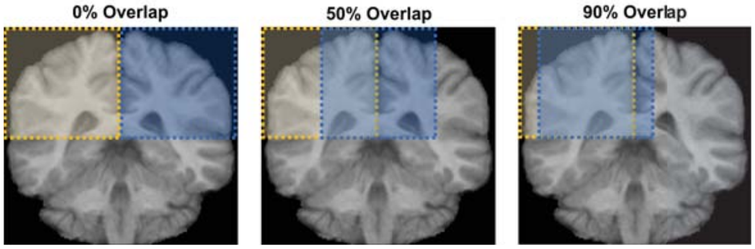


Figure 4.2: Patch extraction with null, medium and high overlap. Yellow and blue areas corresponds to the first and second blocks to consider. When there is overlap among patches, voxels are seen in different neighbourhoods each time.

scaling. However, if the output dimension is not equal to the input size, other options can be considered. Although the main advantage of patch-based FCNNs is their dense prediction, a single pass on a particular area may produce inaccurate outputs as (i) block boundary artefacts may appear – direct consequence of tiling volumes up – and (ii) patches may not contain sufficient information to produce an accurate verdict – e.g. on the boundaries of the input. For instance, patches can be extracted from the input volumes with a certain extent of overlap and, thus, the same voxel would be seen several times surrounded by different neighbourhoods. As each patch contains a specific part of the region of interest, each voxel would be classified according to the information it contains. An example of patch extraction with three extents of overlap is depicted in Fig. 4.2. In such a way, more information would be taken into account to produce a more consented and smoother response. Summarising, the strategy is beneficial as (i) more samples are gathered, and (ii) networks are provided with information that may improve spatial consistency as illustrated in Fig. 4.3. Of note, the overlap degree is determined by the overlap between adjacent output patches and not input ones.

The sampling strategy aforementioned can be enhanced by overlaying predictions, i.e. obtain a consented prediction per voxel from the segmentation of different overlapping patches. Unlike sophisticated post-processing techniques, the network itself is used to improve its segmentation. As depicted in Fig. 4.3 (e-h), the leading property of this post-processing technique is that small segmentation errors – e.g. holes and block boundary artefacts – are corrected. The consensus among outputs can be addressed through majority voting, for instance.

Input modalities

Depending on the number of modalities available in a dataset, approaches can be either single- or multi-modality. If many modalities were acquired, networks could be adapted to process them all at the same time either using different channels or

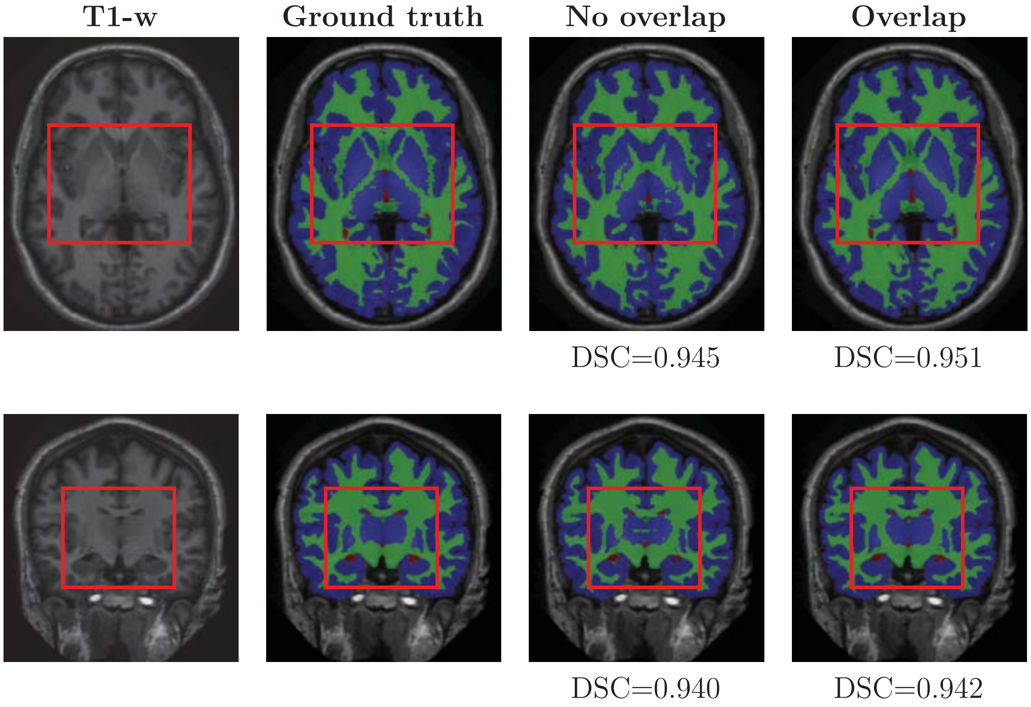


Figure 4.3: Segmentation using overlapping sampling in training (top row) and testing (bottom row). From left to right, T1-w volume, ground truth, segmentation without overlap, and with overlap. The basal ganglia area (inside the red box) depicts notable changes between strategies. Results obtained with overlapping sampling appear more similar to the ground truth. Colours for cerebrospinal fluid, grey matter, and white matter are red, blue, and green, respectively. DSC: Dice similarity coefficient.

various processing paths – also referred in the literature as early and late fusion schemes [167], respectively. Naturally, the former strategy is desirable regarding computational resources, but the latter may extract more valuable features. In this work, we consider the early fusion only. Regardless of the fusion scheme, merging different sources of information may provide models with complementary features and, hence, lead to enhanced outputs [67].

Patch size

A pivotal hyperparameter of CNNs is the input patch size. Experiments in this regard have shown that the larger the input patch, the more contextual information the network can mine to produce the final response. Nevertheless, the greater the

Table 4.1: Details of implemented architectures. The items into consideration appear on the first column. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. Note that there are two inputs for KK as the network has two processing branches.

	Item	**DM**$_{2D}$	**DM**$_{3D}$	**KK**$_{2D}$	**KK**$_{3D}$	**UN**$_{2D}$	**UN**$_{3D}$	**URN**$_{2D}$	**URN**$_{3D}$
General	Input size	27×27	$27 \times 27 \times 27$	$32 \times 3220 \times 20$	$32\times32\times32$ $20\times20\times20$	32×32	$32 \times 32 \times 32$	32×32	$32 \times 32 \times 32$
	Output size	9×9	$9 \times 9 \times 9$	16×16	$16\times16\times16$	32×32	$32 \times 32 \times 32$	32×32	$32 \times 32 \times 32$
	Number of parameters	547 278	3 333 270	569 678	7 101 038	1 931 620	5 606 308	995 108	2 623 844
No. components	Convolutional	13	13	19	19	18	18	18	18
	Batchnorm	0	0	0	0	18	18	9	9
	Max pooling	0	0	0	0	3	3	3	3
	Deconvolution	0	0	1	1	3	3	3	3
	Residual connections	0	0	0	0	0	0	12	12
	Concatenations	1	1	1	1	3	3	0	0

patch, the more resources needed to train the network successfully and the more parameters to be optimised during training. Thus, a trade-off between these factors is needed to obtain the best response.

Network dimensionality

There are two main streams of FCNN regarding its input dimensionality: 2D and 3D. On the one hand, 2D architectures are fast, flexible, and scalable; however, they ignore completely data from neighbouring slices, i.e. implicit information is reduced compared to 3D approaches. On the other hand, 3D networks acquire valuable implicit contextual information from orthogonal planes. Even though labelling is carried out slice-by-slice, these strategies tend to lead to better performance than 2D. Nevertheless, they are computationally demanding due to the exponential increase in parameters and resource consumption and may require larger training sets. Therefore, depending on the data itself, one approach would be more suitable than the other.

4.2.3 Implementation details

General pipeline

General tissue segmentation pipelines contemplate four essential components: pre-processing, data preparation, classification, and post-processing. Specific implementations of each one of these elements can be plugged and unplugged as required to achieve the best performance. First, pre-processing is carried out by (i) removing skull, and (ii) normalising intensities between scans. We use the ground truth masks to address the former tasks and standardise our data to have zero mean and unit variance. Second, data is prepared by extracting useful and overlapping patches – containing information from one of the three tissues. Third, each patch is classified. Fourth, no post-processing is considered.

Network training

The steps to train a model on a given dataset are as follows. First, for each dataset, the training set is split into training and validation at random (80% and 20% of the volumes, respectively). Both training and validation sets are fixed for all networks to ensure they were trained under similar conditions. Second, the networks are trained in batches of 32 elements for a maximum of 20 epochs. In this particular case, we observed experimentally that the loss function of all networks converged to their lowest values for both training and validation collections within 20 epochs and overfitted afterwards. Third, at the end of each epoch, the loss function value on the validation set is computed. The training stopping criterion is no improvement in validation accuracy after n epochs, which is monitored using an early stopping policy with patience n equal to 2. We adopted this strategy to guarantee that all deep networks were trained in the best way possible while avoiding over-fitting to the training set and increasing the chances of achieving the best performance on unknown collections. The models are optimised for the categorical cross-entropy loss function using the Adam [132] optimisation method with an initial learning rate of 1×10^{-3}, a decay of 0.0, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ (i.e. default parameter values, as suggested in the original paper). Of note, we considered this particular optimiser as it showed empirically improved performance in comparison to other stochastic optimisation methods and favourable performance in problems with noisy gradients and, also, we used its default hyperparameter values since the authors found that little tuning was needed to reach acceptable results in most of the cases. All voxels laying on the background region are given a weight of zero to avoid considering them in the optimisation process. This decision was taken as non-brain regions were removed during pre-processing.

Network testing

The steps to test a trained model on a given input MR volume are as follows. First, the whole volume is divided into patches. These patches are extracted from the entire input and not from specific regions. Second, the different patches are passed through the network to obtain a segmentation. Third, as there might be a degree of overlap between output probability maps, the final segmentation is provided through means of majority voting. The mode of the votes for each voxel is selected as consensed classification value. Convolutional-only networks classify only a subset of voxels. Commonly, networks dispense with outermost voxels and predict centermost ones only. For instance, the DM2D model receives a 27 × 27 patch and outputs classification values for voxels within a 9×9 rectangular region delimited by $(9, 9) - (9, 18) - (18, 18) - (18, 9)$. Thus, patches must be extracted with a step in between them of at most the output size of the network to be able to produce a valid whole brain segmentation, i.e. an incoming MR volume is tiled up so that the resulting output maps are adjacent to each other. Once patches are extracted from the scan, they are passed through the network and rearranged to reconstruct the segmentation volume.

Software and hardware

All the architectures were implemented from scratch in Python, using the Keras library. Relevant information per architecture is summarised in Table 4.1. All the experiments were run on a GNU/Linux machine box running Ubuntu 16.04, with 128GB RAM. CNN training and testing were carried out using a single TITAN-X PASCAL GPU (NVIDIA corp., United States) with 8GB RAM. The developed framework for this work is currently available to download at our research website. The source code includes architecture implementation and experimental evaluation scripts.

4.3 Experimental results

4.3.1 Considered datasets

We consider one publicly available repository and two challenges: Internet Brain Segmentation Repository 18 (IBSR18)[3], MICCAI Multi-Atlas Labeling challenge 2012 (MICCAI 2012)[4] and 6-month infant brain MRI segmentation (iSeg2017) [322][5], respectively. The datasets were chosen since they have been widely used in the literature to compare different methods and, also, they contain infants and adults data, with different voxel spacing and a different number of scans. We believe that these two factors allow us to see how robust, general, and useful in different scenarios can be the algorithms. The organisers of the MICCAI 2012 challenge split the data into training and testing (10 and 13 volumes, respectively). To be consistent with the challenge and allow comparison with other strategies, we followed the same evaluation procedure. To use annotations of MICCAI 2012, we mapped all the labels to form the three tissue classes. Specific details of these datasets are presented in

[3] http://www.nitrc.org/projects/ibsr

[4] http://masi.vuse.vanderbilt.edu/workshop2012

[5] http://iseg2017.web.unc.edu

Table 4.2: Relevant information from the considered datasets. In the table, the elements to be considered are presented in the first column and the corresponding information from IBSR18, MICCAI 2012 and iSeg2017 are detailed in the following ones. In the row related to the number of scans (with GT), the number of training and test volumes is separated by a + sign. For both IBSR18 and iSeg2017, the evaluation is carried out using leave-one-out cross-validation.

Item	**IBSR18**	**MICCAI 2012**	**iSeg2017**
Target	Adult	Adult	Infant
Number of scans	18	15 + 20	10
Bias-field corrected	Yes	Yes	Yes
Intensity corrected	No	Yes	No
Skull stripped	No	No	Yes
Voxel spacing	0.8×0.8×1.5 0.9×0.9×1.5 1.0×1.0×1.5	0.5 × 0.5 × 0.5	1.0 × 1.0 × 1.0
Modalities	T1-w	T1-w	T1-w, T2-w

Table 4.2.

4.3.2 Evaluation measurements

We used the Dice similarity coefficient (DSC) [323, 324] and the modified Hausdorff distance [325] to compare segmentation outputs against the ground truths. The DSC is used to determine the extent of overlap between a given segmentation and the ground truth. Given an input volume V, its corresponding ground truth $G = \{g_1, g_2, ..., g_n\}$, $n \in \mathbb{Z}$ and obtained segmentation output $S = \{s_1, s_2, ..., s_m\}$, $m \in \mathbb{Z}$ the DSC is mathematically expressed as

$$DSC\,(G, S) = 2\ \frac{|G \cap S|}{|G| + |S|}, \tag{4.1}$$

where $| \cdot |$ represents the cardinality of the set. The values for DSC lay within $[0, 1]$, where the interval extremes correspond to null or exact similarity between the compared surfaces, respectively.

The MHD evaluates the distance between the sets of points forming the segmented and ground truth surfaces. Using the same notation as in Eq. 4.1, the MDH is calculated as follows

$$MHD\,(G, S) = \max\left\{ {}^{95}K^{th}_{g_i \in G} d(g_i, S), {}^{95}K^{th}_{s_i \in S} d(s_i, G) \right\}, \tag{4.2}$$

where $d(a, \mathcal{B})$ corresponds to the minimum Euclidean distance between the point a and all the points in set $\mathcal{B}$ and ${}^{x}K^{th}_{b \in \mathcal{B}}$ represents the K-th ranked distance such that

$K/|\mathcal{B}| = x\%$ [325]. For example, $x = 50$ corresponds to the median of the distances. We use the 95-th percentile MHD calculation over the original HD ($x = 100$) as the former is more robust to outliers in the segmentation. The values for MHD are positive decimal numbers greater or equal to zero, where zero indicates that the two surfaces exactly coincide – neglecting eccentric observations.

We consider the Wilcoxon signed-rank test to assess and report the statistical differences among architectures.

4.3.3 Evaluation results

The evaluation we conducted is four-fold. First, we investigate the effect of overlapping patches in both training and testing stages. Second, we assess the improvement of multi-modality architectures over single-modality ones. Third, we study whether patch size has any influence on the performance. Fourth, we compare the different models on the three considered datasets. Note that, for the sake of simplicity, the network' dimensionality is shown as a subscript (e.g. URN$_{2D}$ denotes the 2D version of the URN architecture). The exact evaluation results are attached in the Appendix A.1.

Overlapping

To evaluate the effect of extracting overlapping patches in training and testing, we ran all the architectures on the three datasets contemplating three levels: null, medium and high (approximately 0%, 50% and 90%, respectively). On IBSR18 and iSeg2017, we carried out the evaluation using a leave-one-out cross-validation scheme. On MICCAI2012, we used the given training and testing sets.

The number of patches and average processing times for training, validating and testing each architecture in MICCAI2012, IBSR18, and iSeg2017 are condensed in Table 4.3. The average response time per voxel for DM$_{2D}$, DM$_{3D}$, KK$_{2D}$, KK$_{3D}$, UN$_{2D}$, UN$_{3D}$, URN$_{2D}$, and URN$_{3D}$ was $0.14\mu s$, $0.11\mu s$, $0.16\mu s$, $0.09\mu s$, $1.70\mu s$, $0.81\mu s$, $0.79\mu s$, and $0.48\mu s$, respectively. On the one hand, 3D architectures output more voxels at a time and, hence, their voxel-wise classification response time is lower than their 2D analogues. On the other hand, the latter set of networks provides a considerably faster whole volume segmentation compared to their counterpart, in accordance with the literature [122]. Additionally, the fact that the overlapping policy led to a vast amount of training patches could explain why the networks converged in a few epochs: the more the patches, the longer the epochs, but the more the information provided to the network in a single pass.

The first test consisted of quantifying improvement between networks trained

Table 4.3: Number of patches and average processing time for training, validating and testing each model in each dataset. The values for training and validation are of each one of the epochs and the ones for testing are of each volume. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality.

		Overlap	**Item**	**DM$_{2D}$**	**DM$_{3D}$**	**KK$_{2D}$**	**KK$_{3D}$**	**UN$_{2D}$**	**UN$_{3D}$**	**URN$_{2D}$**	**URN$_{3D}$**
MICCAI2012	Training	Null (0%)	Patches	2 666 496	291 648	835 584	52 224	196 608	6 144	196 608	6 144
			Time (s)	176	360	104	170	21	29	19	31
		Intermediate (50%)	Patches	8 930 304	1 779 084	3 440 640	430 080	835 584	52 224	835 584	52 224
			Time (s)	591	2 195	428	1 397	90	244	79	267
		High (90%)	Patches	56 229 888	28 114 944	56 229 888	28 114 944	13 959 168	3 489 792	13 959 168	3 489 792
			Time (s)	3 720	34 684	7 000	91 356	1 504	16 335	1 316	17 820
	Validation	Null (0%)	Patches	666 624	72 912	208 896	13 056	49 152	1 536	49 152	1 536
			Time (s)	44	90	26	42	5	7	5	8
		Intermediate (50%)	Patches	2 232 576	444 771	860 160	107 520	208 896	13 056	208 896	13 056
			Time (s)	148	549	107	349	23	61	20	67
		High (90%)	Patches	14 057 472	7 028 736	14 057 472	7 028 736	3 489 792	872 448	3 489 792	872 448
			Time (s)	930	8 671	1 750	22 839	376	4 084	329	4 455
	Testing	Null (0%)	Patches	222 208	24 304	69 632	4 352	16 384	512	16 384	512
			Time (s)	15	30	9	14	2	2	2	3
		Intermediate (50%)	Patches	744 192	148 257	286 720	35 840	69 632	4 352	69 632	4 352
			Time (s)	49	183	36	116	8	20	7	22
		High (90%)	Patches	4 685 824	2 342 912	4 685 824	2 342 912	1 163 264	290 816	1 163 264	290 816
			Time (s)	310	2 890	583	7 613	125	1 361	110	1 485
IBSR18	Training	Null (0%)	Patches	1 304 576	142 688	425 984	26 624	106 496	3 328	106 496	3 328
			Time (s)	86	176	53	87	11	16	10	17
		Intermediate (50%)	Patches	4 243 200	845 325	1 703 936	212 992	425 984	26 624	425 984	26 624
			Time (s)	281	1 043	212	692	46	125	40	136
		High (90%)	Patches	27 262 976	13 631 488	27 262 976	13 631 488	6 815 744	1 703 936	6 815 744	1 703 936
			Time (s)	1 804	16 817	3 394	44 294	734	7 976	642	8 701
	Validation	Null (0%)	Patches	401 408	43 904	131 072	8 192	32 768	1 024	32 768	1 024
			Time (s)	27	54	16	27	4	5	3	5
		Intermediate (50%)	Patches	1 305 600	260 100	524 288	65 536	131 072	8 192	131 072	8 192
			Time (s)	86	321	65	213	14	38	12	42
		High (90%)	Patches	8 388 608	4 194 304	8 388 608	4 194 304	2 097 152	524 288	2 097 152	524 288
			Time (s)	555	5 174	1 044	13 629	226	2 454	198	2 677
	Testing	Null (0%)	Patches	100 352	10 976	32 768	2 048	8 192	256	8 192	256
			Time (s)	7	14	4	7	1	1	1	1
		Intermediate (50%)	Patches	326 400	65 025	131 072	16 384	32 768	2 048	32 768	2 048
			Time (s)	22	80	16	53	4	10	3	10
		High (90%)	Patches	2 097 152	1 048 576	2 097 152	1 048 576	524 288	131 072	524 288	131 072
			Time (s)	139	1 294	261	3 407	56	614	49	669
iSeg2017	Training	Null (0%)	Patches	602 112	65 856	193 536	12 096	43 008	1 344	43 008	1 344
			Time (s)	40	81	24	39	5	6	4	7
		Intermediate (50%)	Patches	1 906 688	379 848	774 144	96 768	193 536	12 096	193 536	12 096
			Time (s)	126	469	96	314	21	57	18	62
		High (90%)	Patches	12 386 304	6 193 152	12 386 304	6 193 152	3 096 576	774 144	3 096 576	774 144
			Time (s)	820	7 640	1 542	20 124	334	3 624	292	3 953
	Validation	Null (0%)	Patches	172 032	18 816	55 296	3 456	12 288	384	12 288	384
			Time (s)	11	23	7	11	1	2	1	2
		Intermediate (50%)	Patches	544 768	108 528	221 184	27 648	55 296	3 456	55 296	3 456
			Time (s)	36	134	28	90	6	16	5	18
		High (90%)	Patches	3 538 944	1 769 472	3 538 944	1 769 472	884 736	221 184	884 736	221 184
			Time (s)	234	2 183	441	5 750	95	1 035	83	1 129
	Testing	Null (0%)	Patches	75 264	8 232	24 192	1 512	5 376	168	5 376	168
			Time (s)	5	10	3	5	1	1	1	1
		Intermediate (50%)	Patches	238 336	47 481	96 768	12 096	24 192	1 512	24 192	1 512
			Time (s)	16	59	12	39	3	7	2	8
		High (90%)	Patches	1 548 288	774 144	1 548 288	774 144	387 072	96 768	387 072	96 768
			Time (s)	102	955	193	2 515	42	453	36	494

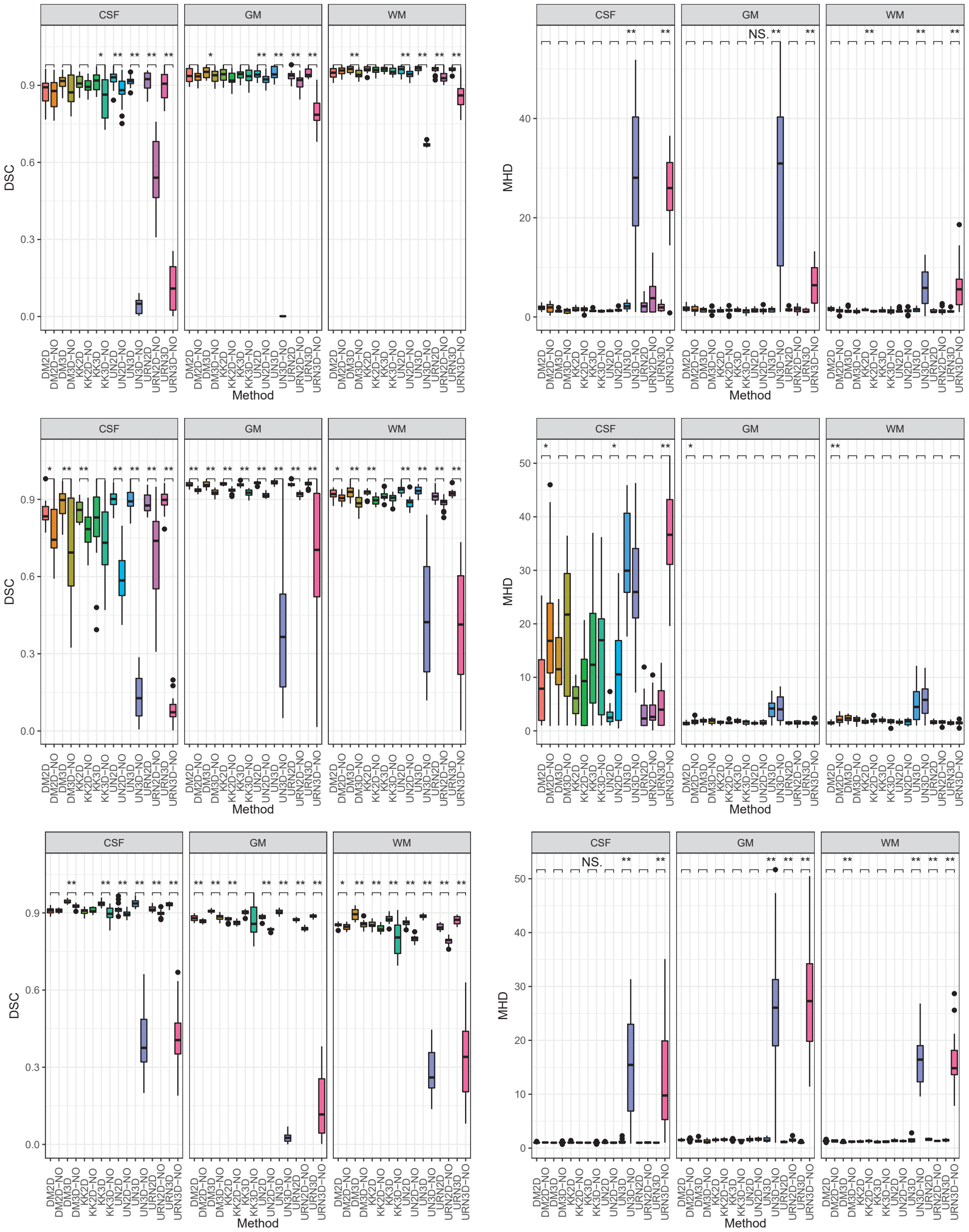


Figure 4.4: DSC (left column) and MHD (right column) values obtained using the null and high overlapping sampling in training. The suffix "-NO" on the name of the method means that the architecture was not trained using the sampling strategy. From top to bottom, boxplots for MICCAI2012, IBSR18, and iSeg2017, respectively. Differences between both versions of the same baseline architecture are highlighted with NS, *, and ** indicating a *p*-value > 0.1, < 0.05 and < 0.01, respectively. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter.

with either null or high degrees of overlap on training. The distribution of segmentation scores obtained on the three datasets is depicted in Fig. 4.4. In general, the models trained with patches extracted with a high extent of overlap yielded higher DSC and lower MHD values compared to when they were not. On the one hand, the sampling technique led to significantly higher DSC scores (p-value < 0.05) in 58 out of the 72 comparisons. On the other hand, overall, the precision of the method (measured in terms of inter-quartile range) regarding MHD increases but improvements were not significant in most of the cases (p-value > 0.05 in 51 out of 71 comparisons). Of note, there are enhancements in the boundaries but without taking into account most eccentric observations, MHD values are fairly similar. These three observations imply that the methods improve their segmentation, are more precise, but, in general, the borders of the segmentation masks do not change dramatically. In IBSR18, most of the models exhibited low DSC and notably high MHD scores when segmenting CSF. This outcome might be a consequence of the reduced number of samples available for this class (only the ventricular region). For iSeg2017, although the models trained with the overlapping sampling strategy yielded high DSC scores for the three classes, the MHD values show that the models had problems with delineating the limits between GM and WM accurately. The two groups of architectures exhibited opposite behaviours. U-shaped networks exhibited topmost improvements. This outcome is related to the fact that non-overlap may mean not enough samples. Instead, convolutional-only models evidenced the least increase. Since output patches are smaller, additional data can be extracted and used during training. Therefore, they can provide already accurate results. This fact is illustrated by the results of DM_{2D} and KK_{2D}.

The second test contemplated quantifying the improvement of extracting patches using combinations of the three considered degrees of overlap during training and testing. As mentioned previously, results were fused using a majority voting technique. We noted that the general trend was that the difference between results using null and high extends of overlap on testing time was not significant (p-values > 0.05). Also, the interquartile range remained similar regardless of the method or dataset. Nevertheless, the general trend was an improvement of mean DSC of at least 1% in the overlapping cases. Another important observation from our experiments is that zero impact or slight degradation of the DSC and MHD values was noted when training with null overlap and testing with high overlap. Naturally, this situation is a consequence of merging predictions of a poorly trained classifier.

Medium level of overlap patch extraction, in both training and testing, led to improvement with respect to null degree cases but yielded lower values than when using a considerable extent of overlap. The general trend is: the more the extent of overlap, the higher the overall performance of the method. The price to pay for using further levels of overlap is computational time and power since the number of samples to process increases exponentially. For example, given an input volume with dimensions 256 × 256 × 256 and a network producing output size of 32 × 32 × 32, the number of possible patches to be extracted following the null, medium and high overlap policies are 512, 3 375 and 185 193, respectively.

As overlapping sampling proved useful, the results showed in following sections correspond to the ones obtained using a high overlap in both training and testing.

Single and multiple modalities

We performed leave-one-out cross-validation on the iSeg2017 dataset using the implemented 2D and 3D architectures to assess the effect of single and multiple imaging sequences on the final segmentation. The results of this experiment are shown in Fig. 4.5. Overall, the more the input modalities, the better the segmentation. In this case, two modalities not only allowed the network to achieve higher mean but also to reduce the IQR, i.e. networks are more accurate and precise. This behaviour was evidenced regardless of architectural design or tissue type. For instance, while the best single modality strategy scored 0.937 ± 0.011, 0.891 ± 0.010 and 0.868 ± 0.016 for CSF, GM and WM, respectively; its multi-modality analogue yielded 0.944 ± 0.008, 0.906 ± 0.008 and 0.887 ± 0.017 for the same classes. Furthermore, in most of the cases, the strategies using both T1-w and T2-w obtained significantly higher DSC and lower MHD values compared to their single-modality counterparts. These results imply that multi-modality architectures obtained enhanced segmentation maps similar to the ground truth compared to the single-modality analogues as a direct consequence of providing the network with additional tissue contrast information (e.g. DSC increased and MHD decreases for CSF due to the contrast between this class and the other two in T2-w).

Effect of patch size

The effect of patch size in the overall performance has been investigated previously [326–329] and the overall trend has been that the larger the patch size, the more the contextual information provided to the network and, thus, the more enhanced the segmentation per se. Nonetheless, this particular experiment has not been carried out on 2D and 3D networks for tissue segmentation to the knowledge of the authors. We modified the baseline architectures by changing the input – and, consequently, output – patch size to study this matter. The size of the patches was selected in light of computational requirements (namely, the larger the patch, the more resources needed) and conditions imposed by the architectures (e.g. u-shaped networks may require input patch dimensions to be multiple of two due to max pooling modules). Information regarding the resulting designs is condensed in Table 4.4.

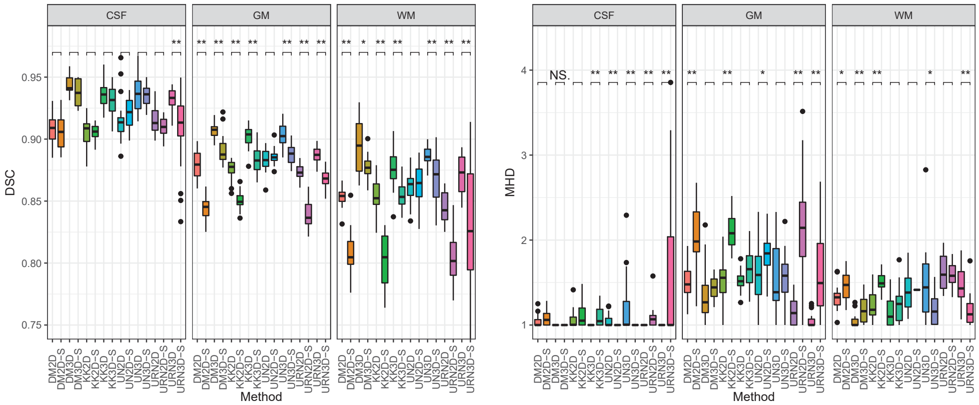


Figure 4.5: DSC (left) and MHD (right) values obtained using single and multiple input modalities. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. The suffix "-S" on the name of the method means that the architecture was single modality. Differences between both versions of the same baseline architecture are highlighted with NS, *, and ** indicating a p-value > 0.1, < 0.05 and < 0.01, respectively. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter.

Table 4.4: Implemented architectures to test patch size influence. The items into consideration appear on the first column. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. Note that there are two inputs for KK as the network has two processing branches.

	Item	**DM$_{2D}$**	**DM$_{3D}$**	**KK$_{2D}$**	**KK$_{3D}$**	**UN$_{2D}$**	**UN$_{3D}$**	**URN$_{2D}$**	**URN$_{3D}$**
Small	Input size	23×23	$23 \times 23 \times 23$	28×28 22×22	$28 \times 28 \times 28$ $22 \times 22 \times 22$	8×8	$8 \times 8 \times 8$	8×8	$8 \times 8 \times 8$
	Output size	5×5	$5 \times 5 \times 5$	14×14	$14 \times 14 \times 14$	8×8	$8 \times 8 \times 8$	8×8	$8 \times 8 \times 8$
	Number of parameters	458 254	1 835 654	466 830	4 345 966	1 931 620	5 606 308	995 108	2 623 844
Medium	Input size	27×27	$27 \times 27 \times 27$	32×32 20×20	$32 \times 32 \times 32$ $20 \times 20 \times 20$	16×16	$16 \times 16 \times 16$	16×16	$16 \times 16 \times 16$
	Output size	9×9	$9 \times 9 \times 9$	16×16	$16 \times 16 \times 16$	16×16	$16 \times 16 \times 16$	16×16	$16 \times 16 \times 16$
	Number of parameters	547 278	3 333 270	569 678	7 101 038	1 931 620	5 606 308	995 108	2 623 844
Large	Input size	37×37	$37 \times 37 \times 37$	36×36 26×26	$36 \times 36 \times 36$ $26 \times 26 \times 26$	32×32	$32 \times 32 \times 32$	32×32	$32 \times 32 \times 32$
	Output size	19×19	$19 \times 19 \times 19$	20×20	$20 \times 20 \times 20$	32×32	$32 \times 32 \times 32$	32×32	$32 \times 32 \times 32$
	Number of parameters	938 398	13 773 790	695 054	11 116 014	1 931 620	5 606 308	995 108	2 623 844

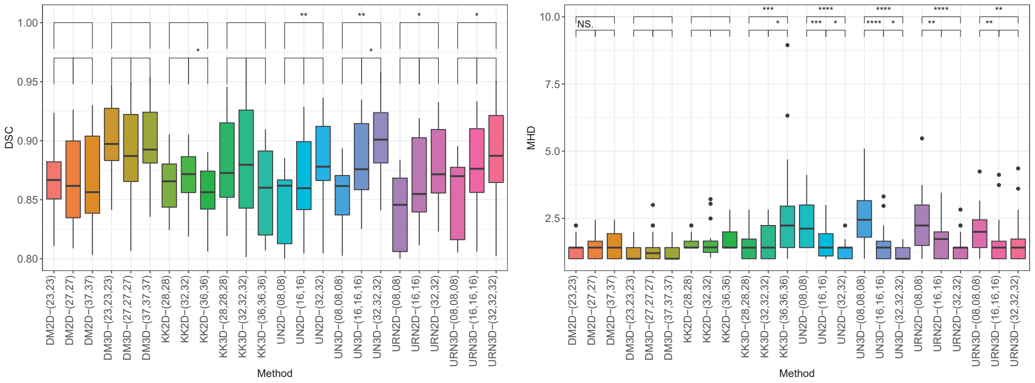


Figure 4.6: DSC (left) and MHD (right) values obtained by three variations of the baseline architectures concerning input and output patch size on the iSeg2017 dataset. Displayed results correspond to the average of scores obtained per class. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The suffix indicates the input block dimensions. Differences between variations of the same baseline architecture are highlighted with ****, ***, **, *, and NS indicating a p-value < 0.0001, < 0.001, < 0.01, < 0.05, and > 0.10, respectively.

We performed a leave-one-out cross-validation on the iSeg2017 dataset using the various architectures to study the effect of patch size. The averaged DSC and MHD results of this trial are displayed in Fig. 4.6. On the one hand, the large u-shape architectures performed better than their medium-size counterpart (improved DSC and MHD mean and, in some cases, standard deviation as well) and significantly better than their small analogues (p-value < 0.05). On the other hand, convolutional-only networks did not exhibit the same pattern. In some cases, the small DM and KK architectures outperformed their medium and large versions, but improvements were not statistically significant. In some other cases, the medium variants led to the best segmentation outcomes. Overall, the large convolutional-only architectures led to inferior performance. This situation might have to do with the fact that the number of trainable parameters increases substantially between network adaptations. For instance, there is an increase in the number of parameters of approximately 256% between the smallest and the largest implementations of KK_{3D}. Consequently, these sizeable networks require more training samples to surpass their tinier versions.

We opted for using the largest u-shaped designs (i.e. patch dimensions equal to 32) and the intermediate convolutional-only networks (i.e. patch dimensions equal to 27 for DM and 32-20 for KK).

Comparison between 2D and 3D FCNN architectures

The eight architectures were evaluated using their best parameters according to the previous sections on the three different datasets: MICCAI2012, IBSR18, and iSeg2017. The distribution of segmentation scores for DSC and MHD is shown in Fig 4.7. The observations for each dataset vary. In MICCAI2012, the difference between 2D and 3D methods can be mostly perceived in the distance between data points, forming the CSF segmentation masks. In IBSR18, 3D algorithms yielded similar or lower performance than their 2D analogues. Taking into account the information in Table 4.2, 3D architectures might be slightly more affected by heterogeneity in voxel spacing. One of the reasons explaining this outcome is the lack of sufficient data which prevents 3D networks from understanding spacing and resolution variations, i.e. 3D networks might lack enough information to generalise properly. In iSeg2017, the 2D architectures displayed lower performance than their 3D counterparts, mostly concerning DSC. The networks performing the best on MICCAI2012, IBSR18, and iSeg2017 were UN_{3D}, URN_{2D} and UN_{2D}, and DM_{3D}, respectively.

Segmentation outputs obtained by the different methods on one of the volumes of the IBSR18 dataset are displayed in Fig. 4.8. Note that architectures using 2D information were trained with axial slices. Since 2D networks process each slice independently, the final segmentation is not necessarily accurate nor consistent: (i) subcortical structures exhibit unexpected shapes and holes, and (ii) sulci and gyri are not segmented finely. Thus, even if segmentation was carried out slice-by-slice, 3D approaches exhibit a smoother segmentation presumably as they exploit the 3D nature of the MR volumes directly.

Another thing to note in Fig. 4.8f is that segmentation provided by KK_{3D} seems worse than the rest – even than its 2D analogue. The problem does not appear to be related to the number of parameters since KK_{3D} has less trainable elements compared to DM_{3D} and UN_{3D}, according to Table 4.1. This issue might be a consequence of the architectural design itself. Anisotropic voxels and heterogeneous spacing may be affecting the low-resolution path of the network considerably. Hence, the overall performance is degraded.

Comparison with the state of the art and conventional methods

We compared our best results for each dataset against two commonly used methods: SPM and FAST. In testing time, SPM, FAST and our models could reach a whole brain segmentation within 6 min. More importantly, SPM and FAST did not require GPUs as deep learning methods do. The results are shown in Fig. 4.9. Overall, SPM and FAST led to significantly lower segmentation results compared to our best model

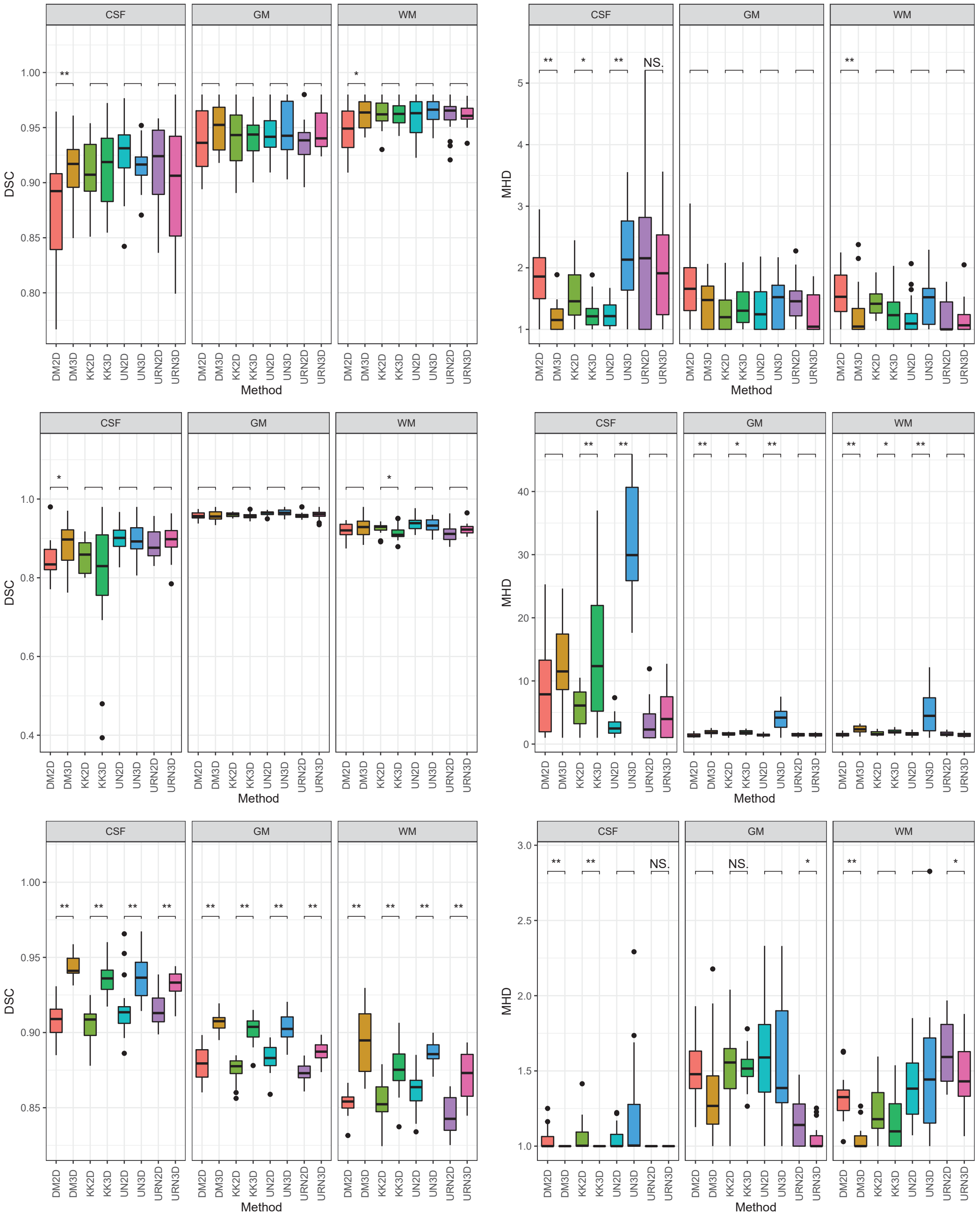
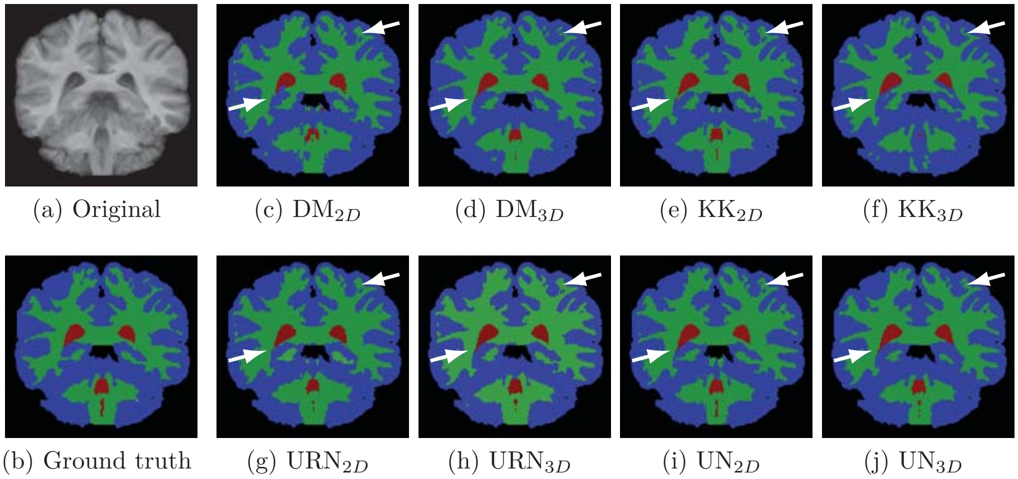


Figure 4.7: DSC (left column) and MHD (right column) values obtained using 2D and 3D versions of the same architecture. From top to bottom, boxplots for MICCAI2012, IBSR18, and iSeg2017, respectively. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. Differences between both versions of the same baseline architecture are highlighted with NS, *, and ** indicating a p-value > 0.1, < 0.05 and < 0.01, respectively. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter.


(a) Original

(c) DM_{2D}

(d) DM_{3D}

(e) KK_{2D}

(f) KK_{3D}

(b) Ground truth

(g) URN_{2D}

(h) URN_{3D}

(i) UN_{2D}

(j) UN_{3D}

Figure 4.8: Segmentation output of the eight considered methods. The ground truth is displayed in (a) and the corresponding segmentation in (b-i). DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. The colours for cerebrospinal fluid, grey matter and white matter, are red, blue and green, respectively. White arrows point out areas, where differences compared to the ground truth, are more noticeable. Architectures using 2D information were trained with axial slices.

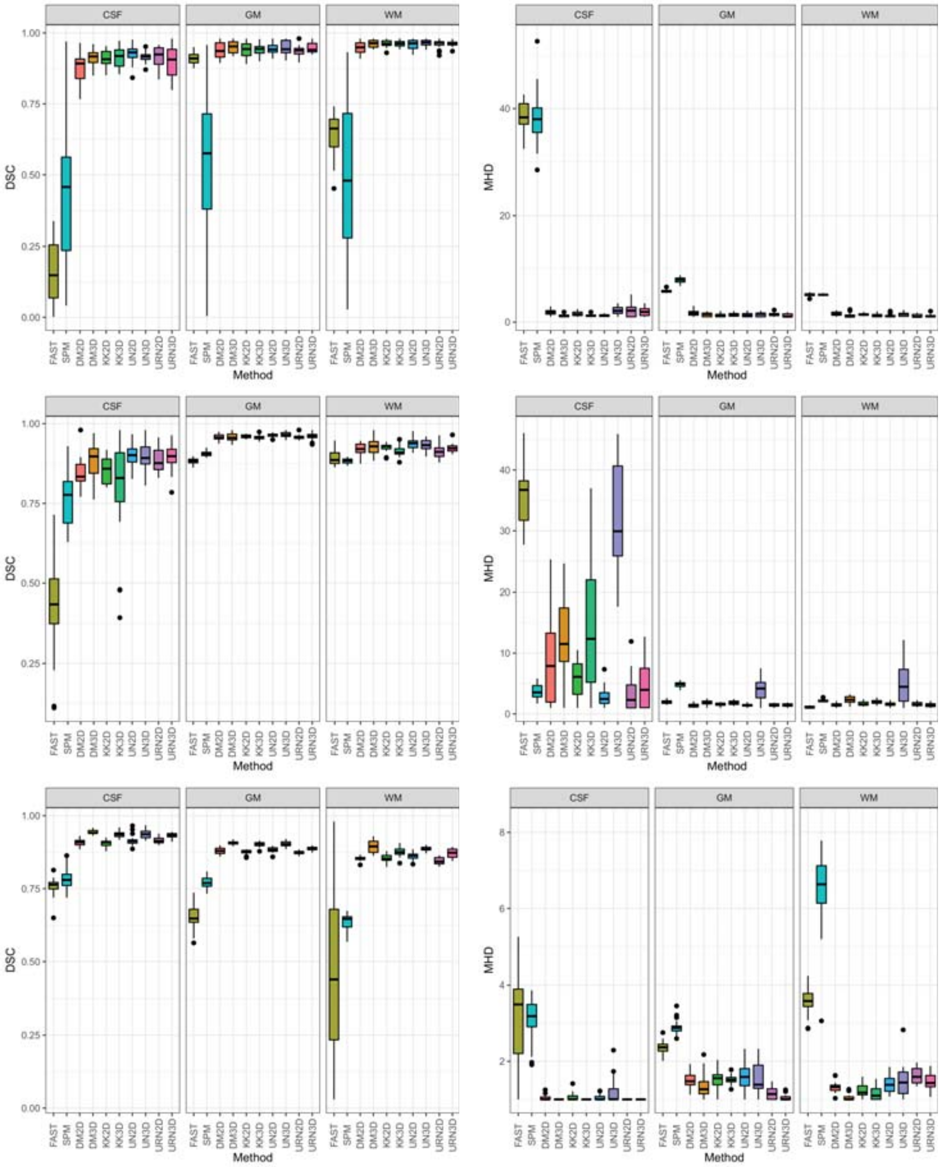


Figure 4.9: DSC (left column) and MHD (right column) values obtained by fully convolutional networks and SPM and FAST. From top to bottom, boxplots for MICCAI2012, IBSR18, and iSeg2017, respectively. DM, KK, UN, and URN refer to the networks inspired by the works of Dolz *et al.* [103], Kamnitsas *et al.* [152], Çiçek *et al.* [148], Guerrero *et al.* [159], respectively. The subindex indicates the network dimensionality. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter.

(p-value < 0.001). Nonetheless, it is essential to understand the pros and cons of each strategy. On the one hand, conventional methods are suitable for many domains, but noise, intensity inhomogeneities [94–97], overlap between tissue distributions, and variations in shape (baby brain vs adult brain) and labelling protocols, hinder obtaining accurate outputs. On the other hand, the accuracy of CNN methods tends to decrease when the distribution of the test set differs significantly from one of the training set due to variations in imaging and labelling protocols. For example, a network trained on one of the datasets would not yield top results if tested on any of the other two since the voxel spacing, image quality and delineation of the different tissues would be different; a workaround would be to adapt the weights of the network to the new domain through transfer learning or, in a practical scenario, to map all the volumes to a standard template (e.g. MNI). We believe that fusing different approaches into a single framework (e.g., convolutional neural networks with tissue segmentation priors [18, 330]) is a promising area to explore to reach robustness.

In comparison with the state of the art, our methods showed similar or enhanced performance. First, the best DSC scores for IBSR18 were collected by Valverde *et al.* [94]. The highest values for CSF, GM and WM were 0.83 ± 0.08, 0.88 ± 0.04 and 0.81 ± 0.07; while our 2D U-Net model scored 0.90 ± 0.03, 0.96 ± 0.01 and 0.93 ± 0.02, for the same classes. Second, the best-known values for tissue segmentation using the MICCAI 2012 dataset, were reported by Moeskops *et al.* [171]. Their strategy – a multi-path CNN – obtained 0.85 ± 0.04 and 0.94 ± 0.01 for CSF and WM, respectively; while our 3D U-Net model yielded 0.92 ± 0.03 and 0.96 ± 0.01. In this case, we cannot establish a direct comparison of GM scores since in Moeskops' case, this class was subdivided into (a) cortical GM and (b) basal ganglia and thalami.

4.4 Discussion

We analysed quantitatively eight FCNN architectures inspired by the literature of brain segmentation related tasks. The networks were assessed through three experiments studying the importance of (i) overlapping patch extraction, (ii) multiple modalities, and (iii) network dimensionality. To ensure that all networks were evaluated under similar and favourable conditions, we used exactly the same pipeline (i.e. pre-processing, data preparation, segmentation, and post-processing), same optimiser, and same training and validation collections, and controlled overfitting by monitoring the network performance on the validation sets.

Our first experiment evaluated the impact of overlapping as sampling strategy at training and testing stages. This overlapping sampling is explored as a workaround to the commonly used data augmentation techniques in medical image tasks. This procedure can be used in this case as none of these networks processes a whole volume at a time, but patches of it. Based on our results, the technique proved beneficial as most of the strategies obtained significantly higher values than when not considered. In particular, the four u-shaped architectures exhibited a remarkable influence of this approach, presumably since more samples are used during training and the same area is seen with different neighbouring regions, enforcing spatial consistency. Overlapping sampling in testing acted as a denoising technique. We observed that this already-incorporated tool led to better performance than when absent as it helped filling small holes in areas expected homogeneous. The improvement was found to be at least 1%. Naturally, the main drawback of this technique was the expertise of the classifier itself, since it could produce undesired outputs when poorly trained.

Our second experiment assessed the effect of single and multiple imaging sequences on the final segmentation. We observed that regardless of the segmentation network, the inclusion of various modalities led to significantly better segmentations that when using a single imaging sequence. This situation may be a consequence of networks being able to extract valuable contrast information. Improvements were noted concerning the mean as well as the dispersion of the values yielded by the methods. Although this outcome is aligned with the literature [67], further trials on more datasets should be carried out to draw stronger conclusions. Future work should consider evaluating tissue segmentation in the presence of pathologies and using more imaging sequences such as FLAIR and PD.

Our third experiment examined the influence of patch size on the final segmentation. Although the literature reports that the larger the patch size, the better the segmentation due to additional contextual information [326–329], we observed that this trend is only followed when there are enough training samples to train such a larger network. This outcome is expected as the number of parameters increases substantially as the input patch dimensions augment. Unexpectedly, small-scale versions of the u-shaped networks were able to distinguish between classes and even though the performance was significantly lower than the large variants, the median DSC and MHD values were above 80% and below 2.50 pixels, respectively. However, it is crucial to recognise that this outcome might not hold on other tasks where tissues are split into sub-classes (e.g. whole brain parcellation or subcortical structure segmentation) as more contextual information might be needed to distinguish one class from another.

Our fourth experiment evaluated significant differences between 2D and 3D methods on the three considered datasets. Although 3D architectures tend to outperform their 2D analogues, the differences may not be significant. Moreover, in one of our datasets, IBSR18, 2D versions of the same baseline architecture could reach better segmentation scores than their 3D analogues. This outcome is a consequence of the heterogeneity of the data in IBSR18, i.e. 2D methods seem to be more resilient to issues regarding voxel spacing than 3D ones. Naturally, the immediate workaround to this issue is to re-sample during pre-processing. Additionally, the situation is likely to worsen when processing highly anisotropic volumes as there is less information in the third dimension.

According to our evaluation results, the segmentation performance is not strictly conditioned by the number of trainable parameters. For example, in IBSR18, 2D networks performed better than 3D networks due to issues of 3D networks to adapt to voxel spacing variations and image quality; in MICCAI2012, the differences between the performance of 2D networks in comparison to 3D networks were not significant overall; in MICCAI2012 and IBSR18, DM3D performed almost similar or worse than u-shaped networks even though it has at least 120% additional parameters. These outcomes suggest that some inherent architectural weaknesses and strengths define the overall performance of a network. Instead, we noted that specific modules allowed some networks to outperform some others. First, we observed that models using information from shallower layers in deeper ones achieved higher performance than those using multi-resolution information directly from the input volume, namely KK$_{2D}$ and KK$_{3D}$. The difference was far more evident in datasets with heterogeneous volumes, e.g. in IBSR18 where scans vary in voxel spacing and image quality, where the latter strategy performed worse on average. This situation underlines the relevance of internal connections (e.g. residual connections and concatenation) for fusing multi-resolution information to segment more accurately. Second, we observed that concatenation and residual layers are present in all of the state-of-the-art networks. This might be related to the fact that these types of connections help in dealing with the degradation problem (i.e. deep networks tend to saturate and degrade rapidly) [85]. As the residual layers reduce the number of parameters to optimise, they should be preferred over concatenation modules. In fact, our experiments showed that two similar u-shaped networks using both approaches achieved similar results. Third, although u-shaped networks tended to outperform convolutional-only networks, no significant/remarkable difference was seen between both design patterns, except for processing times. In both training and testing, u-shaped networks segmented faster than convolutional-only networks: u-shaped models require extracting less number of patches and provide a more prominent output at a time.

Regarding general performance, two methods, DM$_{3D}$ and UN$_{3D}$, obtained the best results. Of note, our specific implementation of the latter architecture required 30% fewer parameters to be set than the former and classified $\approx$ 32K voxels more at a time and completed a whole volume segmentation in half of the time or less. Although URN networks use slightly fewer parameters than UN architectures, both of them have comparable response times. In general, should the priority be overall processing time (training and testing), u-shaped networks are a suitable and recommended approach to address tissue segmentation instead of convolutional-only approaches.

4.5 Participation in international tissue segmentation challenges

During the development of this doctoral thesis, we proposed networks for cross-sectional tissue segmentation that we submitted to three Grand Challenges of the International Conference on Medical Image Computer and Computer-Assisted Interventions (2017-2019). Our proposals reflect different time points of our work: we started with Dolz Multi in 2017 and shifted towards U-Nets based on the findings of our quantitative comparison.

4.5.1 Six-month old infant brain MRI tissue segmentation 2017

The aim of the Grand Challenge on six-month old infant brain MRI tissue segmentation at MICCAI 2017 (iSeg2017) was to provide a platform for comparing tissue segmentation algorithms on baby brains during their isointense phase. In this phase, brains go through a maturation and myelination process which results in an inverted signal between white and grey matter [331] and a reduced contrast between these two brain tissues [322]. The dataset consisted of T1-w and T2-w scans from 23 subjects which were acquired using the same imaging protocol. The challenge organisers segmented brain tissues automatically and edited them subsequently to correct segmentation errors. All 23 pairs of T1-w and T2-w scans were released to the public, 10 of them included the ground truth segmentation and the remaining 13 were used for testing.

For this challenge, we addressed the problem using a fully automatic pipeline consisting of four steps: pre-processing, data preparation, classification, and post-processing. First, we normalised intensities of all training volumes using the z-score approach (i.e. zero mean and unit standard deviation). Second, we extracted overlapping patches of $27 \times 27 \times 27$ from all the pre-processed training volumes along with their corresponding labels (central $9 \times 9 \times 9$ block). Third, using the patches and the labels, we trained our 3D multi-sequence multi-path FCNN based on the DM_{3D} network [103]. We used this specific architecture since it worked well for segmenting tissues in baby brains according to our tests in Section 4.3. We trained our network with 75% of the volumes and the remaining 25% for validation. Fourth, for each case in the test set, we standardised intensities using the mean and the standard deviation calculated in Step 1, extracted overlapping patches, and classified. Fifth, once we processed all patches, we reconstructed the segmentation.

We performed a leave-one-out cross-validation process on the training set and obtained DSC values of 0.973 ± 0.010, 0.917 ± 0.012, 0.887 ± 0.018 for CSF, GM, and WM, respectively. We used the resulting model to segment testing cases and submitted our results to the challenge organisers obtaining DSC values of 0.951 ± 0.005, 0.910 ± 0.008, 0.885 ± 0.015 for the same classes and achieving top-5 (out of 21) performance in 6 out of 9 performance metrics[6].

4.5.2 MR brain segmentation 2018

The goal of the Grand Challenge on MR Brain Segmentation at MICCAI 2018 (MRBrainS18) was to evaluate the robustness of different tissue segmentation methods against brain lesions (particularly, white matter hyperintensities and infarctions). The dataset consisted of T1-w, T1-IR, and FLAIR scans from 30 subjects which were acquired using the same imaging protocols. The challenge organisers segmented regions of interest into cortical and deep grey matter, white matter, white matter hyperintensities, sulcal and ventricular cerebrospinal fluid, cerebellum, brain stem, and infarction. Only data from seven subjects were released to the public while the rest were used for testing (23).

For this challenge, we addressed the problem using a fully automatic pipeline consisting of four steps: pre-processing, data preparation, classification, and post-processing. Pre-processing consisted of skull stripping with ROBEX [208], and tissue segmentation using SPM [56] and FAST [57]. First, we removed non-brain areas using the preprocessed T1-w. Second, we input the obtained volume into the two segmentation algorithms. Note that ROBEX removes vessels and non-brain structures (e.g. cerebral falx and choroid plexus) which are labelled in the challenge dataset as CSF. Thus, we solely considered this mask as guide of the brain area. Data preparation corresponded to tiling volumes up and selecting relevant blocks. For both training and testing, blocks were extracted with 50% overlap. For training only, patches were considered if their brain content corresponded to at least 30% of the whole block. Data were extracted from nine sources: T1-w, FLAIR, brain mask, and three tissue segmentation outputs obtain with both FAST and SPM. We segmented incoming volumes using an ensemble of multi-path u-shaped networks where each network was composed of two u-shaped paths inspired by the work of Guerrero *et al.* [159]. We resorted to use U-Nets as baseline since we needed to make fast tests in short time and they were the best option according to our experimental results in Section 4.3. Unlike the original work, 3D volumes were processed directly, PReLU activations were used instead of ReLU, and activations were used after every

[6] URL: `http://iseg2017.web.unc.edu/rules/results/`. Team name “nic_vicorob”

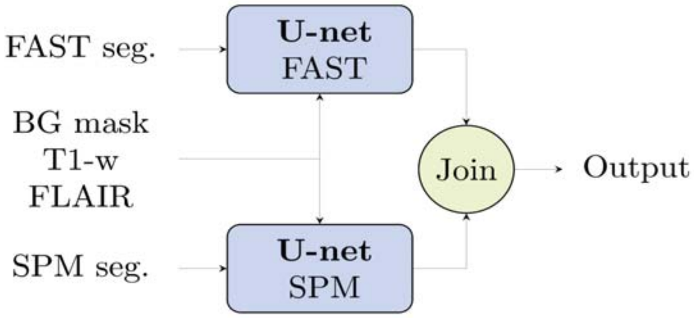
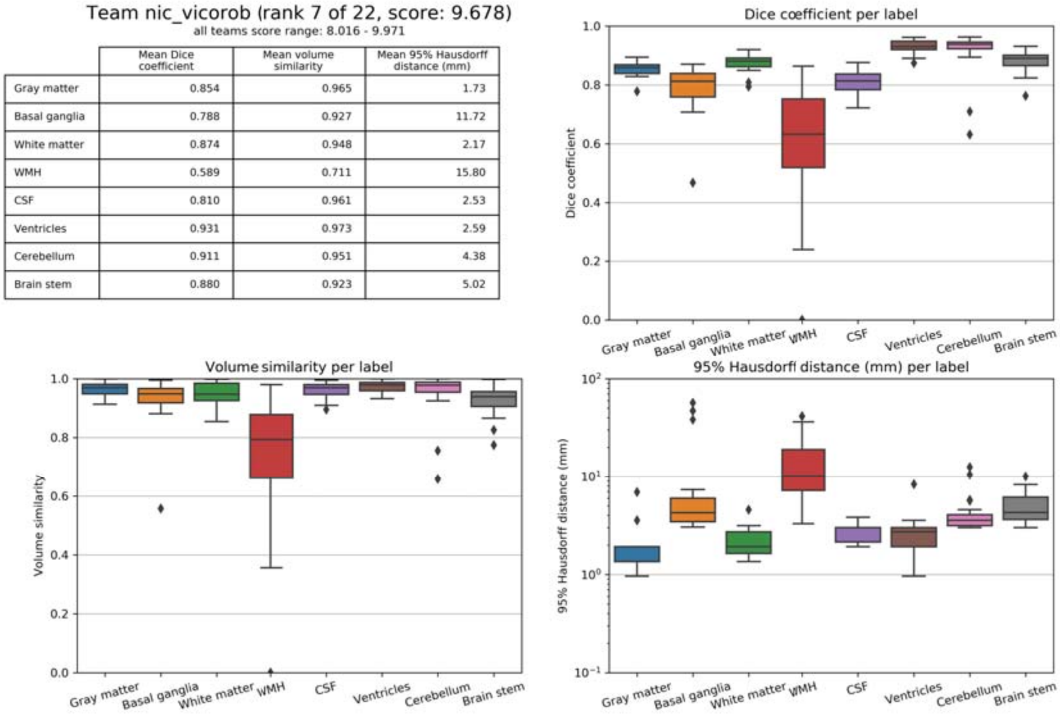


Figure 4.10: High-level scheme of each network in the ensemble of U-Nets. Our proposal for addressing the MRBrainS18 challenge consisted of an ensemble of seven U-Nets using tissue segmentation priors obtained through validated cross-sectional tissue segmentation tools: FAST and SPM. Each network was formed by two U-Nets, both of them using the information provided by T1-w, FLAIR, and brain mask. We gave FAST segmentations to one of the paths, and SPM segmentations to the other one. The outputs of both paths were fused in a late fusion fashion. Similarly, the outputs of each network within the ensemble were combined in the same way to provide a final verdict.

addition module. The two paths were input with T1-w, FLAIR and brain mask, as illustrated in Fig. 4.10. While one path was provided with FAST segmentation, the other one was given the ones of SPM. The outputs of both paths were fused in a late fusion fashion. Similarly, the outputs of each network within the ensemble were combined in the same way to provide a final verdict. Postprocessing consisted of reconstructing the segmented volume by overlaying neighbouring predictions. As output patches overlap, we provided voxel labels through majority voting.

As there are seven training cases in the MRBrainS18 challenge, we trained seven different multi-path u-nets using a leave-one-out cross-validation strategy and put them together to achieve a robust segmentation outcome. We trained all networks for a maximum of 100 epochs using an early stopping policy with patience equal to 10. We submitted the ensemble as a Docker to the challenge and ranked 7 among 22 groups, as shown in Fig. 4.11[7].

[7] URL: https://mrbrains18.isi.uu.nl/results/eight-label-segmentation-results/. Team name “nic_vicorob”



Team nic_vicorob (rank 7 of 22, score: 9.678)

all teams score range: 8.016 - 9.971

	Mean Dice coefficient	Mean volume similarity	Mean 95% Hausdorff distance (mm)
Gray matter	0.854	0.965	1.73
Basal ganglia	0.788	0.927	11.72
White matter	0.874	0.948	2.17
WMH	0.589	0.711	15.80
CSF	0.810	0.961	2.53
Ventricles	0.931	0.973	2.59
Cerebellum	0.911	0.951	4.38
Brain stem	0.880	0.923	5.02

Figure 4.11: Performance of our ensemble of U-Nets on the MRBrainS18 challenge. Results extracted from the MRBrainS18 challenge webpage.

4.5.3 Six-month old infant brain MRI tissue segmentation 2019

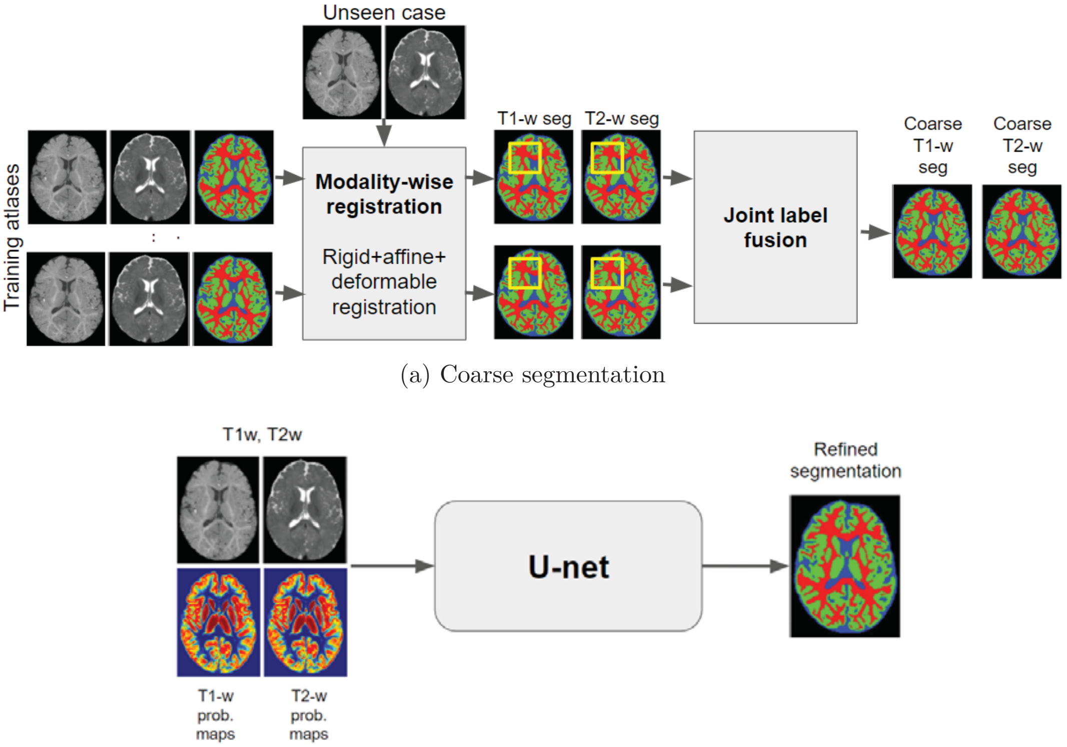
The aim of the Grand Challenge on six-month old infant brain MRI tissue segmentation at MICCAI 2019 (iSeg2019) was to evaluate the robustness of different methods to variations in the acquisitions (different sites, scanners and imaging protocols). Like the iSeg2017, these baby brains exhibited the low contrast and inverted signal between brain tissues. The dataset consisted of T1-w and T2-w scans from 39 subjects which were acquired using three different imaging protocols from three different sites. The challenge organisers segmented brain tissues automatically and edited them subsequently to correct segmentation errors. All 39 pairs of T1-w and T2-w scans were released to the public, 10 of them included the ground truth segmentation, the remaining 19 were used for testing.

Despite the outstanding performance of deep learning in many fields [122], domain shift (e.g. intensity range variations) continues being a challenge for these kinds of techniques [198]. Thus, for this challenge, we resorted to using a hybrid approximation leveraging convolutional neural networks and multi-atlas segmentation to segment brain tissues. Our processing pipeline consisted of two main steps: coarse segmentation, and refinement. We illustrate them in Fig. 4.12. First, we applied the multi-atlas segmentation with joint label fusion [332] to obtain a coarse segmentation of the grey and white matter and cerebrospinal fluid spaces. We registered the ten training atlases to each volume in the validation and the test sets using rigid, affine, and deformable transformations, in that order, and, subsequently, propagated and fused the corresponding labels to get a coarse response. We used this technique on both T1w and T2w since we found multi-modal information to be beneficial [19]. Thus, we obtained two segmentation probability maps for each subject, one per modality. Second, we used a u-shaped fully convolutional neural network based on the implementation of Guerrero *et al.* [159] to refine these segmentation maps to produce a smoother/enhanced response (similar to the network devised for MRBrainS18 in Section 4.3). We used U-Nets as baseline since we needed to make fast tests in short time and they were the best option according to our experimental results in Section 4.3 and Section 4.5.2. In such a way, the network did not face intensity variation problems per se as probability maps are already “normalised”.

The goal of the network was to learn the errors the multi-atlas segmentation approach makes and compensate for them [18, 333]. The training process consisted of the following steps. First, we split the given training dataset into training, validation, and testing (80%, 10%, 10%). Second, we registered both training and validation sets to the test case and segmented tissues to obtain the corresponding probability maps. Third, we tiled up each volume in each set into overlapping blocks. Fourth, we used them along with the corresponding segmentation labels to train the network. We used batches of 5 for a maximum of 500 epochs. At the end of each epoch, we computed the performance on the validation set. The training phase stopped after 20 consecutive epochs of no improvement and we kept the model leading to the lowest loss function value. We considered random offsets, flips, and permutations for data augmentation during training.

The steps to test a trained model on a given input MR volume are as follows. First, we divide the baseline input volume into overlapping blocks. Second, we input the patches to the network to obtain refined segmentation maps. Third, as there is overlap between output blocks, we provide the final segmentation through means of averaging. We rearrange all patches to reconstruct the corresponding segmentation volume.

We performed a leave-one-out cross-validation process on the training set to verify whether the segmentation improved after refining as hypothesised. The results we obtained are shown in Fig. 4.13. We noticed that in most cases the performance of the hybrid framework was significantly superior to that of obtained through the atlas-based segmentation method. We segmented the test set cases and submitted our results to the challenge organisers obtaining the following Dice similarity co-



(a) Coarse segmentation

(b) Segmentation refinement

Figure 4.12: Processing pipeline considered for the iSeg2019 challenge. First, we applied the multi-atlas segmentation with joint label fusion [332] to obtain a coarse segmentation of the cerebrospinal fluid, grey matter, and white matter. Second, we used a u-shaped fully convolutional neural network based on the implementation of Guerrero *et al.* [159] to refine these segmentation maps to produce a smoother/enhanced response.

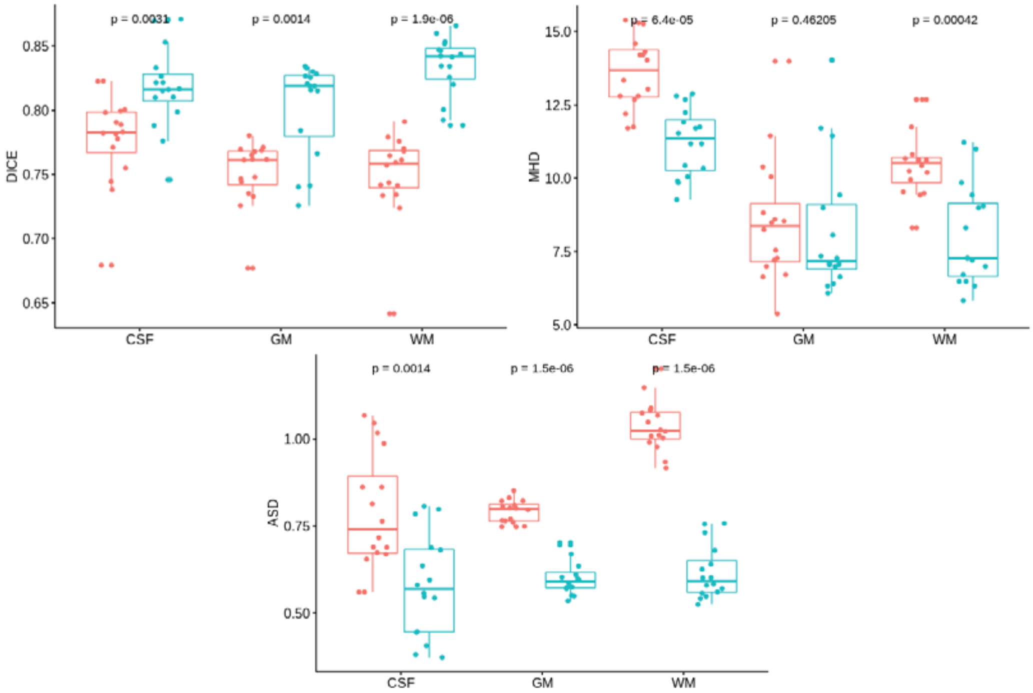


Figure 4.13: Leave-one-out cross-validation performance obtained using the multi-atlas segmentation method only (pink) and our proposed hybrid multi-atlas and convolutional neural network based framework (blue). We tested for differences in the performance of the methods using the Wilcoxon signed-rank test. DICE: Dice similarity coefficient. MHD: modified Hausdorff distance. ASD: average surface distance. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter.

efficient (DSC), modified Hausdorff distance (MHD), and average surface distance (ASD): DSC - CSF: 0.778±0.035, DSC - GM: 0.751±0.025, DSC - WM: 0.749±0.034; MHD - CSF: 13.604 ± 1.232, MHD - GM: 8.522 ± 2.126, MHD - WM: 10.507 ± 1.145; ASD - CSF: 0.789 ± 0.168, ASD - GM: 0.791 ± 0.032, ASD - WM: 1.037 ± 0.073[8].

4.6 Summary

In this chapter, we studied the relevance of patch sampling, multiple modalities, and dimensionality on the performance of the network. We implemented literature-inspired fully convolutional neural networks for tissue segmentation on brain MRI. The networks were compared under a common evaluation framework to establish a direct comparison between the different methods and, consequently, understand the underlying properties of the various architecture directives.

In general, we observed that extracting patches with a certain degree of overlap among themselves, processing multiple sources of information (e.g. multiple modalities) and at multiple scales, and larger patch size led to improved segmentation performance. Although 3D networks tended to outperform their 2D counterparts, they were more more affected by variations in image resolution and voxel spacing. U-shaped networks reached higher DSC and lower MHD values than convolutional-only architectures overall and had the best response time. Therefore, we concluded that U-Nets are particularly suitable for brain MRI tissue segmentation.

We achieved compelling performance for IBSR18, MICCAI2012, iSeg2017, MRBrainS18, and iSeg2019 with our implemented and proposed approaches. Two important things to note in this Chapter. First, we did not tweak any of these networks; a common processing pipeline has been used. Hence, it was possible to compare them under similar conditions. Approaches expressly tuned for challenges may win, but it does not imply they will work identically – using the same set-up – on real-life scenarios. Second, although these strategies showed acceptable results, the performance of these networks might be compromised due to the domain shift problem. However, recent advances in domain adaptation and transfer learning have demonstrated that a trained convolutional network can be fine-tuned successfully to a new domain as long as there are a few training cases from it [198, 199].

Evaluating the performance of cross-sectional tissue segmentation methods can be carried out without complication since there are various publicly available and well-annotated datasets. However, to our knowledge, there is no annotated dataset for evaluating the accuracy of longitudinal methods. In the next chapter, we propose a framework for generating longitudinal atrophy datasets and allowing evaluating

[8] URL: `http://iseg2019.web.unc.edu/evaluation-results`. Team name “nic_vicorob”

the accuracy of atrophy quantification methods and training deep learning methods for performing such a task.

Chapter 5

Generating longitudinal atrophy evaluation datasets

In this chapter, we propose and validate a framework for generating longitudinal cerebral atrophy datasets. This work has been submitted to **Neuroinformatics**, where it is now under second revision.

Paper under second revision in **Neuroinformatics** Submitted: April 2020

JCR IF 5.127, Q1(9/106)

5.1 Introduction

Now that we have seen that deep learning for cross-sectional tissue segmentation and brain volumetry works, we dig into how can it be used for longitudinal atrophy quantification. Although some longitudinal brain MRI datasets are available to the public[1], accuracy is rarely assessed since manual segmentation is tedious, time-consuming, and error-prone, and conventional automatic segmentation tools exhibit inaccuracies [334]. Instead, the evaluation is carried out at the level of scan-rescan error and statistical power. To examine the former aspect, patients are scanned multiple times in different scanners in short periods of time, to ensure minimal brain changes, and brain volumetry methods are judged based on their precision. The latter aspect aims to determine whether these approximations can discern between patients undergoing different treatments/pathologies. Commonly, the exercise consists of examining how well can algorithms discern between populations ongoing different treatments/diseases (e.g. dementia versus control). Nonetheless, such an evaluation does not reflect the accuracy of the methods. Synthetic image generation could be used to generate controlled evaluation environments where ground truth is available and known beforehand.

In medical image analysis, image generation approaches have been applied to assess registration, estimate and correct bias in longitudinal atrophy analyses, generate absent modalities and augment training sets [169, 170, 191, 335–342]. The techniques range from transformation models mimicking brain tissue loss to adversarial/generative networks with problem-specific loss functions. Karaçali *et al.* [335] devised a method for deforming MRI scans such that the atrophy extent corresponded to the requested one[2]. The downfall of such an approach is that resulting deformation patterns cannot be controlled locally and follow a topology-preserving strategy which might not permit mimicking multiple pathologies. Roy *et al.* [337] used patch-based dictionary learning to estimate a mapping function between two imaging sequences or image acquisition protocols[3], e.g. making it appealing in retrospective harmonisation pipelines. However, its direct usage for atrophy generation might not be feasible since the technique does not deform the brain but finds matching intensity values between imaging modalities. Chartsias *et al.* [169] proposed a framework to synthesise MR modalities from others using encoder-decoder CNNs and modality-invariant latent spaces[4]. Apart from the modality synthesis, the authors showed the potential of the framework to in-paint white matter hyperintensities onto normal-appearing tissue and the usage of multiple losses to achieve realistic synthesis. Inspired by their work, Salem *et al.* [170] devised a proposal to generate

[1] See `http://freesurfer.net/fswiki/LongitudinalData`

[2] Available at `http://web.iyte.edu.tr/~bilgekaracali/VoxelVolumeMatching.tar.gz`

[3] Available at `https://www.nitrc.org/projects/image_synthesis/`

[4] Available at `https://github.com/agis85/multimodal_brain_synthesis.`

synthetic yet realistic MS lesions as an image augmentation strategy[5]. Evidently, a similar principle could be considered for generating atrophy. Shin *et al.* [191] developed a proposal in which realistic MRI scans were generated from brain anatomy and tumour segmentation masks using conditional generative adversarial networks (CGAN). The authors showed that their approach could be used for dealing with the lack of diverse, sufficient, and correctly annotated data. Although their code is not available in principle, their proposal is inspired by the image-to-image translation with CGAN [343][6]. Up to our knowledge, these types of architectures have not been considered for longitudinal data generation, but they can be extended for this purpose by giving the network the baseline scan and the segmentation map of the follow-up acquisition.

In this chapter, we use a cascaded U-Net trained with our own region-wise loss function to deform a given T1-w scan based on the information provided through tissue probability maps. This setting allows building longitudinal collections for assessing atrophy quantification methods as the tissue loss between original and generated scans is controlled, induced, and known beforehand. Note that our aim is not to predict the atrophy that a patient will suffer in a certain amount of time, but a prediction of what would be the brain appearance given a tissue change (segmentation). The relevance of this work is two-fold. First, our proposal allows comparing atrophy quantification tools quantitatively. Second, it can serve as ground truth for training deep learning approaches for atrophy quantification.

5.2 Methods

Our proposed atrophy generation framework is depicted in Fig. 5.1. Given a baseline T1-w scan and its modified tissue probability maps, the goal of our framework is to alter the input such that brain tissues are altered as requested. In such a way the atrophy between the baseline and generated images is known in advance. We take a T1-w scan, segment its regions using conventional tissue segmentation tools, alter its segmentation probability maps manually or automatically, and plug both the baseline T1-w scan and the resulting probability maps into the generation network to create a synthetic volume.

Note that the way the framework has been structured is advantageous as a plethora of scans can be generated by modifying the input tissue segmentation maps (e.g. manually, using morphological operations, or pathology-related deformation fields [344]). We apply real deformation fields to alter the original segmentation probability maps. Further details of the approach are discussed in the following

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

[6] Available at `https://github.com/phillipi/pix2pix`

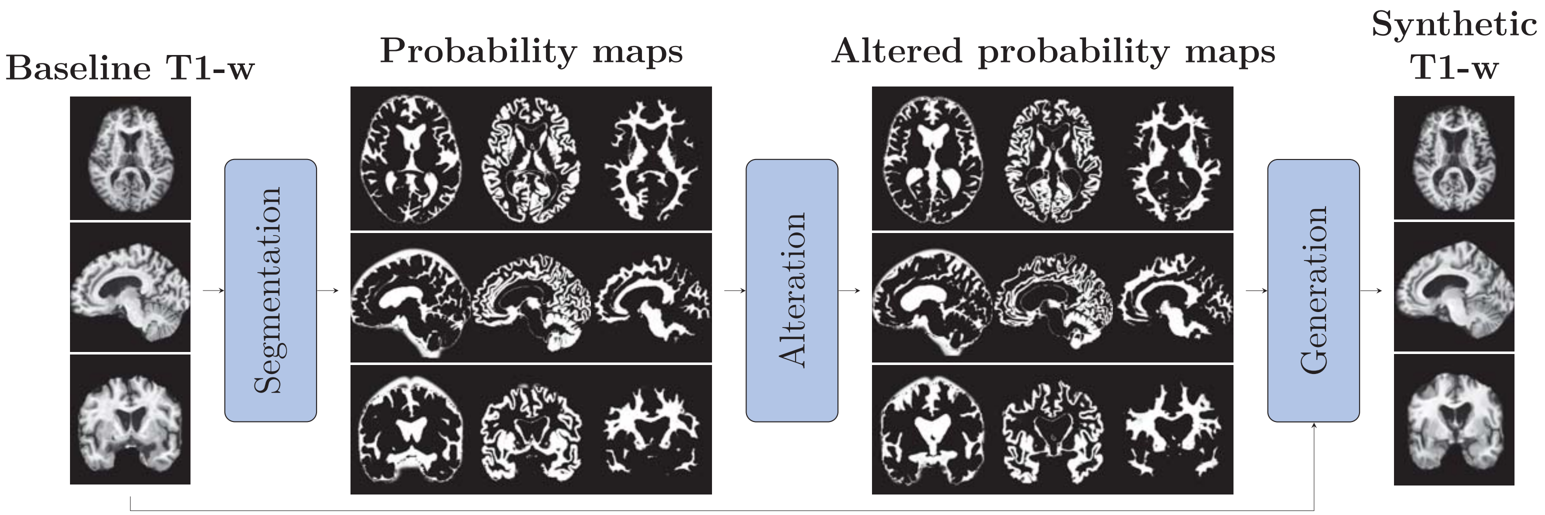


Figure 5.1: Inducing controlled tissue variations. We take a baseline T1-w scan, segment it, alter its segmentation probability maps manually or automatically, and plug it into the generation network to create a synthetic volume. We apply conventional tools for tissue segmentation and real deformation fields to alter original segmentation probability maps. Given a baseline T1-w input image and modified tissue probability maps, the goal of our framework is to generate a T1-w scan in which the tissues are altered as requested. In this example, tissue changes were requested in both cortical and periventricular regions, e.g. lateral ventricles appear enlarged (all three views), Sylvian fissures have been altered (axial and coronal), and the third ventricle seems more atrophied (coronal).

sections.

5.2.1 Processing pipeline

Our processing pipeline contemplates four essential components: pre-processing, data preparation, processing, and reconstruction.

Pre-processing consists of (i) skull stripping with ROBEX [208], (ii) histogram matching [345] to fix voxel values to a common range, and (iii) registration to the MNI space as harmonising step. The first step allows discarding non-relevant areas that may affect the generation process as they are commonly hyperintense in T1-w. We chose ROBEX since it is an unsupervised method that delivered consistent and robust results when compared to conventional methods. The second step allows mapping voxel intensities to a reference range. This procedure is essential to reduce issues regarding generalisability due to intensity shifts [74]. The third step permits using the same network on various datasets as reducing the heterogeneity of voxel spacing may enhance the overall performance [19].

Data preparation consists of splitting input volumes into patches. For both training and testing, we extract overlapping blocks to gather more samples, reduce block boundary artefacts, and enforce spatial consistency [19]. Additionally, we discard empty or partially empty training patches to prevent building background-biased generators. We set the minimum content rate and overlap extent to 30% and 50%, respectively. Both values were favourable experimentally.

In the processing step, we pass each tuple of patches extracted from the baseline scan and modified probability maps through the network in batches of 32 elements at a time. We did not increase this parameter due to hardware constraints.

We overlay neighbouring predictions to reconstruct the synthetic volume and provide voxel-wise responses through averaging. We run histogram matching on the reconstructed volume to ensure intensity range similarity. No further post-processing is required.

5.2.2 Generation architecture

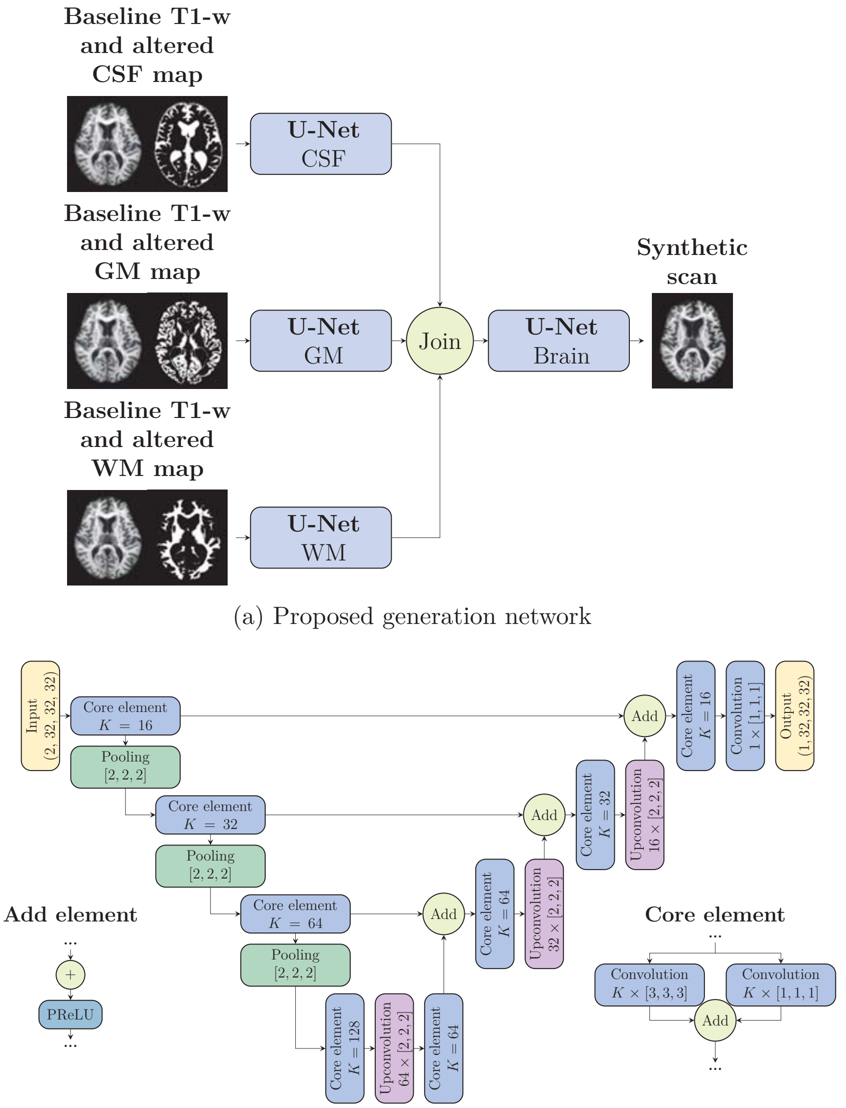
Our proposed network follows a cascaded U-Net construction scheme, as illustrated in Fig. 5.2(a). First, we input the baseline scan and its modified tissue probability maps into three networks arranged in parallel. Each one of these networks accounts separately for changes in cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM). Second, we append and pass the resulting individual latent representations to another u-shaped network which merges them effectively to produce the final output. In our implementation, the input and output patches have the same height, width and depth, 32 voxels in each dimension. The overall cascaded network is trained end-to-end, i.e. none of the sub-nets is trained independently.

Each U-Net module comprises a contracting path, performing consecutive convolution and down-sampling operations, and an expansive path, carrying successive up-sampling and convolutions. In this way, it is possible to output a patch with the same dimensions as the input while reducing response times. The architecture is illustrated in Fig. 5.2(b). The network consists of $8 \times 2 + 1$ convolutional layers – eight pairs occur in *parallel*, as shown in the lower right corner of Fig. 5.2(b) – three down-sampling modules and three backward strided convolutions. The number of kernels doubles per contracting path layer from 2^4, in its shallowest, to 2^7, in its latent space, and afterwards halves per expansive path layer until the kernels are 2^4, in its deepest level. Strides for down-sampling and up-convolutions are set to $2 \times 2 \times 2$.

The U-Nets are equipped with filter banks of varied sizes in a Network-in-Network (NIN) resembling scheme [140,147]. These modules, implemented as $1 \times 1 \times 1$-kernel layers, act similar to embedded multi-layer perceptrons which enhance the discriminant and representation power of the overall model. These processing components are referred to as *core elements* in Fig. 5.2(b).

Each sub-module uses residual connections to merge feature maps from higher-resolution layers with de-convolved maps to preserve localisation details and improve back-propagation [85]. Moreover, each sub-module combines feature maps by adding them and not concatenating them as widespread [148,150]. This option is preferred to reduce the cardinality of the trainable parameter set. Note the different channels are processed in an early fusion fashion [167].

The design of the sub-modules is inspired by the work of [159]. The main differences are the dimensionality of the network, the downsampling approach, and the type and location of non-linear activation layers. First, the network is extended to process 3D data directly. This strategy is considered instead of a slice-by-slice approach to exploit the nature of MRI, incorporate contextual information from the three orthogonal planes, and produce more consistent results. Second, strided convolutions are used instead of max-pooling layers [117] to achieve improved performance. Third, the Rectified Linear Unit (ReLU) layers used in the original work are exchanged for Parametric ReLU (PReLU) [113]. This asset helps the model to cope with issues regarding the gradient update and empirical performance [86, 106, 122]. Fourth, these rectifier layers are used after every addition of feature maps. This choice promotes sparsity within the network, i.e. a more resilient representation [104].



(a) Proposed generation network

(b) Specific implementation of each U-Net

Figure 5.2: High level design of the proposed generation network. In (a), the model receives four inputs: a baseline T1-w acquisition and three tissue probability maps. This information is processed by three u-shaped networks, as illustrated in (b), each one specialised in generating cerebrospinal fluid, grey matter and white matter areas, and then merged by a fourth U-Net (U-Net Brain in (a)) to produced smooth reconstructions. Our specific implementation requires optimising approximately 10M parameters. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter.

5.2.3 Region-wise loss function

Atrophy quantification algorithms perform tissue segmentation and/or linear and non-linear registration. These widespread practices impose three constraints on the generation: (i) tissue contrast should be sufficiently high to be segmentation-feasible, (ii) synthesised volumes should appear visually similar to the actual scans at intensity level, and (iii) brain boundaries should be well-defined. We propose a four-objective loss function to fulfil these needs and train the whole model properly. Each objective evaluates the similarity between the expected and synthesised volume in the CSF, GW, WM, and whole intracranial volume. Given a real scan, y, its corresponding tissue probability maps, s_{CSF}, s_{GM}, and s_{WM}, and an approximation obtained with our model, $\tilde{y}$, the region-wise mean square error (RWMSE) loss function is defined as follows

$$\mathcal{L}(y, \tilde{y}) = \underbrace{L(y, \tilde{y}; \sum_{\text{ROI}} s_{\text{ROI}})}_{\textit{Combined}} + \underbrace{\sum_{\text{ROI}} L(y, \tilde{y}; s_{\text{ROI}})}_{\textit{Individual}}, \tag{5.1}$$

$$L(y, \tilde{y}; s) = \frac{1}{M \cdot N \cdot P} \sum_{v=1}^{M \cdot N \cdot P} H(s(v)) \cdot ||y_v - \tilde{y}_v||_1, \tag{5.2}$$

where $H(a)$ is the discrete heaviside step function. While the loss for overall reconstruction is back-propagated from the last layer of the network, the others affect the parallel U-Nets disjointly – i.e. one loss per path. Hence, the parallel sub-modules are in charge of generating tissue changes and the merging network of combining them smoothly.

This loss function requires segmentation priors of the follow-up volume, s_i in Eq. 5.1. This information is passed to the network to provide notions of the CSF, GM, and WM regions and specialise each path of the network towards generating realistic T1-w scans. This input can be obtained using a ground truth if available, validated segmentation tools (FAST or SPM) or cross-sectional deep learning models fine tuned to the incoming data. In our case, we use FAST to obtain tissue probability maps as it does not require any retraining.

5.2.4 Generating controlled evaluation environments

Once the network is trained using real baseline and follow-up acquisitions, we use it to generate controlled atrophy change evaluation environments, as illustrated in Fig. 5.1. The process consists of gradually increasing the overall tissue loss to establish whether our tool can generate various extents of deformation accurately. Segmentation maps can be altered in various ways. For instance, they could be dilated or eroded using morphological operations. However, this will not mimic

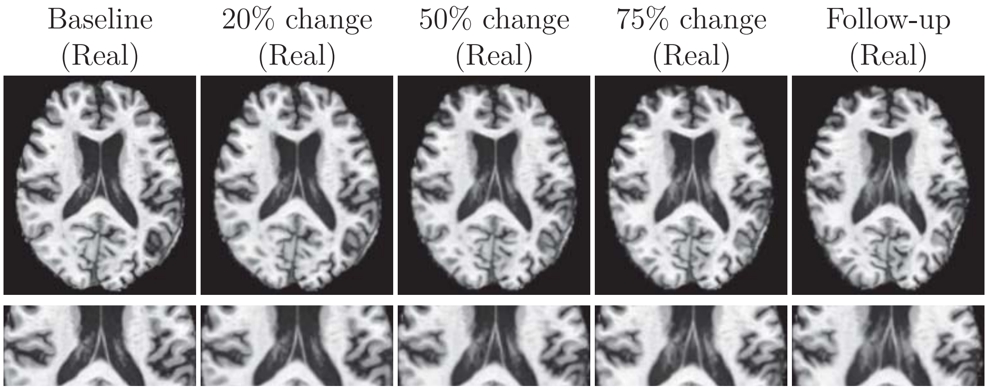


Figure 5.3: Examples of intermediate atrophy extents generated using real deformation fields.

pathological processes altering brain tissue as atrophy changes are not necessarily even in all brain regions. Alternatively, real atrophy deformation fields could be used to modify the segmentation maps. We compute real deformation fields, using FNIRT [346], from patients exhibiting the largest tissue loss and use them to alter baseline tissue segmentation maps. We multiply the resulting deformation vectors by scalars to obtain intermediate stages, as depicted in Fig. 5.3.

5.2.5 Implementation details

Network training

The steps to train our model on a given dataset are as follows. First, we split the training set into training and validation at random – 70% and 30% of the volumes, respectively. Second, we train the network in batches of 32 (default parameter value) for a maximum of 100 epochs. At the end of each epoch, we compute the performance on the validation set. The training phase stops after 10 consecutive epochs without improvement. We retain the model leading to the lowest loss function value. We optimise the models using the Adam [132] optimisation method with an initial learning rate of 1×10^{-3}, a decay of 0, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ (i.e. default parameter values, as suggested in the original paper).

Network testing

The steps to test a trained model on a given input MR volume are as follows. First, we divide the baseline input volume and the modified probability segmentation maps into patches. We extract these patches from the entire input and not from specific regions. Second, we input the patches to the network to obtain synthetic blocks. Third, as there is overlap between output blocks, we provide the final segmentation through means of averaging. We rearrange all synthetic patches to reconstruct the corresponding synthetic volume.

Software and hardware

We implement all the architectures from scratch in Python, using the Keras library. We run all the experiments on a GNU/Linux machine box running Ubuntu 16.04, with 128GB RAM. We train and test our models using a single GeForce GTX 1080-TI GPU (NVIDIA Corp., United States) with 11GB RAM. The developed framework is available to download at our research website[7].

5.3 Experiments and results

In this section, we describe the considered datasets, performance evaluation measurements, implementation details, and experiments evaluating our proposed model and corresponding results. The experiments assess loss function and architecture selection, image generation quality, and whether induced changes are detectable by conventional brain volumetry methods. Further details of each experiment and the outcomes are described in the following sections.

5.3.1 Considered datasets

We considered two publicly available longitudinal MRI repositories: the Open Access Series of Imaging Studies (OASIS) [347] and the Alzheimer's Disease Neuroimaging Initiative (ADNI)[8]. Relevant information of each dataset is presented in Table 5.1. The OASIS2 dataset was split, for easing downloading, into two sets. We refer to those as $O1$ and $O2$ from hereon. The former set contains 169 pairs of baseline follow-up cases and the second one 126. The ADNI collection contains a plethora of longitudinal cases and, hence, we opted to filter some cases. We used only cases of ADNI2 subjects with Alzheimer's disease which scans were bias field corrected

[7] `https://github.com/NIC-VICOROB/atrophy_generation`

[8] `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). For up-to-date information, see `www.adni-info.org`.

Table 5.1: Relevant information from the two considered datasets. The items to describe each dataset are listed in the first column. Although the average reconstruction matrix of the ADNI dataset is the one indicated below, the actual dimensions vary. Pairs refer to tuples of baseline and follow-up acquisitions.

Item	**OASIS2**	**ADNI**
No. of pairs	295	289
No. of time-points (max.)	5	5
Voxel spacing	$1.0 \times 1.0 \times 1.3$	$1.2 \times 1.0 \times 1.0$
Reconstruction matrix	$256 \times 256 \times 128$	$196 \times 256 \times 256$
Bias-field corrected	No	Yes
Intensity standardised	No	No
Skull stripped	No	No
Sets and no. of pairs	$O_1 : 169$, $O_2 : 126$	$A_1 : 153$, $A_2 : 136$

and coregistered correctly using FLIRT [348, 349]. Unlike in the OASIS2 case, the database was not divided in principle. Thus, we split it into two sets, $A1$ and $A2$, with 153 and 136 pairs of cases, respectively. For the sake of reproducibility, we attach the list of selected cases in Appendix A.2.

The distribution of relative CSF change between baseline and follow-up scans for OASIS and ADNI2 is illustrated in Fig. 5.4. The majority of cases were concentrated within $[0.45, 0.55]$ for the OASIS2 dataset and $[0.30, 0.50]$ for the ADNI2 dataset, but ADNI contained more cases with values above 1.00.

5.3.2 Evaluation metrics

Our generation framework should produce synthetic scans of such a quality that they resemble real ones. In this work, we scrutinised generation quality by comparing real and synthetic scans in terms of their perceptual properties and their tissue segmentation and cerebral atrophy quantification results.

Image quality

We assessed the quality of our generations with respect to that of real scans locally and globally. Locally, we measured voxel-wise intensity differences between a real scan, y, and its approximation, $\tilde{y}$, using the following expression

$$MAE(y, \tilde{y}) = \text{median } |y - \tilde{y}|. \tag{5.3}$$

The MAE approaches zero as voxel-wise differences between y and $\tilde{y}$ decrease. Globally, we quantified similarity between images through the structural similarity index

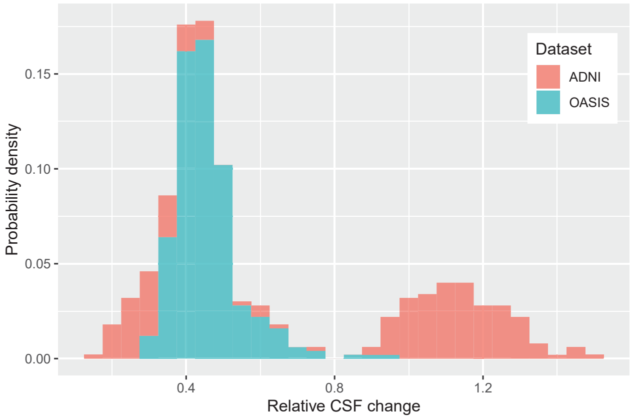


Figure 5.4: Distribution of relative cerebrospinal fluid enlargement among pairs of baseline and follow-up volumes on the OASIS and ADNI datasets. Of note, these values may be affected by skull stripping results. CSF: cerebrospinal fluid.

(SSIM) [350] as it has been found correlated with the quality of perception of the human visual system [351] and accounts jointly for variations in luminance, contrast, and structure (correlation):

$$SSIM(y,\tilde{y}) = \underbrace{\frac{2\mu_y\mu_{\tilde{y}} + c_1}{\mu_y^2\mu_{\tilde{y}}^2 + c_1}}_{\text{Luminance}} \cdot \underbrace{\frac{2\sigma_y\sigma_{\tilde{y}} + c_2}{\sigma_y^2\sigma_{\tilde{y}}^2 + c_2}}_{\text{Contrast}} \cdot \underbrace{\frac{\operatorname{cov}(y,\tilde{y}) + c_3}{\sigma_y\sigma_{\tilde{y}} + c_3}}_{\text{Structure}}, \tag{5.4}$$

where μ and σ denote the mean and standard deviation values of the luminance of the images, $\operatorname{cov}(y,\tilde{y})$ the covariance between y and $\tilde{y}$, and c_i constants to avoid a null denominator [351]. The SSIM values range within zero and one, where the former indicates null similarity while the latter implies that y and $\tilde{y}$ are equal. We expected our framework to produce synthetic scans of such perceptual quality that MAE and SSIM values tended to zero and one, respectively.

Segmentation agreement

Segmentation-based atrophy quantification algorithms segment brain tissues and measure volumetric differences [70, 98] or brain boundary shifts [60, 71–73]. This situation requires our framework to produce synthetic brain scans in which tissue contrast is good enough for algorithms to detect grey matter, white matter, and cerebrospinal fluid. For that, we segmented brain tissues in both real and generated scans using FAST [57] and measured their agreement using the Dice similarity coefficient (DSC) [323, 324]. With the DSC, we determine the extent of overlap between a given segmentation and the ground truth. Given binary tissue segmentation masks for a real scan, m_{CSF}, m_{GM}, and m_{WM}, and those for its corresponding approximation, $\tilde{m}_{CSF}$, $\tilde{m}_{GM}$, and $\tilde{m}_{WM}$, the DSC is mathematically expressed as

$$DSC_{ROI}\left(s, \tilde{s}\right) = 2\ \frac{\sum_{v \in ROI} m_v \cdot \tilde{m}_v}{\sum_{v \in ROI} m_v \cdot \sum_{v \in ROI} \tilde{m}_v}. \tag{5.5}$$

The values for DSC range from zero to one, where zero indicates null similarity between segmentation masks and one exact agreement. We expected our framework to produce synthetic scans such that their segmentations are comparable to those used for generating them in the first place, i.e. DSC values close to one.

Cerebral atrophy

As the ultimate goal of our generation framework is to predict the appearance of a baseline T1-w scan after being altered as requested, we studied whether induced variations matched the request. We considered two atrophy quantification methods for assessing this aspect: SIENA [60] and the Jacobian determinant integration method [71] – segmentation-based and registration-based methods, respectively. Once our model deformed the baseline scan according to the input probability maps, we used these two tools to quantify potential atrophy variations between the generated and real scans. Ideally, the percentage of whole-brain volume change (PBVC) yielded by SIENA and the integral of Jacobian determinants yielded by the Jacobian method should be close to zero and one, respectively. Since these two methods address atrophy quantification from two different perspectives, they allow us to verify whether tissue variations were induced effectively and whether brain boundaries were well-defined.

Statistical differences

We used the Wilcoxon signed-rank test to assess statistical significance of differences among methods. We considered p-values below 0.01 to be statistically significant.

5.3.3 Architectural directives and loss functions

The first experiment compared the generation quality of four strategies: two of them inspired by state-of-the-art data generation strategies and our network optimised with two different loss functions. Some details as follows:

- **3D CGAN - MSE**: A network inspired by the work of Shin *et al.* [191], consisting of a U-Net generating three brain regions and a discriminator determining whether the generated scan is realistic enough or not. We optimised

Table 5.2: Generation quality scores obtained with four different strategies. The results were obtained from training on O_2 and testing on O_1 ($O_2 \rightarrow O_1$) and vice versa ($O_1 \rightarrow O_2$). The variable n represents the number of test cases. The values in bold are significantly higher (p-value < 0.01) than the ones yielded by the other three approaches. MAE: median absolute error. SSIM: structural similarity. DSC: Dice similarity coefficient. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter. PBVC: percentage of brain volume change. CGAN: conditional generative adversarial network. MSE: mean square error. RWMSE: region-wise mean square error.

	n	**Approach**	**Intensity**		**Segmentation**			**Atrophy**	
			MAE	**SSIM**	**DSC - CSF**	**DSC - GM**	**DSC - WM**	**PBVC**	**Jacobian Int**
$O2 \rightarrow O1$	126	3D CGAN - MSE	0.03 ± 0.01	0.95 ± 0.02	0.83 ± 0.16	0.69 ± 0.21	0.78 ± 0.20	2.19 ± 5.70	$\mathbf{0.99 \pm 0.01}$
		Baseline - MSE	0.08 ± 0.04	0.90 ± 0.04	0.92 ± 0.02	0.87 ± 0.03	0.90 ± 0.02	2.60 ± 1.01	1.13 ± 0.06
		Cascaded - MSE	0.05 ± 0.03	0.96 ± 0.01	0.93 ± 0.02	0.87 ± 0.05	0.91 ± 0.04	0.33 ± 0.25	1.16 ± 0.06
		Cascaded - RWMSE	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.95 \pm 0.02}$	0.27 ± 0.16	1.14 ± 0.05
$O1 \rightarrow O2$	169	3D CGAN - MSE	0.03 ± 0.01	0.95 ± 0.01	0.80 ± 0.15	0.71 ± 0.20	0.79 ± 0.21	2.19 ± 5.70	$\mathbf{1.00 \pm 0.02}$
		Baseline - MSE	0.15 ± 0.05	0.87 ± 0.08	0.92 ± 0.03	0.88 ± 0.02	0.91 ± 0.01	2.73 ± 0.43	1.12 ± 0.05
		Cascaded - MSE	0.11 ± 0.03	0.95 ± 0.02	0.94 ± 0.02	0.90 ± 0.02	0.91 ± 0.03	0.21 ± 0.16	1.14 ± 0.05
		Cascaded - RWMSE	$\mathbf{0.01 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.95 \pm 0.01}$	0.19 ± 0.14	1.14 ± 0.05

such networks using the mean square error (generator) and the categorical cross-entropy (discriminator).

- **Baseline U-Net - MSE**: A network inspired by the work of Chartsias *et al.* [169] and Salem *et al.* [170], consisting of three parallel U-Nets generating three brain regions separately and a final addition module to merging them into a single T1-w scan. Each u-shaped subnetwork resembles the design illustrated in Fig. 3 in [170]. We optimised such a network using a mean square error loss as in the original papers.

- **Cascaded U-Nets - MSE**: Our proposed network, as depicted in Fig. 5.2, consisting of three parallel U-Nets generating three brain regions separately and a final U-Net merging them into a single T1-w scan. We optimised this network using a mean square error loss.

- **Cascaded U-Nets - RWMSE**: Our proposed network, as above, optimised using our proposed region-wise mean square error, described in Eq. 5.1.

We implemented the aforementioned strategies and compared their generation quality. We provided the networks with baseline volumes and actual follow-up tissue segmentation probability maps and evaluated the similarity between the actual follow-up and the approximated one. We trained all networks using the same scheme, i.e. same optimiser, training data, training stopping policy, and machine. Data were taken from the O_2 collection and tested on the O_1 set and vice versa. The results of this trial are presented in Table 5.2.

The cascaded U-Net trained with the mean square error loss performed significantly better than its baseline in most cases ($n = 295$; Cascaded-MSE vs Baseline-MSE, p-value; MAE: 0.08 ± 0.04 vs 0.12 ± 0.06, $p < 0.01$; SSIM: 0.95 ± 0.02 vs 0.88 ± 0.07, $p < 0.01$; DSC-CSF: 0.94 ± 0.02 vs 0.92 ± 0.03, $p < 0.01$; DSC-GM: 0.89 ± 0.04 vs 0.88 ± 0.03, $p < 0.01$; PBVC: 0.26 ± 0.21 vs 2.68 ± 0.74, $p < 0.01$), except in terms of the segmentation of white matter, where they both obtained similar Dice scores (DSC-WM: 0.91 ± 0.04 vs 0.91 ± 0.02, $p = 1.00$), and Jacobian integral, where the latter outperformed the former (Jacobian Int: 1.15 ± 0.06 vs 1.12 ± 0.6, $p < 0.01$). This outcome suggested that scans generated with the cascaded U-Nets trained with the mean square error appear more similar to real follow-up acquisitions and exhibit better tissue contrast than those generated with the baseline U-Net, but brain edges are more blurred.

The use of the region-wise mean square error resulted in significantly improved performance compared to that obtained using the original mean square error ($n = 295$; Cascaded-RWMSE vs Cascaded-MSE, p-value; MAE: 0.01 ± 0.01 vs 0.08 ± 0.04, $p < 0.01$; SSIM: 0.99 ± 0.01 vs 0.95 ± 0.02, $p < 0.01$; DSC-CSF: 0.96 ± 0.02 vs 0.94 ± 0.02, $p < 0.01$; DSC-GM: 0.94 ± 0.03 vs 0.89 ± 0.04, $p < 0.01$; DSC-WM: 0.95 ± 0.01 vs 0.91 ± 0.04, $p < 0.01$; PBVC: 0.22 ± 0.15 vs 0.26 ± 0.21, $p < 0.01$), except for the Jacobian integral, where the difference between their scores was not significant (Jacobian Int: 1.14 ± 0.05 vs 1.15 ± 0.06, $p > 0.01$). These results suggest that the proposed loss function allows the network to generate more faithful reconstructions versus the accustomed loss. However, the proposed loss did not seem to help to sharpen brain edges.

Notably, the image-to-image translation conditional adversarial network inspired by the work of Isola *et al.* [343] and Shin *et al.* [191] obtained Jacobian integration values close to one (1.00 ± 0.02), i.e. brain edges were delineated almost perfectly according to this metric. In this regard, this network outperformed all other approaches significantly ($p < 0.01$). Nevertheless, its performance according to the rest of the metrics was significantly lower than our cascaded U-Net trained with the region-wise mean square loss function ($n = 295$; Cascaded-RWMSE vs 3D CGAN-MSE, p-value; MAE: 0.01 ± 0.01 vs 0.03 ± 0.01, $p < 0.01$; SSIM: 0.99 ± 0.01 vs 0.95 ± 0.01, $p < 0.01$; DSC-CSF: 0.96 ± 0.02 vs 0.81 ± 0.15, $p < 0.01$; DSC-GM: 0.94 ± 0.03 vs 0.70 ± 0.20, $p < 0.01$; DSC-WM: 0.95 ± 0.01 vs 0.79 ± 0.21, $p < 0.01$; PBVC: 0.22 ± 0.15 vs 2.19 ± 5.70, $p < 0.01$). Compared to the rest of the models, scans generated using the adversarial model presented lower tissue contrast that prevented them from being segmented properly.

An example of generated scans using the five strategies is presented in Fig. 5.5. We displayed the generation on the case 157 of the OASIS2 dataset as it exhibited the maximum relative CSF change in this dataset and, thus, generation issues were visually evident. Qualitatively speaking, literature inspired strategies did not lead to

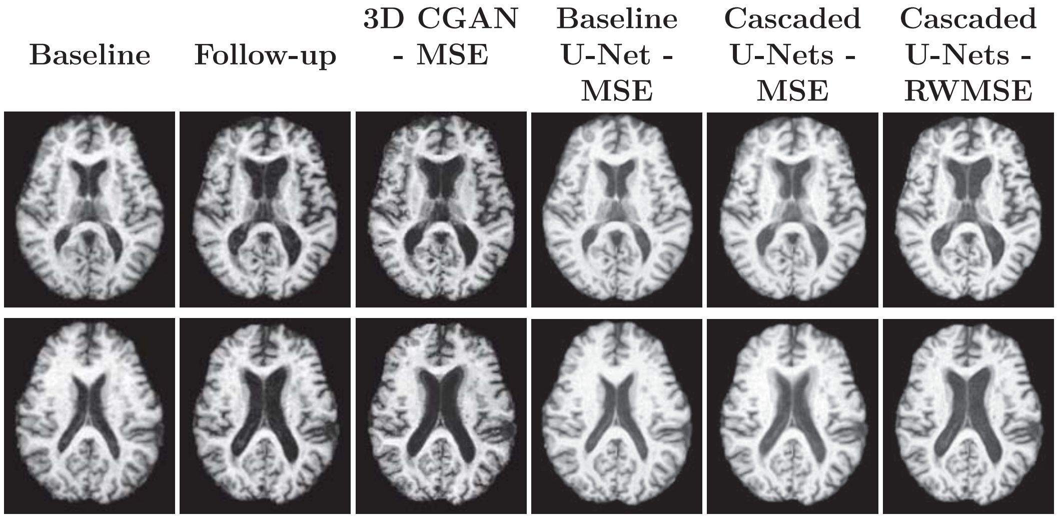


Figure 5.5: Example of scans generated with different architectures and loss functions. The first and second column correspond to the real baseline and follow-up scans. From the third to the sixth column, scans generated with the conditional generative adversarial network trained using the mean square error loss, with the baseline U-Net trained using the mean square error loss, with our proposed design trained with a mean square error, and with our proposed architecture optimised with our region-wise mean square error. CGAN: conditional generative adversarial network. MSE: mean square error. RWMSE: region-wise mean square error.

appealing results. The conditional adversarial network generated scans with sharp yet noisy edges and inaccuracies in the lateral ventricles that appear as if the model laid the ground truth over the baseline and failed at amalgamating intensities accurately. The baseline U-Net learnt identity mapping as the only visual differences are in terms of the noise, reduced in synthetic scans. Scans generated using our cascaded U-Nets trained with the mean square error exhibited artefacts; the reconstructions provided by each branch seem to be merged in an uncoordinated way as tissues seem superimposed. On the contrary, both axial slices generated using our proposed RWMSE loss function appear similar to the expected follow-up scan as tissues were altered as expected. Our proposal reduced speckle noise and delineated better some structures (e.g. sub-cortical structures) compared to the real follow-up scans, i.e. the contrast of the image was enhanced.

Taking the aforementioned information into account, our proposed cascaded U-Net model optimised with the region-wise loss function evidenced improved performance both qualitatively and quantitatively. Henceforth, we computed our results using such a model.

Table 5.3: Comparison between generated and actual volumes concerning intensity, segmentation, and atrophy dissimilarities. The column n shows the cardinality of test set. The segmentation scores correspond to the DSC values computed using FAST masks. MAE: median absolute error. SSIM: structural similarity. DSC: Dice similarity coefficient. CSF: cerebrospinal fluid. GM: grey matter. WM: white matter. PBVC: percentage of brain volume change.

Train → Test	**n**	**Intensity**		**Segmentation**			**Atrophy**	
		MAE	**SSIM**	**DSC - CSF**	**DSC - GM**	**DSC - WM**	**PBVC**	**Jacobian Int**
$O2 \rightarrow O1$	169	0.02 ± 0.01	0.99 ± 0.01	0.96 ± 0.01	0.94 ± 0.03	0.95 ± 0.02	0.27 ± 0.16	1.14 ± 0.05
$O1 \rightarrow O2$	126	0.01 ± 0.01	0.99 ± 0.01	0.96 ± 0.02	0.94 ± 0.02	0.95 ± 0.01	0.19 ± 0.14	1.14 ± 0.05
$A2 \rightarrow A1$	153	0.03 ± 0.03	0.97 ± 0.02	0.95 ± 0.02	0.92 ± 0.02	0.96 ± 0.01	0.24 ± 0.18	1.11 ± 0.05
$A1 \rightarrow A2$	136	0.04 ± 0.03	0.98 ± 0.01	0.95 ± 0.02	0.93 ± 0.03	0.96 ± 0.01	0.23 ± 0.16	1.13 ± 0.03
$OASIS \rightarrow ADNI$	289	0.03 ± 0.01	0.97 ± 0.02	0.96 ± 0.01	0.94 ± 0.03	0.96 ± 0.01	0.15 ± 0.15	1.12 ± 0.04
$ADNI \rightarrow OASIS$	295	0.02 ± 0.01	0.97 ± 0.01	0.96 ± 0.01	0.94 ± 0.02	0.95 ± 0.02	0.22 ± 0.35	1.15 ± 0.06

5.3.4 Generation quality (same dataset)

We ran a second experiment to evaluate the quality of the generation of our tool. We assessed generation when synthesising a scan from a baseline, i.e. we provide the network with a baseline T1-w volume and three tissue probability maps of the corresponding follow-up T1-w acquisition.

The results obtained by our proposal on the considered datasets are displayed in Table 5.3. Our model generated volumes that were quantitatively similar to the actual follow-up scans at intensity, segmentation and atrophy levels. Regarding intensity, our method yielded MAE values below 0.11 and SSIM values above 0.90. Concerning segmentation, our tool produced images with tissue masks comparable to the ones of the actual volumes as all DSC values are above 0.80. Nevertheless, the obtained segmentation errors were within the FAST accuracy and reproducibility ranges [334]. Regarding the volume change detected by atrophy quantification algorithms, our method reported low values overall and within reproducibility rates [11].

Our method yielded better results intensity-wise on the OASIS set than on the ADNI one. This might be a consequence of increased lousy skull stripping of ROBEX on the latter set in comparison to the former. If a synthetic scan is compared to a follow-up volume which skull has not been entirely removed, the scores for MAE and SSIM will be lower than when non-brain areas have been completely masked out. The error might also be caused by the atrophy levels in ADNI, which are higher than in OASIS.

5.3.5 Generation quality (cross-dataset)

The third experiment consisted of evaluating the performance of the proposal when training on a certain dataset and testing on a different one (OASIS→ADNI and

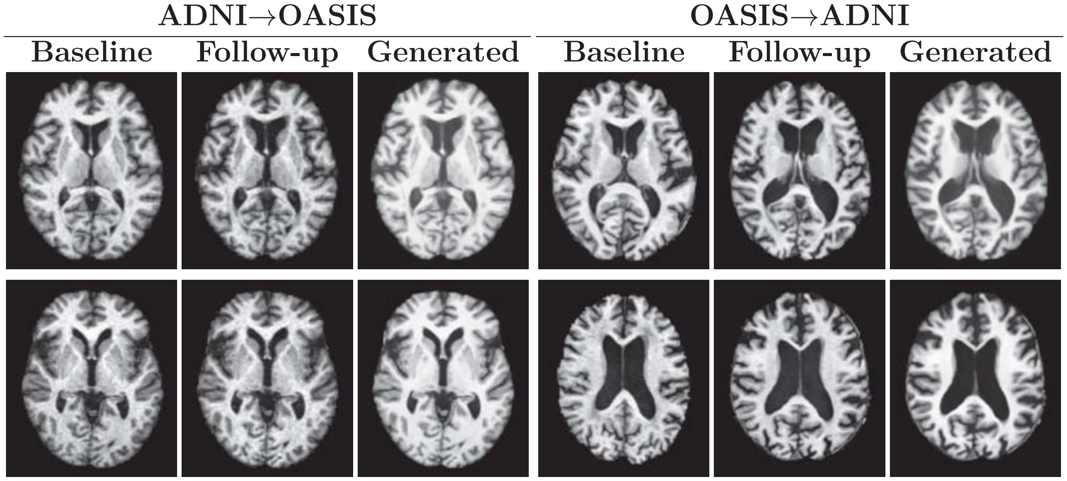


Figure 5.6: Cross-dataset generation examples: OASIS follow-up scans using a network trained on ADNI (left) and ADNI follow-up scans using a network trained on OASIS (right).

ADNI→OASIS). The results are shown in Table 5.3 and two cases depicted in Fig. 5.6. The generation per se did not seem significantly affected as none of the intensity, segmentation, or atrophy values differed significantly from the performance measurements obtained when training and testing on the same dataset. This outcome makes our proposal appealing as it shows that by pre-processing the incoming data (e.g. harmonisation by registering to a common space and matching intensity histograms), the network might be used in a different domain without requiring retraining.

5.3.6 Assessing induced changes with volumetry methods

The fourth experiment consisted of exploring whether induced tissue variations could be detected by atrophy quantification algorithms. We created the dataset as follows. Initially, we selected ADNI subjects which exhibited the maximum atrophy over time. The atrophy was measured as relative enlargement of the CSF region. We computed the deformation field between the baseline and latest follow-up scans. Then, we multiplied the resulting deformation vectors by scalars between zero and one to obtain intermediate scans. We considered five scalars: 0%, 25%, 50%, 75%, and 100%. Of note, this is an approximation to the pathological process as we would assume that atrophy change varies spatially at the same time in all directions. Afterwards, we ran FAST on the baseline scans to segment each tissue and altered the resulting tissue probability maps using the various deformation fields. Finally, we input each pair of baseline volume and modified tissue maps to generate a synthetic scan. In total, we generated 216 synthetic scans.

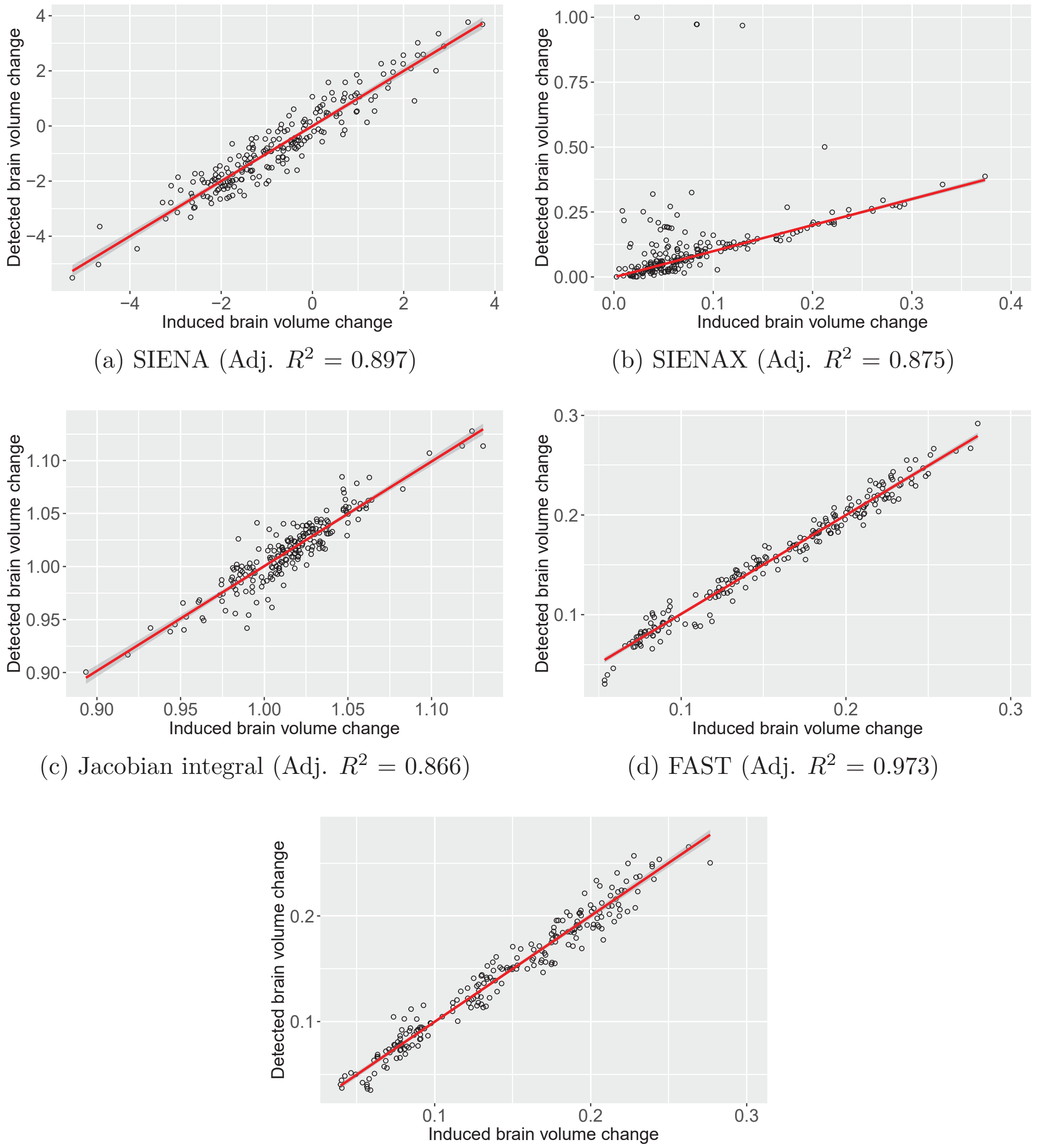
We evaluated the capacity of our framework to generate detectable tissue variations using four methods: three atrophy quantification algorithms, SIENA, SIENAX, and the Jacobian determinant integration method, and two tissue segmentation algorithms, FAST and SPM. We computed a robust multiple linear regression model [352] using relative absolute volumetric differences, tissue-wise average symmetric surface changes [353], and Dice coefficients (as surrogate measures for tissue displacement) as predictor variables and detected or observed brain volume change as a response variable. The results are shown in Fig. 5.7. Overall, our induced tissue variations correlated well with the detected volume change (adjusted correlation coefficient R^2 above 0.86). For SPM, the linear model was close to $x = y$ as $R^2 \approx 1$ and y-intercept ≈ 0. This outcome implies that the induced tissue variations were detected correctly by conventional cross-sectional and longitudinal atrophy quantification tools.

5.4 Discussion

In this chapter, we proposed a CNN-based framework for creating longitudinal evaluation environments given a set of T1-w baseline scans and follow-up tissue probability maps. Our pipeline contemplates four stages: preprocessing, data preparation, generation and postprocessing. Initially, we skull-stripped, intensity corrected and registered all volumes to a common space. Then, we tiled up the baseline and altered tissue probability maps into overlapping blocks and passed them through our network, a cascaded u-shaped network. Finally, once all blocks had been processed, we reconstructed and intensity corrected the resulting synthetic volume.

The network consisted of four processing modules: one dedicated to generating changes on each class (namely, CSF, GM, and WM) and the last one in charge of fusing them. We optimised all components end-to-end using a region-aware multi-objective loss function. We followed state-of-the-art design patterns to devise our network. Overall, the devised framework showed being capable of producing accurate synthetic scans in terms of intensity, segmentation and tissue volume similarity. The proposal was assessed through four experiments exploring architecture directives and loss functions, generation quality when training the network on a particular dataset and testing on the same or different one, and the ability of our framework to generate acceptable and detectable changes.

The first experiment compared our proposal against literature-inspired networks and tested the efficacy of the proposed region-wise loss function versus the conventional mean square error loss. The two literature-inspired networks were based on the work on latent space representations using U-Nets of Chartsias *et al.* [169] and



(a) SIENA (Adj. $R^2 = 0.897$)

(b) SIENAX (Adj. $R^2 = 0.875$)

(c) Jacobian integral (Adj. $R^2 = 0.866$)

(d) FAST (Adj. $R^2 = 0.973$)

(e) SPM (Adj. $R^2 = 0.957$)

Figure 5.7: Real versus fitted values obtained using robust linear regression models for the five different methods. The models were built using the average symmetric surface distance, the relative absolute volumetric difference, and Dice coefficients between original and deformed tissue maps as predictor variables and detected volume change as response variable. Data points and regression lines are represented by empty circles and red lines, respectively.

Salem *et al.* [170] and the label-to-image translation conditional generative adversarial network described by Shin *et al.* [191]. We generated 295 follow-up scans using baseline data and real follow-up tissue segmentation maps. We observed that our proposed network outperformed both state-of-the-art approximations and that the region-wise mean square error objective function led to superior performance.

The second experiment gauged the capacity of our proposal to generate synthetic follow-up scans which were similar to the actual images when training and testing on the same domain. The similarity was evaluated regarding intensity using MAE and SSIM, tissue segmentation mask overlap using FAST and DSC, and atrophy change using SIENA and the Jacobian integration method. The experiment considered four collections: two from the OASIS and two from the ADNI. In all of them, our proposal yielded high similarity scores. We observed that skull stripping errors resulted in increased dissimilarity scores, as indicated previously in the literature [62]. Nonetheless, all the values were within the reproducibility ranges reported in the literature.

The third experiment explored whether the framework could be used in unseen and different data collections. We trained our network on a particular selection (e.g. OASIS2) and tested on another one (e.g. ADNI) and vice versa to determine how robust was the entire framework to these sort of variations. Our preliminary results showed that our framework may cope with this situation without affecting its performance considerably and without requiring additional adjustments, but further testing in this regard is needed. Evidently, this outcome is appealing as our ultimate goal is to apply our pipeline to datasets with possibly varying acquisition parameters.

The fourth experiment examined whether conventional tools detected synthetically induced changes. This is key in this research as our primary goal is to create high-quality synthetic scans for which tissue variations (loss) with respect to the baseline scans are known beforehand. We used real tissue displacement vectors to alter baseline segmentation masks, input them into our framework, and gauged changes using SIENA, SIENAX, the Jacobian integration method, SPM, and FAST. All changes detected by these five tools highly correlate with our induced changes (Adj. R^2 values above 0.86), showing common tissue segmentation and volumetry methods can detect brain alterations generated by our proposal. Note that even algorithms that were not used at any point within our framework correlated with the induced changes.

A direct and fair comparison with other works in the area is not straightforward as inputs and generation mechanisms vary. For example, in [335], the tool is provided with an MR scan and a number indicating the desired level of expected tissue loss and the tool outputs another scan in which the brain volume has been altered to match the requested value. The deformation of the volume follows the topology of the brain rather than a pathology-oriented pattern per se. Khanal *et al.* [339] proposed a tool for prescribing local atrophy changes given segmentation and atrophy maps, in which the user indicates modifiable and not modifiable regions and the expected degree of atrophy, respectively. We did not compare to their work since we would need to build both maps appropriately and accurately. Thus, we compared our proposal against networks inspired by previous works on image and lesion synthesis [169,170] and data augmentation [191] since their code was either publicly available and/or used established and well-known strategies.

The motivation behind our proposal is two-fold. First, we aim to generate controlled environments to evaluate atrophy quantification strategies. Following the urgent challenges in GM atrophy measurement exposed by Amiri *et al.* [13], pipelines could be compared under the same settings, and their pros and cons could be adequately analysed using our tool. This would be a way to extend the clinical validation of existing tools. Second, we target using the deep learning power to craft a more precise and accurate method for measuring tissue loss. As it is well-known in the literature, deep learning has outperformed traditional machine learning methods in scenarios where lots of data are available. Thus, we could train networks to achieve improved measurements using our tool.

Our proposal exhibits limitations regarding segmentation, model assumptions, domain dependence, and bias. First, it is well-known that the segmentation performance of FAST in basal ganglia is not accurate enough. Although we did not observed problems in this regard (see Fig. 5.5 and Fig. 5.6), better segmentation strategies need to be considered. Second, unlike model-based proposals [335], there are no assumptions on how tissues are altered to match the input segmentation maps. On the one hand, this favours the flexibility of the generation scheme. On the other hand, it does not follow a specific pathology-oriented deformation strategy. Third, the core network may produce undesired outcomes when the intensity range of an input scan differs considerably from the training intensity interval. Nonetheless, this issue was mitigated by performing intensity standardisation and registering input scans to the training space. Fourth, the current strategy for generating controlled environments requires image segmentation and registration, i.e. generation is biased towards them. Nonetheless, we observed that our method could generate tissue changes that were highly correlated with SPM, a method that was not considered in the training pipeline.

5.5 Summary

The lack of a ground truth prevents testing the accuracy of longitudinal atrophy quantification methods and training deep learning methods to perform these types of assessments. In this chapter, we proposed a deep learning framework to generate controlled evaluation environments as a way for addressing this problem. Our framework deformed T1-w brain MRI scans as requested through segmentation maps. Experimental results showed that our framework could produce synthetic scans that resemble real ones, that induced changes highly correlated with measurements detected by validated cross-sectional and longitudinal segmentation algorithms, and that its performance was significantly superior to that of two literature-inspired works overall. Moreover, our experiments on harmonised datasets evince the potential of our framework to be applied to various data collections without further adjustment.

In the next chapter, we show an application of our generation framework as data augmentation strategy for training a deep learning based method for assessing longitudinal cerebral atrophy and determining its suitability in patients with multiple sclerosis and dementia.

Chapter 6

Deep learning for quantifying longitudinal brain atrophy

In this chapter, we present a deep learning based framework for quantifying longitudinal brain atrophy from baseline and follow-up T1-w brain magnetic resonance imaging scans using the longitudinal data generation approach proposed in Chapter 5.

ARTICLE IN PREPARATION. EMBARGO UNTIL PUBLICATION DATE

Chapter 7

Conclusions

7.1 Thesis summary

Cerebral brain atrophy corresponds to the loss of neurons and their connections as a result of the ageing process or brain pathologies, such as dementia and multiple sclerosis. Therefore, their accurate quantification may help medical doctors to diagnose these brain diseases more timely, monitor their progression, and evaluate the effectiveness of novel treatments.

Qualitative and quantitative methods were developed prior to the deep learning era. Qualitative clinical ratings were devised to rate cerebral atrophy into normal or abnormal based on prior anatomical knowledge. However, the rating process is subjective, requires serious training, may not generalise to various cohorts, and is limited by its discrete nature. Quantitative methods reduced the subjectivity of the assessment. Nonetheless, their performance was compromised by their assumptions, the craftsmanship of the image analyst at engineering the methods, the presence of brain lesions or unseen cases, and extrinsic and intrinsic imaging variations. Due to the astonishing performance of deep learning in applications involving image processing, we questioned whether deep learning could be used for evaluating brain tissue volumes at cross-sectional levels and their variations over time at longitudinal levels using brain MRI scans from both healthy subjects and patients.

To accomplish our primary goal, we reviewed relevant works on deep learning for brain medical image analysis to get familiar with the topic at hand; understand the needs from the medical perspective; study trends, processing pipelines, and applications for which deep learning had been used; analyse their potential strengths and pitfalls; and comprehend general challenges that needed to be addressed in the field. We noticed that more than two thirds of the works targeted either segmentation or classification, potentially due to the availability of popular competitions in these regards. On the one hand, these types of events permit demonstrating how capable – or not – is each method when evaluated under the same evaluation framework (dataset, metrics, task). On the other hand, a top-performing method may have been overfit to a certain evaluation framework and data and, hence, it might not be able to generalise well to unseen data or may involve a series of steps limiting or making them impractical in real-life scenarios. Moreover, these types of challenges does not allow determining whether certain design patterns are more beneficial than others.

In light of the aforementioned limitations, we benchmarked applicable deep learning methods for brain tissue segmentation using not only the same evaluation framework (dataset, metrics, and tasks) but also the same processing pipeline (i.e. pre-processing, data preparation, segmentation, and post-processing). In such a way, we ensured all models were being evaluated under similar yet favourable conditions. Moreover, this setup allowed us to compare deep learning architectures themselves and determine whether specific design directives explained their performance. Overall, we observed that deep learning could be used for cross-sectional tissue segmentation and could produce state-of-the-art results compared to traditional computational methods when trained and tested on the same domain. Further development in regards to transfer learning and domain adaptation may enable their use in cross-dataset scenarios and, perhaps more importantly, in clinical scenarios. Other relevant findings were:

- Overlapping patch sampling seemed beneficial as they allow extracting more training cases and enforcing spatial consistency.
- Networks performed their best when provided with various modalities due to the complementary data that various information sources supply.
- The larger the patch size, the more the contextual information the network uses and, hence, the better the segmentation performance.
- Architectures exploiting the 3D nature of the MRI scans tend to outperform their 2D analogues. However, differences may not be statistically significant. Also, the former group of networks is less resilient to heterogeneity in voxel spacing than the latter.
- The number of trainable parameters does not automatically translate into segmentation performance, but it does explain processing times as networks with more parameters take more time during training and testing.
- U-Nets are suitable for brain MRI tissue segmentation due to their balance between performance and processing speed.

Based on the findings of our quantitative comparison, we developed three different deep learning based methods for segmenting brain tissues cross-sectionally in three international challenges: iSeg2017, MRBrainS18, and iSeg2019. In these three events, we processed scans from six-month-old babies acquired with the same and varied scanning protocols and adult brains with brain lesions (white matter hyperintensities and/or brain infarcts). Our proposals incorporated densely connected networks (iSeg2017), ensemble of U-Nets using tissue segmentation priors computed with validated tissue segmentation tools (MRBrainS18), and hybrid approximation leveraging U-Nets and multi-atlas segmentation (iSeg2019). In all three challenges, we achieved compelling performance, suggesting that our benchmark was useful for creating appropriate models for this particular task.

In contrast to the almost ideal scenario for cross-sectional tissue segmentation, longitudinal brain atrophy quantification proposals are difficult to devise and validate due to the lack of publicly available datasets with ground truth. This issue prevents assessing the accuracy of any development (deep learning based or not) and training any deep learning approach to specifically detect brain changes over time. To overcome this limitation, we proposed a deep learning based method for generating longitudinal datasets by deforming a baseline T1-w scan as requested through segmentation maps. Our proposal incorporated a cascaded three-path U-Net which synthesised a rough generation (first U-Net) and a refined one (second U-Net). We optimised our model using a multi-objective mean square error loss function to force each path to produce a cerebrospinal fluid, grey matter, and white matter accurately. We tested the capacity of our framework to synthesise scans that looked realistic and to produce changes that were detected by validated tissue segmentation and atrophy quantification tools. First, we provided our model with baseline scans and real follow-up segmentation maps from two longitudinal datasets of patients with Alzheimer's disease and dementia and observed that our framework could produce synthetic follow-up scans that matched the real ones regarding their perceptual properties (luminance, contrast, and structure), segmentation, and cerebral atrophy. Our preliminary cross-dataset results suggest that the performance was consistent even when training and testing on different but harmonised datasets. Compared to two relevant works generating brain lesions using U-Nets and conditional generative adversarial networks, our proposal outperformed them significantly in most cases, except in the delineation of brain edges where the generative network took the lead. Second, we examined whether changes induced with our framework were detected by FAST, SPM, SIENAX, SIENA, and the Jacobian integration method. In all cases, we noticed that induced and detected changes were highly correlated.

We used our generation framework as data augmentation strategy for training a deep learning based method for longitudinal atrophy quantification. Our proposal incorporated recent advances in deep learning for non-linear registration (namely, VoxelMorph) to produce deformation fields and the Jacobian integration method to convert deformation vectors into measures of atrophy. We compared our proposal against validated tools: SIENA and the Jacobian integration method using ANTs for non-linear registration. Moreover, we compared our work against a pre-trained VoxelMorph and a model trained on the original training dataset to show the effect of training on harmonised data and using our data augmentation proposal. We consider evaluating brain changes around edges between brain regions where atrophy is likely to happen and not on grey matter as accustomed in the literature. This approach allowed us to measure changes in both healthy and pathological regions of the brain simultaneously. We assessed the suitability of our proposal based on the scan-rescan error and its ability to produce different atrophy measurements for subjects ongoing different pathologies. In scan-rescan assessments, we observed that the error yielded by our proposal was comparable to those of ANTs and the pretrained VoxelMorph and lower than that SIENA. In longitudinal assessments, we noted that of SIENA discerned better between Alzheimer's disease patients vs control and dementia patients vs control, but our proposal was particularly suitable for multiple sclerosis vs control subjects. From our analysis, we concluded that deep learning shows promising results for longitudinal atrophy quantification, but further testing needs to be carried out to determine a suitable approximation for demented patients.

7.1.1 Contributions

The aim of this thesis was to develop deep learning methods for quantifying brain atrophy, at cross-sectional and longitudinal levels from brain MRI, to provide medical doctors with accurate and precise cerebral atrophy measurements. This work has the potential to shed light into the relationship between brain atrophy and neurological diseases, monitoring disease progression, and assessing treatment effectiveness.

The principal contributions of this thesis for both the scientific and medical community are:

- An extensive review of deep learning based strategies for brain medical image analysis in which we discussed common architectures, tasks in which they have been applied, common processing pipelines, and current challenges that need to be addressed in the field.
    - **Bernal, J.**, Kushibar, K., Clèrigues, A., Oliver, A., & Lladó, X. (2020). Deep learning for medical imaging. In: Bacciu D., Lisboa P.J.G., Vellido A., eds. *Deep Learning in Biology and Medicine.* Singapore: World Scientific Publishing. Under review.
    - **Bernal, J.**, Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., & Lladó, X. (2019). Deep convolutional neural networks for brain

image analysis on magnetic resonance imaging: a review. *Artificial intelligence in medicine*, 95, 64-81. Quality index: [JCR IF 3.574, Q1(5/26)].

- A quantitative comparison of unsupervised and supervised state-of-the-art for tissue segmentation in brain MRI to establish a direct comparison between them and understand their experimental strengths and weaknesses based on their segmentation performance on three relevant datasets (IBSR18, MICCAI2012, and iSeg2019).

  - **Bernal, J.**, Kushibar, K., Cabezas, M., Valverde, S., Oliver, A., & Lladó, X. (2019). Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access*, 7, 89986-90002. Quality index: [JCR IF 4.098, Q1(23/155)]

- Three deep learning based methods developed based on our experimental findings to segment brain tissues cross-sectionally in six-month-old subjects and diabetic, demented, and control adult subjects as part of Grand Challenges of the Medical Image Computing and Computer Assisted Intervention Conference 2017-2019.

- A fully automatic deep learning based framework for generating controlled longitudinal cerebral atrophy on brain MRI to cope with the lack of ground truth and a dataset containing 216 scans for evaluating longitudinal cerebral atrophy in brain MRI.

  - **Bernal, J.**, Valverde, S., Kushibar, K., Oliver, A., & Lladó, X. (2019). Generating controlled atrophy change evaluation environments on brain MR using convolutional neural networks and segmentation priors. Under review in Neuroinformatics. Quality index: [JCR IF 5.127, Q1(9/106)]

- A fully automatic deep learning based framework for quantifying longitudinal cerebral atrophy in healthy subjects and multiple sclerosis and dementia patients.

  - **Bernal, J.**, Oliver, A., & Lladó, X. (2020). Deep learning for quantifying longitudinal cerebral atrophy in brain magnetic resonance imaging. To be submitted.

- A prototype of a toolbox for segmenting brain tissues using state-of-the-art deep learning methods, generating longitudinal cerebral atrophy datasets, and quantifying longitudinal cerebral atrophy, partially available in our GitHub repository `github.com/NIC-VICOROB`.

7.2 Future work

The cross-sectional and longitudinal quantification of cerebral atrophy using deep learning is a complex topic involving multiple aspects and many research lines. This thesis is a clear example of the multidisciplinary mixture of medicine, medical physics, computer science, computer vision, statistics, among other disciplines, that are involved in this process. In this thesis, we showed how deep learning could be applied to segment brain tissues in babies and adults and in healthy and unhealthy subjects, and to quantify longitudinal brain atrophy in both control subjects and Alzheimer's disease, dementia, and multiple sclerosis patients. Nonetheless, these strategies can be extrapolated to other tasks taking advantage of MRI and other pathologies/treatments.

In this section, we discuss those aspects concerning the work presented in this work that need to be addressed in short-term and relevant challenges and research lines that have not been investigated during this thesis and could be further explored.

7.2.1 Short-term proposal improvements

In Chapter 4, we did not study the performance of these networks when trained on a specific domain and tested on another one. Due to the substantial heterogeneity presented in the datasets and the well-known domain shift problem of deep learning methods caused by variations during acquisition (e.g. different scanner and imaging protocol), we do not expect them to perform well. However, transfer learning or domain adaptation strategies could be explored for adapting cross-sectional tissue segmentation methods to other unseen domains. We think that this is a vital step that should be carried out in the future to evaluate the deployability of deep learning solutions to clinical practice and determine whether certain architectures adapt better than others.

In Chapter 5, we mentioned that our approach for inducing tissue variations does not reflect the atrophy patterns that brain diseases exhibit. In the future, we plan to use deep learning to learn pathology specific tissue deformations using a conditional generative network and add this module to our framework. This will open the doors to develop novel tools that can be later used in investigating atrophy-related pathologies. Recently, Amiri *et al.* [13] discussed current challenges in grey matter atrophy quantification: lack of public and well-annotated reference datasets, lesion-sensitive processing pipelines, data heterogeneity, and lack of research investigating the effect of abnormalities in current processing pipelines. We could use our generation framework to target the first challenge, i.e. generate cross-sectional and longitudinal atrophy quantification benchmarks, allowing us to develop novel atrophy quantification algorithms and establish their pros and cons and, ultimately, extend the clinical validation of existing tools.

In Chapter 6, we showed that deep learning could be used for atrophy quantification strategy and demonstrated that it could compete with the state-of-the-art. To further validate our proposal, we need to compare against other relevant segmentation-based and registration-based methods. Despite being criticised because of its vulnerability against imaging variations, SIENA continues being a suitable method for measuring brain atrophy. To improve our proposal, we could scrutinise small regions of interest, particularly those damaged by the pathology under examination; process positive and negative values separately to account for enlargement and shrinkage brain processes independently; use other alternative statistics to reflect the possibly multimodal nature of the atrophy; and carry out a two-way inspection (baseline to follow-up and vice versa) to account for potential regression of abnormal tissues. We could use the approach presented in Section 4.5.3 where we used a deep learning method for correcting errors made by another tool, SIENA in this case.

Cerebral atrophy is one of many neuroimaging features of brain diseases. Therefore, the tools developed in this thesis should be integrated with others developed by our group. In such a way, we would create a system which would examine the brain and produce holistic measures of brain status that might help medical doctors to understand better the mechanisms of these brain diseases, predict diseases, patient outcome, assess treatment effectiveness and treatment response.

7.2.2 Future research lines

As discussed in our literature review, there are general issues that need to be addressed prior to deploying any deep learning based method for routine clinical practice regarding data harmonisation, data availability, generalisation, and interpretability. In the long term, there are various research lines our research group could study.

Magnetic resonance imaging is prone to imaging artefacts because of intrinsic and extrinsic factors, such as scanner instability, truncation and motion artefacts, aliasing, among others [309–312]. These issues manifest in visual artefacts which may compromise any subsequent assessment and, in severe cases, may even result in useless data. In this thesis, we observed that data harmonisation could help models to generalise better to unseen cases. Testing of current image enhancement and image standardisation techniques or development of new ones could ease domain shift problems [196]. Moreover, deep learning could be used for reducing these imaging artefacts [313–317, 366].

According to our literature revision and our own experience, deep learning can achieve outstanding results as long as there are enough manually annotated cases. Models do not necessarily need to be trained from scratch, instead legacy models can be fine-tuned to process new data [200, 201]. Even a single case can allow pre-trained models to perform similar to fully trained ones on another domain [198,199]. However, these approaches are not yet widespread in the literature nor are cross-dataset evaluations. In some tasks such as classification and segmentation, it might be about time we, researchers, move from showing how capable deep learning is and concentrate on how practical or deployable are our deep learning based solutions.

Part of the generalisability problem that deep learning based methods experience in the medical domain is that large, heterogeneous, and well-labelled datasets are scarce. Although valuable datasets have been released in recent years, more publicly available datasets are needed not only for training networks in such a way they generalise better but also for the sake of reproducibility. In this thesis, we showed that image generation could be used to generate controlled evaluation environments and train a deep learning model better. Further development on image synthesis or data augmentation could help to overcome these limitations. Additionally, more frameworks for testing generalisability blindly are needed. Up to now, only a few challenges, such as the Medical Image Segmentation Decathlon and the six-month infant brain MRI segmentation challenge, allow assessing the generalisability and robustness to imaging variations and processing tasks.

The lack of interpretability in deep learning models may prevent their application in medical practice as medical doctors and patients need to understand the decision-making process. However, this task is not trivial in deep learning as models contain thousands of parameters. Efforts for interpreting how and why the network reaches specific verdicts have been made [316,318], but further development in this regard is needed. Additionally, the quantification of uncertainty of the decision may help to improve their reliability and accuracy [267]. Furthermore, uncertainty quantification can help to understand what a deep learning model does not know, a fundamental asset for designing robust models [319].

Appendices

Appendix A

Benchmarking brain tissue segmentation methods

A.1 Segmentation accuracy values

Supplementary table 1. DSC and MHD statistics scored by each one of the considered methods in IBSR18, MICCAI2012, and iSeg2017.

Method	Dataset	ROI	DSC		MHD	
			Mean	**Std**	**Mean**	**Std**
DM2D	IBSR18	CSF	0.87	0.05	6.36	9.22
DM2D	IBSR18	GM	0.96	0.01	1.32	0.3
DM2D	IBSR18	WM	0.92	0.02	1.53	0.4
DM2D-NO	IBSR18	CSF	0.78	0.1	16.02	17.71
DM2D-NO	IBSR18	GM	0.94	0.01	1.76	0.44
DM2D-NO	IBSR18	WM	0.91	0.02	1.97	0.67
DM3D	IBSR18	CSF	0.86	0.07	9.49	10.78
DM3D	IBSR18	GM	0.96	0.01	1.78	0.39
DM3D	IBSR18	WM	0.93	0.02	2.01	0.59
DM3D-NO	IBSR18	CSF	0.69	0.25	11.71	12.95
DM3D-NO	IBSR18	GM	0.93	0.02	1.79	0.4
DM3D-NO	IBSR18	WM	0.89	0.03	2.12	0.59
FAST	IBSR18	CSF	0.47	0.18	35.04	5.54
FAST	IBSR18	GM	0.88	0.01	2.01	0.34
FAST	IBSR18	WM	0.89	0.02	1.06	0.21
KK2D	IBSR18	CSF	0.88	0.04	4.62	3.2
KK2D	IBSR18	GM	0.96	0.01	1.52	0.27
KK2D	IBSR18	WM	0.92	0.02	1.61	0.38

KK2D-NO	IBSR18	CSF	0.77	0.08	8.04	8.55
KK2D-NO	IBSR18	GM	0.93	0.01	1.57	0.28
KK2D-NO	IBSR18	WM	0.9	0.02	1.69	0.42
KK3D	IBSR18	CSF	0.8	0.2	15.13	10.85
KK3D	IBSR18	GM	0.96	0.01	1.8	0.26
KK3D	IBSR18	WM	0.92	0.02	1.98	0.39
KK3D-NO	IBSR18	CSF	0.71	0.17	10.47	12.38
KK3D-NO	IBSR18	GM	0.92	0.01	1.65	0.28
KK3D-NO	IBSR18	WM	0.9	0.02	1.83	0.42
SPM	IBSR18	CSF	0.77	0.08	3.55	1.35
SPM	IBSR18	GM	0.91	0.01	4.84	0.45
SPM	IBSR18	WM	0.88	0.01	2.13	0.23
UN2D	IBSR18	CSF	0.9	0.03	2.37	1.72
UN2D	IBSR18	GM	0.96	0.01	1.44	0.38
UN2D	IBSR18	WM	0.93	0.02	1.54	0.46
UN2D-NO	IBSR18	CSF	0.62	0.14	4.93	8.99
UN2D-NO	IBSR18	GM	0.92	0.01	1.48	0.37
UN2D-NO	IBSR18	WM	0.88	0.02	1.6	0.52
UN3D	IBSR18	CSF	0.88	0.05	30.56	12.09
UN3D	IBSR18	GM	0.96	0.01	4	1.87
UN3D	IBSR18	WM	0.93	0.02	5.21	3.5
UN3D-NO	IBSR18	CSF	0.08	0.15	29.93	11.39
UN3D-NO	IBSR18	GM	0.41	0.31	4.59	2.61
UN3D-NO	IBSR18	WM	0.42	0.23	5.22	3.73
URN2D	IBSR18	CSF	0.89	0.04	2.77	2.39
URN2D	IBSR18	GM	0.96	0.01	1.43	0.41
URN2D	IBSR18	WM	0.92	0.02	1.58	0.46
URN2D-NO	IBSR18	CSF	0.68	0.14	3.05	2.47
URN2D-NO	IBSR18	GM	0.92	0.01	1.47	0.4
URN2D-NO	IBSR18	WM	0.88	0.02	1.6	0.45
URN3D	IBSR18	CSF	0.9	0.04	4.91	6.41
URN3D	IBSR18	GM	0.96	0.01	1.43	0.32
URN3D	IBSR18	WM	0.93	0.02	1.46	0.33
URN3D-NO	IBSR18	CSF	0.05	0.06	40.23	12
URN3D-NO	IBSR18	GM	0.69	0.31	1.48	0.36
URN3D-NO	IBSR18	WM	0.45	0.28	1.5	0.41
DM2D	iSeg2017	CSF	0.91	0.01	1	0.13
DM2D	iSeg2017	GM	0.87	0.01	1.5	0.27
DM2D	iSeg2017	WM	0.86	0.01	1.33	0.17
DM2D-NO	iSeg2017	CSF	0.91	0.01	1.04	0
DM2D-NO	iSeg2017	GM	0.87	0.01	1.53	0.2

DM2D-NO	iSeg2017	WM	0.84	0.02	1.33	0.17
DM2D-S	iSeg2017	CSF	0.91	0.01	1.04	0.13
DM2D-S	iSeg2017	GM	0.85	0.01	2.02	0.38
DM2D-S	iSeg2017	WM	0.81	0.02	1.5	0.2
DM3D	iSeg2017	CSF	0.94	0.01	1	0
DM3D	iSeg2017	GM	0.91	0.01	1.27	0.33
DM3D	iSeg2017	WM	0.89	0.02	1.04	0.13
DM3D-NO	iSeg2017	CSF	0.92	0.01	1.04	0.13
DM3D-NO	iSeg2017	GM	0.88	0.01	1.35	0.3
DM3D-NO	iSeg2017	WM	0.86	0.02	1.08	0.17
DM3D-S	iSeg2017	CSF	0.94	0.01	1	0
DM3D-S	iSeg2017	GM	0.89	0.01	1.5	0.2
DM3D-S	iSeg2017	WM	0.87	0.02	1.12	0.2
FAST	iSeg2017	CSF	0.76	0.03	3	1.52
FAST	iSeg2017	GM	0.65	0.04	2.42	0.17
FAST	iSeg2017	WM	0.58	0.37	3.49	0.38
KK2D	iSeg2017	CSF	0.91	0.01	1.04	0.13
KK2D	iSeg2017	GM	0.87	0.01	1.56	0.27
KK2D	iSeg2017	WM	0.85	0.02	1.25	0.21
KK2D-NO	iSeg2017	CSF	0.91	0.01	1.04	0.13
KK2D-NO	iSeg2017	GM	0.87	0.01	1.56	0.27
KK2D-NO	iSeg2017	WM	0.84	0.01	1.29	0.2
KK2D-S	iSeg2017	CSF	0.91	0.01	1.12	0.2
KK2D-S	iSeg2017	GM	0.85	0.01	2.04	0.22
KK2D-S	iSeg2017	WM	0.81	0.02	1.48	0.13
KK3D	iSeg2017	CSF	0.94	0.01	1	0
KK3D	iSeg2017	GM	0.9	0.01	1.51	0.15
KK3D	iSeg2017	WM	0.88	0.02	1.12	0.2
KK3D-NO	iSeg2017	CSF	0.9	0.03	1.04	0.13
KK3D-NO	iSeg2017	GM	0.83	0.07	1.51	0.15
KK3D-NO	iSeg2017	WM	0.8	0.07	1.12	0.2
KK3D-S	iSeg2017	CSF	0.94	0.01	1.04	0.13
KK3D-S	iSeg2017	GM	0.88	0.01	1.54	0.21
KK3D-S	iSeg2017	WM	0.86	0.01	1.25	0.21
SPM	iSeg2017	CSF	0.78	0.03	3.12	0.56
SPM	iSeg2017	GM	0.77	0.02	2.89	0.2
SPM	iSeg2017	WM	0.64	0.03	6.42	0.93
UN2D	iSeg2017	CSF	0.92	0.01	1.04	0.13
UN2D	iSeg2017	GM	0.88	0.01	1.55	0.31
UN2D	iSeg2017	WM	0.86	0.02	1.25	0.21
UN2D-NO	iSeg2017	CSF	0.9	0.01	1.04	0.13

UN2D-NO	iSeg2017	GM	0.84	0.01	1.62	0.28
UN2D-NO	iSeg2017	WM	0.8	0.01	1.33	0.17
UN2D-S	iSeg2017	CSF	0.92	0.01	1	0
UN2D-S	iSeg2017	GM	0.88	0.01	1.86	0.23
UN2D-S	iSeg2017	WM	0.86	0.02	1.41	0
UN3D	iSeg2017	CSF	0.93	0.01	1.14	0.46
UN3D	iSeg2017	GM	0.9	0.01	1.58	0.74
UN3D	iSeg2017	WM	0.88	0.01	1.29	0.39
UN3D-NO	iSeg2017	CSF	0.43	0.12	12	10.34
UN3D-NO	iSeg2017	GM	0.02	0.03	21.53	15.23
UN3D-NO	iSeg2017	WM	0.3	0.16	15.43	5.02
UN3D-S	iSeg2017	CSF	0.93	0.01	1	0
UN3D-S	iSeg2017	GM	0.89	0.01	1.65	0.32
UN3D-S	iSeg2017	WM	0.87	0.02	1.21	0.22
URN2D	iSeg2017	CSF	0.91	0.01	1.04	0.13
URN2D	iSeg2017	GM	0.87	0.01	1.55	0.3
URN2D	iSeg2017	WM	0.85	0.01	1.25	0.17
URN2D-NO	iSeg2017	CSF	0.9	0.01	1.04	0.13
URN2D-NO	iSeg2017	GM	0.84	0.01	1.62	0.34
URN2D-NO	iSeg2017	WM	0.8	0.02	1.33	0
URN2D-S	iSeg2017	CSF	0.91	0.01	1.04	0.13
URN2D-S	iSeg2017	GM	0.85	0.01	2.23	0.62
URN2D-S	iSeg2017	WM	0.8	0.02	1.51	0.15
URN3D	iSeg2017	CSF	0.93	0.01	1	0
URN3D	iSeg2017	GM	0.89	0.01	1.54	0.21
URN3D	iSeg2017	WM	0.87	0.02	1.17	0.21
URN3D-NO	iSeg2017	CSF	0.43	0.15	11.34	13.12
URN3D-NO	iSeg2017	GM	0.07	0.21	25.34	12.53
URN3D-NO	iSeg2017	WM	0.22	0.24	17.24	5.34
URN3D-S	iSeg2017	CSF	0.92	0.03	1.57	1.81
URN3D-S	iSeg2017	GM	0.87	0.01	2.03	0.66
URN3D-S	iSeg2017	WM	0.83	0.04	1.45	0.3
DM2D	MICCAI2012	CSF	0.89	0.05	1.66	0.69
DM2D	MICCAI2012	GM	0.93	0.03	1.49	0.74
DM2D	MICCAI2012	WM	0.95	0.02	1.32	0.57
DM2D-NO	MICCAI2012	CSF	0.88	0.05	1.72	0.91
DM2D-NO	MICCAI2012	GM	0.93	0.03	1.52	0.78
DM2D-NO	MICCAI2012	WM	0.96	0.02	1.33	0.58
DM3D	MICCAI2012	CSF	0.92	0.03	1.22	0.36
DM3D	MICCAI2012	GM	0.94	0.02	1.15	0.42
DM3D	MICCAI2012	WM	0.97	0.01	1.14	0.52

DM3D-NO	MICCAI2012	CSF	0.89	0.05	1.21	0.3
DM3D-NO	MICCAI2012	GM	0.93	0.03	1.24	0.47
DM3D-NO	MICCAI2012	WM	0.95	0.02	1.19	0.54
FAST	MICCAI2012	CSF	0.16	0.13	39.62	2.69
FAST	MICCAI2012	GM	0.91	0.02	5.84	0.45
FAST	MICCAI2012	WM	0.66	0.08	5.16	0.24
KK2D	MICCAI2012	CSF	0.91	0.03	1.41	0.49
KK2D	MICCAI2012	GM	0.94	0.02	1.23	0.53
KK2D	MICCAI2012	WM	0.96	0.01	1.46	0.2
KK2D-NO	MICCAI2012	CSF	0.9	0.03	1.44	0.5
KK2D-NO	MICCAI2012	GM	0.93	0.02	1.27	0.53
KK2D-NO	MICCAI2012	WM	0.96	0.01	1.18	0.27
KK3D	MICCAI2012	CSF	0.91	0.04	1.12	0.28
KK3D	MICCAI2012	GM	0.94	0.02	2.04	0.43
KK3D	MICCAI2012	WM	0.96	0.01	1.48	0.45
KK3D-NO	MICCAI2012	CSF	0.86	0.08	1.18	0.24
KK3D-NO	MICCAI2012	GM	0.93	0.02	1.21	0.44
KK3D-NO	MICCAI2012	WM	0.96	0.02	1.12	0.45
SPM	MICCAI2012	CSF	0.39	0.29	38.87	4.56
SPM	MICCAI2012	GM	0.63	0.38	7.91	0.52
SPM	MICCAI2012	WM	0.62	0.37	5.17	0.16
UN2D	MICCAI2012	CSF	0.91	0.03	1.19	0.21
UN2D	MICCAI2012	GM	0.94	0.02	1.27	0.51
UN2D	MICCAI2012	WM	0.96	0.02	1.2	0.52
UN2D-NO	MICCAI2012	CSF	0.87	0.06	1.37	0.29
UN2D-NO	MICCAI2012	GM	0.92	0.02	1.33	0.54
UN2D-NO	MICCAI2012	WM	0.95	0.02	1.24	0.52
UN3D	MICCAI2012	CSF	0.92	0.03	1.76	0.86
UN3D	MICCAI2012	GM	0.94	0.02	1.37	0.54
UN3D	MICCAI2012	WM	0.96	0.01	1.2	0.52
UN3D-NO	MICCAI2012	CSF	0.04	0.02	30.05	18.24
UN3D-NO	MICCAI2012	GM	0	0	24.92	16.35
UN3D-NO	MICCAI2012	WM	0.67	0.01	5.35	5.18
URN2D	MICCAI2012	CSF	0.91	0.03	1.95	1.5
URN2D	MICCAI2012	GM	0.94	0.03	1.27	0.55
URN2D	MICCAI2012	WM	0.96	0.02	1.21	0.55
URN2D-NO	MICCAI2012	CSF	0.58	0.13	4.77	4.51
URN2D-NO	MICCAI2012	GM	0.91	0.03	1.41	0.65
URN2D-NO	MICCAI2012	WM	0.94	0.02	1.25	0.55
URN3D	MICCAI2012	CSF	0.9	0.05	1.79	1.13
URN3D	MICCAI2012	GM	0.95	0.02	1.18	0.46

URN3D	MICCAI2012	WM	0.97	0.01	1.15	0.49
URN3D-NO	MICCAI2012	CSF	0.07	0.1	18.37	11.38
URN3D-NO	MICCAI2012	GM	0.78	0.09	8.35	4.95
URN3D-NO	MICCAI2012	WM	0.87	0.04	7.46	5
DM2D-(23,23)	iSeg2017	CSF	0.9	0.02	1.04	0.12
DM2D-(23,23)	iSeg2017	GM	0.86	0.01	1.41	0
DM2D-(23,23)	iSeg2017	WM	0.84	0.02	1.77	0.3
DM2D-(37,37)	iSeg2017	CSF	0.91	0.02	1.04	0.12
DM2D-(37,37)	iSeg2017	GM	0.85	0.02	1.48	0.13
DM2D-(37,37)	iSeg2017	WM	0.81	0.02	2.11	0.22
DM3D-(23,23,23)	iSeg2017	CSF	0.94	0.01	1	0
DM3D-(23,23,23)	iSeg2017	GM	0.89	0.01	1.08	0.17
DM3D-(23,23,23)	iSeg2017	WM	0.87	0.01	1.42	0.28
DM3D-(37,37,37)	iSeg2017	CSF	0.94	0.02	1	0
DM3D-(37,37,37)	iSeg2017	GM	0.89	0.02	1.17	0.2
DM3D-(37,37,37)	iSeg2017	WM	0.87	0.02	1.42	0.28
KK2D-(28,28)	iSeg2017	CSF	0.86	0.02	1.41	0
KK2D-(28,28)	iSeg2017	GM	0.88	0.02	1.41	0
KK2D-(28,28)	iSeg2017	WM	0.84	0.02	1.88	0.28
KK2D-(36,36)	iSeg2017	CSF	0.86	0.02	1.41	0
KK2D-(36,36)	iSeg2017	GM	0.87	0.02	1.5	0.19
KK2D-(36,36)	iSeg2017	WM	0.83	0.02	2.18	0.26
KK3D-(28,28,28)	iSeg2017	CSF	0.91	0.03	1.04	0.12
KK3D-(28,28,28)	iSeg2017	GM	0.85	0.05	1.52	0.29
KK3D-(28,28,28)	iSeg2017	WM	0.81	0.05	1.97	0.55
KK3D-(36,36,36)	iSeg2017	CSF	0.88	0.03	1.71	0.88
KK3D-(36,36,36)	iSeg2017	GM	0.77	0.05	2.13	0.6
KK3D-(36,36,36)	iSeg2017	WM	0.65	0.18	3.73	2.17
UN2D-(08,08)	iSeg2017	CSF	0.86	0.02	1.41	0
UN2D-(08,08)	iSeg2017	GM	0.77	0.05	2.29	0.38
UN2D-(08,08)	iSeg2017	WM	0.68	0.08	3.11	0.39
UN2D-(16,16)	iSeg2017	CSF	0.9	0.02	1.08	0.17
UN2D-(16,16)	iSeg2017	GM	0.84	0.05	1.5	0.19
UN2D-(16,16)	iSeg2017	WM	0.81	0.04	2.05	0.54
UN3D-(08,08,08)	iSeg2017	CSF	0.84	0.04	2.41	1.48
UN3D-(08,08,08)	iSeg2017	GM	0.71	0.09	2.3	0.47
UN3D-(08,08,08)	iSeg2017	WM	0.65	0.11	3.21	0.69
UN3D-(16,16,16)	iSeg2017	CSF	0.91	0.02	1.1	0.3
UN3D-(16,16,16)	iSeg2017	GM	0.84	0.08	1.45	0.29
UN3D-(16,16,16)	iSeg2017	WM	0.82	0.05	1.96	0.65
URN2D-(08,08)	iSeg2017	CSF	0.85	0.03	1.55	0.55

URN2D-(08,08)	iSeg2017	GM	0.73	0.13	2.34	0.56
URN2D-(08,08)	iSeg2017	WM	0.68	0.1	3.24	0.83
URN2D-(16,16)	iSeg2017	CSF	0.9	0.02	1.07	0.22
URN2D-(16,16)	iSeg2017	GM	0.8	0.12	1.76	0.56
URN2D-(16,16)	iSeg2017	WM	0.77	0.06	2.41	0.53
URN3D-(08,08,08)	iSeg2017	CSF	0.87	0.03	1.31	0.3
URN3D-(08,08,08)	iSeg2017	GM	0.76	0.07	2.01	0.23
URN3D-(08,08,08)	iSeg2017	WM	0.71	0.07	2.92	0.55
URN3D-(16,16,16)	iSeg2017	CSF	0.91	0.02	1	0
URN3D-(16,16,16)	iSeg2017	GM	0.81	0.16	1.69	0.81
URN3D-(16,16,16)	iSeg2017	WM	0.81	0.06	1.96	0.67

A.2 Considered ADNI cases

002_5018, 003_4136, 003_4152, 003_4373, 003_4892, 003_5165, 003_5187, 005_4707, 005_4910, 005_5038, 005_5119, 006_4546, 006_4867, 007_4911, 007_5196, 009_5027, 009_5037, 009_5224, 009_5252, 011_4827, 011_4845, 011_4906, 011_4912, 011_4949, 013_5071, 014_4615, 016_4353, 016_4583, 016_4887, 016_5032, 016_5057, 018_4696, 018_4733, 018_5074, 018_5240, 019_4549, 019_5012, 019_5019, 021_4718, 021_4924, 023_5120, 023_5241, 024_4223, 024_4280, 024_4905, 024_5054, 027_4801, 027_4802, 027_4938, 027_4962, 027_4964, 029_4307, 031_4024, 032_4755, 033_5013, 033_5017, 033_5087, 035_4783, 036_4820, 036_5063, 036_5149, 036_5210, 037_4001, 037_4770, 037_4879, 037_5162, 051_4980, 051_5005, 052_4959, 052_5062, 053_5070, 053_5208, 067_4728, 067_5205, 068_4859, 068_4968, 068_5146, 070_4692, 070_4719, 073_4853, 073_5090, 082_5029, 082_5184, 094_4282, 094_4737, 098_4201, 099_4994, 109_4378, 114_4379, 116_4209, 116_4338, 116_4537, 116_4732, 126_4686, 127_4749, 127_4940, 127_4992, 127_5028, 127_5056, 127_5058, 127_5067, 127_5095, 128_4772, 128_4774, 128_4792, 128_5123, 130_4589, 130_4730, 130_4971, 130_4982, 130_4984, 130_4990, 130_4997, 130_5006, 130_5059, 130_5231, 131_5138, 135_4863, 135_4954, 135_5015, 136_4993, 137_4756

A.3 ADNI cases considered in Chapter 6 for testing

A.3.1 Alzheimer's disease patients

006_4153, 006_4192, 007_4568, 014_4039, 016_4009, 016_4591, 019_4252, 019_4477, 023_4501, 094_4089, 098_4215, 116_4195, 116_4625, 123_4526, 126_4494, 127_4500, 130_4641, 130_4660, 135_4657, 135_4676, 137_4211, 137_4258, 137_4672, 153_4172

A.3.2 Control subjects

002_4262, 003_4288, 007_4516, 009_4612, 023_4448, 029_4384, 029_4385, 029_4585, 035_4464, 053_4578, 094_4503, 094_4560, 099_4104, 116_4483, 128_4832, 129_4369, 129_4371, 129_4422, 941_4292

Bibliography

[1] Simon H Parson. Clinically oriented anatomy. *Journal of Anatomy*, 215(4):474, 2009.

[2] Wojciech Sokołowski, Karolina Barszcz, Marta Kupczyńska, Norbert Czubaj, Michał Skibniewski, and Halina Purzyc. Lymphatic drainage of cerebrospinal fluid in mammals–are arachnoid granulations the main route of cerebrospinal fluid outflow? *Biologia*, 73(6):563–568, 2018.

[3] Bahram Mokri. The Monro–Kellie hypothesis: applications in CSF volume depletion. *Neurology*, 56(12):1746–1748, 2001.

[4] LO Wahlund, I Agartz, Ove Almqvist, H Basun, L Forssell, J Sääf, and L Wetterberg. The brain in healthy aged individuals: MR imaging. *Radiology*, 174(3):675–679, 1990.

[5] Marie C Henry-Feugeas, Philippe Azouvi, Anne Fontaine, Pierre Denys, Bernard Bussel, Fahid Maaz, Yves Samson, and Elisabeth Schouman-Claeys. MRI analysis of brain atrophy after severe closed-head injury: relation to clinical status. *Brain Injury*, 14(7):597–604, 2000.

[6] Christian Hoffmann, Luitpold Distel, Stefan Knippen, Thomas Gryc, Manuel Alexander Schmidt, Rainer Fietkau, and Florian Putz. Brain volume reduction after whole-brain radiotherapy: quantification and prognostic relevance. *Neuro-oncology*, 20(2):268–278, 2018.

[7] Glenn T Stebbins, David L Nyenhuis, Changsheng Wang, Jennifer L Cox, Sally Freels, Katherine Bangen, Leyla deToledo Morrell, Kumar Sripathirathan, Michael Moseley, David A Turner, et al. Gray matter atrophy in patients with ischemic stroke with cognitive impairment. *Stroke*, 39(3):785–793, 2008.

[8] Joanna M Wardlaw, Colin Smith, and Martin Dichgans. Small vessel disease: mechanisms and clinical implications. *The Lancet Neurology*, 2019.

[9] Sander V Haijma, Neeltje Van Haren, Wiepke Cahn, P Cédric MP Koolschijn, Hilleke E Hulshoff Pol, and René S Kahn. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia Bulletin*, 39(5):1129–1138, 2013.

[10] Theo GM van Erp, Derrek P Hibar, Jerod M Rasmussen, David C Glahn, Godfrey D Pearlson, Ole A Andreassen, Ingrid Agartz, Lars T Westlye, Unn K Haukvik, Anders M Dale, et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, 21(4):547–553, 2016.

[11] Keith S Cover, Ronald A van Schijndel, Bob W van Dijk, Alberto Redolfi, Dirk L Knol, Giovanni B Frisoni, Frederik Barkhof, Hugo Vrenken, Alzheimer's Disease Neuroimaging Initiative, et al. Assessing the reproducibility of the SienaX and Siena brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. *Psychiatry Research: Neuroimaging*, 193(3):182–190, 2011.

[12] Maria A Rocca, Marco Battaglini, Ralph HB Benedict, Nicola De Stefano, Jeroen JG Geurts, Roland G Henry, Mark A Horsfield, Mark Jenkinson, Elisabetta Pagani, and Massimo Filippi. Brain MRI atrophy quantification in MS: from methods to clinical application. *Neurology*, 88(4):403–413, 2017.

[13] Houshang Amiri, Alexandra de Sitter, Kerstin Bendfeldt, Marco Battaglini, Claudia AM Gandini Wheeler-Kingshott, Massimiliano Calabrese, Jeroen JG Geurts, Maria A Rocca, Jaume Sastre-Garriga, Christian Enzinger, et al. Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI. *NeuroImage: Clinical*, 19:466–475, 2018.

[14] Àlex Rovira, Mike P Wattjes, Mar Tintoré, Carmen Tur, Tarek A Yousry, Maria P Sormani, Nicola De Stefano, Massimo Filippi, Cristina Auger, Maria A Rocca, et al. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. *Nature Reviews Neurology*, 11(8):471, 2015.

[15] Martijn D Steenwijk, Jeroen JG Geurts, Marita Daams, Betty M Tijms, Alle Meije Wink, Lisanne J Balk, Prejaas K Tewarie, Bernard MJ Uitdehaag, Frederik Barkhof, Hugo Vrenken, et al. Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant. *Brain*, 139(1):115–126, 2016.

[16] Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Alex Rovira, Jaume Sastre-Garriga, Mar Tintorè, Jette L Frederiksen, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology*, 15(3):292–303, 2016.

[17] Loredana Storelli, Maria A Rocca, Elisabetta Pagani, Wim Van Hecke, Mark A Horsfield, Nicola De Stefano, Alex Rovira, Jaume Sastre-Garriga, Jacqueline Palace, Diana Sima, et al. Measurement of whole-brain and gray matter atrophy in multiple sclerosis: assessment with MR imaging. *Radiology*, 288(2):554–564, 2018.

[18] Kaisar Kushibar, Sergi Valverde, Sandra González-Villà, Jose Bernal, Mariano Cabezas, Arnau Oliver, and Xavier Lladó. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Medical Image Analysis*, 48:177–186, 2018.

[19] Jose Bernal, Kaisar Kushibar, Mariano Cabezas, Sergi Valverde, Arnau Oliver, and Xavier Lladó. Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access*, 7:89986–90002, 2019.

[20] Paolo Preziosa, Maria A Rocca, Gianna C Riccitelli, Lucia Moiola, Loredana Storelli, Mariaemma Rodegher, Giancarlo Comi, Alessio Signori, Andrea Falini, and Massimo Filippi. Effects of natalizumab and fingolimod on clinical, cognitive, and magnetic resonance imaging measures in multiple sclerosis. *Neurotherapeutics*, 17(1):208–217, 2020.

[21] World Health Organisation. Dementia. `https://www.who.int/news-room/fact-sheets/detail/dementia`, 2019. Accessed: 23-03-2020.

[22] Murali Vijayan and P Hemachandra Reddy. Stroke, vascular dementia, and Alzheimer's disease: molecular links. *Journal of Alzheimer's Disease*, 54(2):427–443, 2016.

[23] G Fein, V Di Sclafani, J Tanabe, V Cardenas, MW Weiner, WJ Jagust, Bruce R Reed, D Norman, N Schuff, L Kusdra, et al. Hippocampal and cortical atrophy predict dementia in subcortical ischemic vascular disease. *Neurology*, 55(11):1626–1635, 2000.

[24] Dennis Chan, Nick C Fox, Rachael I Scahill, William R Crum, Jennifer L Whitwell, Guy Leschziner, Alex M Rossor, John M Stevens, Lisa Cipolotti, and Martin N Rossor. Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Annals of Neurology*, 49(4):433–442, 2001.

[25] John T O'Brien, Michael J Firbank, Karen Ritchie, Katie Wells, Guy B Williams, Craig W Ritchie, and Li Su. Association between midlife dementia risk factors and longitudinal brain atrophy: the PREVENT-Dementia study. *Journal of Neurology, Neurosurgery & Psychiatry*, 91(2):158–161, 2020.

[26] Belinda Yew, Daniel A Nation, and Alzheimer's Disease Neuroimaging Initiative. Cerebrovascular resistance: effects on cognitive decline, cortical atrophy, and progression to dementia. *Brain*, 140(7):1987–2001, 2017.

[27] Martin Prince, A Wimo, M Guerchet, GC Ali, YT Wu, M Prina, et al. The global impact of dementia. *World Alzheimer Report*, pages 1–82, 2015.

[28] Lawrence Steinman. Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system. *Cell*, 85(3):299–302, 1996.

[29] Ranjan Dutta and Bruce D Trapp. Mechanisms of neuronal dysfunction and degeneration in multiple sclerosis. *Progress in Neurobiology*, 93(1):1–12, 2011.

[30] Nancy D Chiaravalloti and John DeLuca. Cognitive impairment in multiple sclerosis. *The Lancet Neurology*, 7(12):1139–1151, 2008.

[31] Jeffrey M Gelfand. Multiple sclerosis: diagnosis, differential diagnosis, and clinical presentation. In *Handbook of Clinical Neurology*, volume 122, pages 269–290. Elsevier, 2014.

[32] Mitchell T Wallin, William J Culpepper, Emma Nichols, Zulfiqar A Bhutta, Tsegaye Tewelde Gebrehiwot, Simon I Hay, Ibrahim A Khalil, Kristopher J Krohn, Xiaofeng Liang, Mohsen Naghavi, et al. Global, regional, and national burden of multiple sclerosis 1990–2016: a systematic analysis for the global burden of disease study. *The Lancet Neurology*, 18(3):269–285, 2019.

[33] Narges Dargahi, Maria Katsara, Theodore Tselios, Maria-Eleni Androutsou, Maximilian De Courten, John Matsoukas, and Vasso Apostolopoulos. Multiple sclerosis: immunopathology and treatment update. *Brain Sciences*, 7(7):78, 2017.

[34] Natalie Kappus, Bianca Weinstock-Guttman, Jesper Hagemeier, Cheryl Kennedy, Rebecca Melia, Ellen Carl, Deepa P Ramasamy, Mariya Cherneva, Jacqueline Durfee, Niels Bergsland, et al. Cardiovascular risk factors are associated with increased lesion burden and brain atrophy in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(2):181–187, 2016.

[35] Nicola De Stefano, Maria Laura Stromillo, Antonio Giorgio, Maria Letizia Bartolozzi, Marco Battaglini, Mariella Baldini, Emilio Portaccio, Maria Pia Amato, and Maria Pia Sormani. Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(1):93–99, 2016.

[36] Robert A Bermel and Rohit Bakshi. The measurement and clinical relevance of brain atrophy in multiple sclerosis. *The Lancet Neurology*, 5(2):158–170, 2006.

[37] Jaume Sastre-Garriga, Deborah Pareto, and Àlex Rovira. Brain atrophy in multiple sclerosis: clinical relevance and technical aspects. *Neuroimaging Clinics*, 27(2):289–300, 2017.

[38] Paul Browne, Dhia Chandraratna, Ceri Angood, Helen Tremlett, Chris Baker, Bruce V Taylor, and Alan J Thompson. Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. *Neurology*, 83(11):1022–1024, 2014.

[39] U.S. Food and Drug Administration (FDA). Medical X-ray imaging. `https://www.fda.gov/radiation-emitting-products/medical-imaging/medical-x-ray-imaging`, 2017. [Online; accessed 25 Jan 2020].

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

[41] A. Berger. Magnetic resonance imaging. *The BMJ*, 324(7328):35, 2002.

[42] M. A. Brown and R. C. Semelka. *Relaxation*, pages 21–31. John Wiley & Sons, Inc., 2005.

[43] R. Bakshi, S. Ariyaratana, R. H. B. Benedict, and L. Jacobs. Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions. *Archives of Neurology*, 58(5):742–748, 2001.

[44] R. H. Hashemi, W. G. Bradley Jr, D.-Y. Chen, J. E. Jordan, J. A. Queralt, A. E Cheng, and J. N. Henrie. Suspected multiple sclerosis: MR imaging with a thin-section fast FLAIR pulse sequence. *Radiology*, 196(2):505–510, 1995.

[45] Jeffrey A Lieberman, Bernhard Bogerts, Gustav Degreef, Manzar Ashtari, George Lantos, and Jose Alvir. Qualitative assessment of brain morphology in acute and chronic schizophrenia. *The American journal of psychiatry*, 1992.

[46] PH Scheltens, D Leys, F Barkhof, D Huglo, HC Weinstein, P Vermersch, M Kuiper, M Steinling, E Ch Wolters, and J Valk. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(10):967–972, 1992.

[47] CJ Galton, B Gomez-Anson, N Antoun, P Scheltens, K Patterson, M Graves, BJ Sahakian, and JR Hodges. Temporal lobe rating scale: application to Alzheimer's disease and frontotemporal dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 70(2):165–173, 2001.

[48] C Farrell, F Chappell, PA Armitage, P Keston, A MacLullich, S Shenkin, and JM Wardlaw. Development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years. *European Radiology*, 19(1):177–183, 2009.

[49] Elizabeth A Krupinski, Kevin S Berbaum, Robert T Caldwell, Kevin M Schartz, and John Kim. Long radiology workdays reduce detection and accommodation accuracy. *Journal of the American College of Radiology*, 7(9):698–704, 2010.

[50] Tarek N Hanna, Matthew E Zygmont, Ryan Peterson, David Theriot, Haris Shekhani, Jamlik-Omari Johnson, and Elizabeth A Krupinski. The effects of fatigue from overnight shifts on radiology search patterns and diagnostic performance. *Journal of the American College of Radiology*, 15(12):1709–1716, 2018.

[51] Jennifer A. McCabe. Floor and ceiling effects. *The SAGE Encyclopedia of Abnormal and Clinical Psychology*, 2017.

[52] M. J. Cardoso, A. Melbourne, G. S. Kendall, M. Modat, N. J. Robertson, N. Marlow, and S. Ourselin. AdaPT: an adaptive preterm segmentation algorithm for neonatal brain MRI. *NeuroImage*, 65:97–108, 2013.

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

[54] D. L. Pham. Robust fuzzy segmentation of magnetic resonance images. In *IEEE Symposium on Computer-Based Medical Systems*, pages 127–131, 2001.

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

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

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

[58] Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56(3):907–922, 2011.

[59] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[60] Stephen M Smith, Yongyue Zhang, Mark Jenkinson, Jacqueline Chen, PM Matthews, Antonio Federico, and Nicola De Stefano. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479–489, 2002.

[61] Sergi Valverde, Arnau Oliver, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, and Xavier Lladó. Automated tissue segmentation of MR brain images in the presence of white matter lesions. *Medical Image Analysis*, 35:446–457, 2017.

[62] Kunio Nakamura, Simon F Eskildsen, Sridar Narayanan, Douglas L Arnold, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. Improving the SIENA performance using BEaST brain extraction. *PloS one*, 13(9), 2018.

[63] Julio Acosta-Cabronero, Guy B Williams, João MS Pereira, George Pengas, and Peter J Nestor. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. *NeuroImage*, 39(4):1654–1665, 2008.

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

[65] Sergi Valverde, Arnau Oliver, Eloy Roura, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Jaume Sastre-Garriga, Xavier Montalban, Àlex Rovira, and Xavier Lladó. Quantifying brain tissue volume in multiple sclerosis with automated lesion segmentation and filling. *NeuroImage: Clinical*, 9:640–647, 2015.

[66] Sandra González-Villà, Arnau Oliver, Yuankai Huo, Xavier Lladó, and Bennett A Landman. Brain structure segmentation in the presence of multiple sclerosis lesions. *NeuroImage: Clinical*, 22:101709, 2019.

[67] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015.

[68] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deb-

orah Blacker, et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, 46(1):177–192, 2009.

[69] Navid Shiee, Pierre-Louis Bazin, Jennifer L Cuzzocreo, Ari Blitz, and Dzung L Pham. Segmentation of brain images using adaptive atlases with application to ventriculomegaly. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 1–12. Springer, 2011.

[70] RA Rudick, E Fisher, J-C Lee, J Simon, L Jacobs, Multiple Sclerosis Collaborative Research Group, et al. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. *Neurology*, 53(8):1698–1698, 1999.

[71] Kunio Nakamura, Nicolas Guizard, Vladimir S Fonov, Sridar Narayanan, D Louis Collins, and Douglas L Arnold. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. *NeuroImage: Clinical*, 4:10–17, 2014.

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

[73] NC Fox, R Jenkins, SM Leary, VL Stevenson, NA Losseff, WR Crum, Richard J Harvey, MN Rossor, DH Miller, and AJ Thompson. Progressive cerebral atrophy in MS: a serial study using registered, volumetric MRI. *Neurology*, 54(4):807–812, 2000.

[74] Marco Battaglini, Mark Jenkinson, Nicola De Stefano, and Alzheimer's Disease Neuroimaging Initiative. SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI. *Human Brain Mapping*, 39(3):1063–1077, 2018.

[75] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

[76] Abdelghani Bouziane, Djelloul Bouchiha, Noureddine Doumi, and Mimoun Malki. Question answering systems: survey and trends. *Procedia Computer Science*, 73:366–375, 2015.

[77] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas,

Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019.

[78] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[79] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[80] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[81] Maryam Badar, Muhammad Haris, and Anam Fatima. Application of deep learning for retinal image analysis: A review. *Computer Science Review*, 35:100203, 2020.

[82] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

[83] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of Digital Imaging*, 30(4):449–459, 2017.

[84] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[86] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[87] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[88] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[89] Akira Hasegawa, Shih-Chung B Lo, Matthew T Freedman, and Seong K Mun. Convolution neural-network-based detection of lung structures. In *Medical Imaging 1994*, pages 654–662. International Society for Optics and Photonics, 1994.

[90] S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong K Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[91] Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Datong Wei, Mark A Helvie, Dorit D Adler, and Mitchell M Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging*, 15(5):598–610, 1996.

[92] Anders Eklund, Paul Dufort, Daniel Forsberg, and Stephen M LaConte. Medical image processing on the GPU–past, present and future. *Medical Image Analysis*, 17(8):1073–1094, 2013.

[93] Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, et al. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014.

[94] S. Valverde, A. Oliver, M. Cabezas, E. Roura, and X. Lladó. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. *Journal of Magnetic Resonance Imaging*, 41(1):93–101, 2015.

[95] L.P. Clarke, R.P. Velthuizen, M.A. Camacho, J.J. Heine, M. Vaidyanathan, L.O. Hall, R.W. Thatcher, and M.L. Silbiger. MRI segmentation: methods and applications. *Magnetic Resonance Imaging*, 13(3):343–368, 1995.

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

[97] A. W.-C. Liew and H. Yan. Current methods in the automatic tissue segmentation of 3D magnetic resonance brain images. *Current Medical Imaging Reviews*, 2(1):91–103, 2006.

[98] Guang Jia, Steven B Heymsfield, Jinyuan Zhou, Guang Yang, and Yukihisa Takayama. Quantitative biomedical imaging: techniques and clinical applications. *BioMed Research International*, 2016, 2016.

[99] David H Miller, Frederik Barkhof, Joseph A Frank, Geoffrey JM Parker, and Alan J Thompson. Measurement of atrophy in multiple sclerosis: pathological basis, methodological aspects and clinical relevance. *Brain*, 125(8):1676–1695, 2002.

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

[101] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Handwritten digit recognition with a back-propagation network. In *Neural Information Processing Systems (NIPS)*, 1989.

[102] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[103] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170:456–470, 2018.

[104] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[105] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning*, pages 807–814, 2010.

[106] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[107] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, 2013.

[108] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *coRR*, 2015.

[109] Xiaojie Jin, Chunyan Xu, Jiashi Feng, Yunchao Wei, Junjun Xiong, and Shuicheng Yan. Deep learning with s-shaped rectified linear activation units. In *AAAI Conference on Artificial Intelligence*, 2016.

[110] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International Conference on International Conference on Machine Learning*, volume 28, pages III–1319, 2013.

[111] Jost Tobias Springenberg and Martin Riedmiller. Improving deep neural networks with probabilistic maxout units. *coRR*, abs/1312.6116, 2013.

[112] Hao Chen, Lequan Yu, Qi Dou, Lin Shi, Vincent CT Mok, and Pheng Ann Heng. Automatic detection of cerebral microbleeds via deep learning based 3D feature representation. In *IEEE International Symposium on Biomedical Imaging*, pages 764–767. IEEE, 2015.

[113] Ludovic Trottier, Philippe Gigu, Brahim Chaib-draa, et al. Parametric exponential linear unit for deep convolutional neural networks. In *IEEE International Conference on Machine Learning and Applications*, pages 207–214. IEEE, 2017.

[114] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning.* MIT Press, 2016.

[115] Y.T. Zhou and R. Chellappa. Computation of optical flow using a neural network. In *IEEE International Conference on Neural Networks*, volume 1998, pages 71–78, 1988.

[116] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2559–2566. IEEE, 2010.

[117] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (workshop track)*, 2015.

[118] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[119] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[120] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[121] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[122] Jose Bernal, Kaisar Kushibar, Daniel S Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, 95:64–81, 2019.

[123] Richard Sutton. Two problems with back propagation and other steepest descent learning procedures for networks. In *Annual Conference of the Cognitive Science Society*, pages 823–832, 1986.

[124] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

[125] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[126] Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on Learning Theory*, pages 81–102, 2016.

[127] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence O (1/k2). In *Doklady an SSSR*, volume 269, pages 543–547, 1983.

[128] Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. *30th International Conference on Machine Learning*, 28:1139–1147, 2013.

[129] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[130] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[131] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.

[132] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *coRR*, abs/1412.6980, 2014.

[133] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

[134] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[135] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, page 101552, 2019.

[136] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[137] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.

[138] Geert Litjens, Francesco Ciompi, Jelmer M Wolterink, Bob D de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular Imaging*, 12(8):1549–1565, 2019.

[139] Guangming Zhu, Bin Jiang, Liz Tong, Yuan Xie, Greg Zaharchuk, and Max Wintermark. Applications of deep learning to neuro-imaging techniques. *Frontiers in Neurology*, 10:869, 2019.

[140] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[141] Cagdas Ulas, Dhritiman Das, Michael J Thrippleton, Maria del C Valdes Hernandez, Paul A Armitage, Stephen D Makin, Joanna M Wardlaw, and Bjoern H Menze. Convolutional neural networks for direct inference of pharmacokinetic parameters: Application to stroke dynamic contrast-enhanced MRI. *Frontiers in Neurology*, 9:1147, 2019.

[142] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–527. Springer, 2014.

[143] Mark Lyksborg, Oula Puonti, Mikael Agn, and Rasmus Larsen. An ensemble of 2D convolutional neural networks for tumor segmentation. In *Scandinavian Conference on Image Analysis*, pages 201–211. Springer, 2015.

[144] Ariel Birenbaum and Hayit Greenspan. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 58–67. Springer, 2016.

[145] Holger R Roth, Yinong Wang, Jianhua Yao, Le Lu, Joseph E Burns, and Ronald M Summers. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97850P. International Society for Optics and Photonics, 2016.

[146] Jihye Yun, Jinkon Park, Donghoon Yu, Jaeyoun Yi, Minho Lee, Hee Jun Park, June-Goo Lee, Joon Beom Seo, and Namkug Kim. Improvement of fully automated airway segmentation on volumetric computed tomographic images using a 2.5 dimensional convolutional neural net. *Medical Image Analysis*, 51:13–20, 2019.

[147] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *coRR*, abs/1312.4400, 2013.

[148] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[149] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent CT Mok, Lin Shi, and Pheng-Ann Heng. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Transactions on Medical Imaging*, 35(5):1182–1195, 2016.

[150] Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, 35(5):1229–1239, 2016.

[151] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

[152] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[153] Abhijit Guha Roy, Sailesh Conjeti, Sri Phani Krishna Karri, Debdoot Sheet, Amin Katouzian, Christian Wachinger, and Nassir Navab. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical Optics Express*, 8(8):3627–3642, 2017.

[154] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[155] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–429. Springer, 2018.

[156] Hirohisa Oda, Holger R Roth, Kanwal K Bhatia, Masahiro Oda, Takayuki Kitasaka, Shingo Iwano, Hirotoshi Homma, Hirotsugu Takabatake, Masaki Mori, Hiroshi Natori, et al. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest CT images. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 1057502. International Society for Optics and Photonics, 2018.

[157] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*, 51:21–45, 2019.

[158] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571. IEEE, 2016.

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

[160] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-DenseUNet: hybrid densely connected UNet for liver and tu-

mor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

[161] Jingfei Hu, Hua Wang, Shengbo Gao, Mingkun Bao, Tao Liu, Yaxing Wang, and Jicong Zhang. S-UNet: a bridge-style U-Net framework with a saliency mechanism for retinal vessel segmentation. *IEEE Access*, 7:174167–174177, 2019.

[162] Chengqin Ye, Wei Wang, Shanzhuo Zhang, and Kuanquan Wang. Multi-depth fusion network for whole-heart CT image segmentation. *IEEE Access*, 7:23421–23429, 2019.

[163] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng Ann Heng. Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. In *AAAI Conference on Artificial Intelligence*, 2017.

[164] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[165] Mostafa Salem, Sergi Valverde, Mariano Cabezas, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage: Clinical*, 25:102149, 2020.

[166] Albert Clèrigues, Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, and Xavier Lladó. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Computers in Biology and Medicine*, 115:103487, 2019.

[167] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Inge WM van Uden, Clara I Sanchez, Geert Litjens, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, and Bram Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):1–12, 2017.

[168] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.

[169] Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814, 2017.

[170] Mostafa Salem, Sergi Valverde, Mariano Cabezas, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET. *IEEE Access*, 7:25171–25184, 2019.

[171] Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S De Vries, Manon JNL Benders, and Ivana Išgum. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261, 2016.

[172] Liya Zhao and Kebin Jia. Multiscale CNNs for brain tumor segmentation and diagnosis. *Computational and Mathematical Methods in Medicine*, 2016, 2016.

[173] Raghav Mehta, Aabhas Majumdar, and Jayanthi Sivaswamy. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. *Journal of Medical Imaging*, 4(2):024003, 2017.

[174] Saddam Hussain, Syed Muhammad Anwar, and Muhammad Majid. Brain tumor segmentation using cascaded deep convolutional neural network. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1998–2001. IEEE, 2017.

[175] Yufan He, Aaron Carass, Yeyi Yun, Can Zhao, Bruno M Jedynak, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Towards topological correct segmentation of macular OCT from cascaded FCNs. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 202–209. Springer, 2017.

[176] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

[177] Rens Janssens, Guodong Zeng, and Guoyan Zheng. Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks. In *IEEE International Symposium on Biomedical Imaging*, pages 893–897. IEEE, 2018.

[178] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3D fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018.

[179] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, 2018.

[180] Manhua Liu, Danni Cheng, Kundong Wang, Yaping Wang, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics*, 16(3-4):295–308, 2018.

[181] Khosro Bahrami, Islem Rekik, Feng Shi, and Dinggang Shen. Joint reconstruction and segmentation of 7T-like MR images from 3T MRI based on cascaded convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 764–772. Springer, 2017.

[182] Christian Wachinger, Martin Reuter, and Tassilo Klein. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018.

[183] Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1):1–17, 2018.

[184] G Anthony Reina, Ravi Panchumarthy, Siddhesh Pravin Thakur, Alexei Bastidas, and Spyridon Bakas. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. *Frontiers in Neuroscience*, 14:65, 2020.

[185] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Mayra Bergkamp, Joost Wissink, Jiri Obels, Karlijn Keizer, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, et al. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, 14:391–399, 2017.

[186] Herve Lombaert, Jon Sporring, and Kaleem Siddiqi. Diffeomorphic spectral matching of cortical surfaces. In *International Conference on Information Processing in Medical Imaging*, pages 376–389. Springer, 2013.

[187] Christian Wachinger, Matthew Brennan, G Sharp, and Polina Golland. On the importance of location and features for the patch-based segmentation of parotid glands. In *MICCAI Workshop on Image-Guided Adaptive Radiation Therapy*, 2014.

[188] Christian Wachinger, Matthew Brennan, Greg C Sharp, and Polina Golland. Efficient descriptor-based segmentation of parotid glands with nonlocal means. *IEEE Transactions on Biomedical Engineering*, 64(7):1492–1502, 2016.

[189] Haozhe Jia, Yong Xia, Yang Song, Weidong Cai, Michael Fulham, and David Dagan Feng. Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging. *Neurocomputing*, 275:1358–1369, 2018.

[190] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[191] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 1–11. Springer, 2018.

[192] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE Transactions on Medical Imaging*, 38(4):1016–1025, 2018.

[193] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[194] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

[195] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

[196] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[197] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 516–524. Springer, 2017.

[198] Kaisar Kushibar, Sergi Valverde, Sandra González-Villà, Jose Bernal, Mariano Cabezas, Arnau Oliver, and Xavier Lladó. Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Scientific Reports*, 9(1):1–15, 2019.

[199] Sergi Valverde, Mostafa Salem, Mariano Cabezas, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21:101638, 2019.

[200] Claudia Mazo, Jose Bernal, Maria Trujillo, and Enrique Alegre. Transfer learning for classification of cardiovascular tissues in histological images. *Computer Methods and Programs in Biomedicine*, 165:69–76, 2018.

[201] Deisy Chaves, Laura Fernández-Robles, Jose Bernal, Enrique Alegre, and Maria Trujillo. Automatic characterisation of chars from the combustion of pulverised coals using machine vision. *Powder Technology*, 338:110–118, 2018.

[202] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[203] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.

[204] Valeriu Popescu, Marco Battaglini, WS Hoogstrate, Sander CJ Verfaillie, IC Sluimer, Ronald A van Schijndel, Bob W van Dijk, Keith S Cover, Dirk L Knol, Mark Jenkinson, et al. Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. *NeuroImage*, 61(4):1484–1494, 2012.

[205] Jong-Min Lee, Uicheul Yoon, Sang Hee Nam, Jung-Hyun Kim, In-Young Kim, and Sun I Kim. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. *Computers in Biology and Medicine*, 33(6):495–507, 2003.

[206] Suresh A Sadananthan, Weili Zheng, Michael WL Chee, and Vitali Zagorodnov. Skull stripping using graph cuts. *NeuroImage*, 49(1):225–239, 2010.

[207] Mark Jenkinson, Mickael Pechaud, Stephen Smith, et al. BET2: MR-based estimation of brain, skull and scalp surfaces. In *Annual Meeting of the Organization for Human Brain Mapping*, volume 17, page 167. OHBM, 2005.

[208] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, 2011.

[209] Feng Shi, Li Wang, Yakang Dai, John H Gilmore, Weili Lin, and Dinggang Shen. LABEL: pediatric brain extraction using learning-based meta-algorithm. *NeuroImage*, 62(3):1975–1986, 2012.

[210] Simon F Eskildsen, Pierrick Coupé, Vladimir Fonov, José V Manjón, Kelvin K Leung, Nicolas Guizard, Shafik N Wassef, Lasse Riis Østergaard, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. BEaST: brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362–2373, 2012.

[211] László G Nyúl and Jayaram K Udupa. On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999.

[212] Mohak Shah, Yiming Xiao, Nagesh Subbanna, Simon Francis, Douglas L Arnold, D Louis Collins, and Tal Arbel. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis*, 15(2):267–282, 2011.

[213] Jacob C Reinhold, Blake E Dewey, Aaron Carass, and Jerry L Prince. Evaluating the impact of intensity normalization on MR image synthesis. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109493H. International Society for Optics and Photonics, 2019.

[214] Jean-Philippe Fortin, Elizabeth M Sweeney, John Muschelli, Ciprian M Crainiceanu, Russell T Shinohara, Alzheimer's Disease Neuroimaging Initiative, et al. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132:198–212, 2016.

[215] John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998.

[216] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

[217] Maria del C Valdés Hernández, Victor González-Castro, Dina T Ghandour, Xin Wang, Fergus Doubal, Susana Muñoz Maniega, Paul A Armitage, and Joanna M Wardlaw. On the computational assessment of white matter hyperintensity progression: difficulties in method selection and bias field correction performance on images with significant white matter pathology. *Neuroradiology*, 58(5):475–485, 2016.

[218] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):1–13, 2018.

[219] Sitara Afzal, Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, Khalid M Awan, Irfan Mehmood, and Oh-Young Song. A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. *IEEE Access*, 7:115528–115539, 2019.

[220] Samsuddin Ahmed, Kyu Yeong Choi, Jang Jae Lee, Byeong C Kim, Goo-Rak Kwon, Kun Ho Lee, and Ho Yub Jung. Ensembles of patch-based classifiers for diagnosis of Alzheimer diseases. *IEEE Access*, 7:73373–73383, 2019.

[221] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, Alzheimer's Disease Neuroimaging Initiative, et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, 21:101645, 2019.

[222] Ahmad Chaddad, Matthew Toews, Christian Desrosiers, and Tamim Niazi. Deep radiomic analysis based on modeling information flow in convolutional neural networks. *IEEE Access*, 7:97242–97252, 2019.

[223] Rafael Ceschin, Alexandria Zahner, William Reynolds, Jenna Gaesser, Giulio Zuccoli, Cecilia W Lo, Vanathi Gopalakrishnan, and Ashok Panigrahy. A computational framework for the detection of subcortical brain dysmaturation in neonatal MRI using 3D convolutional neural networks. *NeuroImage*, 178:183–197, 2018.

[224] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U Brandt, Klemens Ruprecht, René M Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John-Dylan Haynes, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clinical*, 24:102003, 2019.

[225] Chiyu Feng, Ahmed Elazab, Peng Yang, Tianfu Wang, Feng Zhou, Huoyou Hu, Xiaohua Xiao, and Baiying Lei. Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM. *IEEE Access*, 7:63605–63618, 2019.

[226] Mark S Graham, Ivana Drobnjak, and Hui Zhang. A supervised learning approach for diffusion MRI quality control with minimal training data. *NeuroImage*, 178:668–676, 2018.

[227] Heng Huang, Xintao Hu, Yu Zhao, Milad Makkie, Qinglin Dong, Shijie Zhao, Lei Guo, and Tianming Liu. Modeling task fMRI data via deep convolutional autoencoder. *IEEE Transactions on Medical Imaging*, 37(7):1551–1561, 2017.

[228] Dewen Hu, Zhiguo Luo, and Longfei Zhao. Gender identification based on human brain structural MRI with a multi-layer 3D convolution extreme learning machine. *Cognitive Computation and Systems*, 1(4):91–96, 2019.

[229] Tae-Eui Kam, Han Zhang, Zhicheng Jiao, and Dinggang Shen. Deep learning of static and dynamic brain functional networks for early MCI detection. *IEEE Transactions on Medical Imaging*, 2019.

[230] Eunho Lee, Jun-Sik Choi, Minjeong Kim, Heung-Il Suk, Alzheimer's Disease Neuroimaging Initiative, et al. Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *NeuroImage*, 202:116113, 2019.

[231] Mingxia Liu, Jun Zhang, Dong Nie, Pew-Thian Yap, and Dinggang Shen. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1476–1485, 2018.

[232] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, 66(5):1195–1206, 2018.

[233] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical Image Analysis*, 43:157–168, 2018.

[234] Hongming Li and Yong Fan. Interpretable, highly accurate brain decoding of subtly distinct brain states from functional MRI using intrinsic functional networks and long short-term memory recurrent neural networks. *NeuroImage*, 202:116059, 2019.

[235] Melika Maleki, M Teshnehlab, and M Nabavi. Diagnosis of multiple sclerosis (MS) using convolutional neural network (CNN) from MRIs. *Global Journal of Medicinal Plant Research*, 1(1):50–54, 2012.

[236] Shervin Minaee, Yao Wang, Alp Aygar, Sohae Chung, Xiuyuan Wang, Yvonne W Lui, Els Fieremans, Steven Flanagan, and Joseph Rath. MTBI identification from diffusion MR images using bag of adversarial visual features. *IEEE Transactions on Medical Imaging*, 38(11):2545–2555, 2019.

[237] Fujia Ren, Chenhui Yang, Qi Qiu, Nianyin Zeng, Chunting Cai, Chaoqun Hou, and Quan Zou. Exploiting discriminative regions of brain slices based on 2D CNNs for Alzheimer's disease classification. *IEEE Access*, 2019.

[238] Saman Sarraf, Danielle D Desouza, John AE Anderson, and Cristina Saverino. MCADNNet: Recognizing stages of cognitive impairment through efficient convolutional fMRI and MRI neural network topology models. *IEEE Access*, 7:155584–155600, 2019.

[239] Hossam H Sultan, Nancy M Salem, and Walid Al-Atabany. Multi-classification of brain tumor images using deep neural network. *IEEE Access*, 7:69215–69225, 2019.

[240] Zijian Wang, Yaoru Sun, Qianzi Shen, and Lei Cao. Dilated 3D convolutional neural networks for brain MRI data classification. *IEEE Access*, 7:134388–134398, 2019.

[241] Lin Yuan, Xue Wei, Hui Shen, Ling-Li Zeng, and Dewen Hu. Multi-center brain imaging classification using a novel 3D CNN approach. *IEEE Access*, 6:49925–49934, 2018.

[242] Lulu Yue, Xiaoliang Gong, Jie Li, Hongfei Ji, Maozhen Li, and Asoke K Nandi. Hierarchical feature extraction for early Alzheimer's disease diagnosis. *IEEE Access*, 7:93752–93760, 2019.

[243] Shu Zhang, Huan Liu, Heng Huang, Yu Zhao, Xi Jiang, Brook Bowers, Lei Guo, Xiaoping Hu, Mar Sanchez, and Tianming Liu. Deep learning models unveiled functional difference between cortical gyri and sulci. *IEEE Transactions on Biomedical Engineering*, 66(5):1297–1308, 2018.

[244] Yu Zhao, Qinglin Dong, Shu Zhang, Wei Zhang, Hanbo Chen, Xi Jiang, Lei Guo, Xintao Hu, Junwei Han, and Tianming Liu. Automatic recognition of fMRI-derived functional networks using 3-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 65(9):1975–1984, 2017.

[245] Liang Zou, Jiannan Zheng, Chunyan Miao, Martin J Mckeown, and Z Jane Wang. 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. *IEEE Access*, 5:23626–23636, 2017.

[246] Mikael Agn, Per Munck af Rosenschöld, Oula Puonti, Michael J Lundemann, Laura Mancini, Anastasia Papadaki, Steffi Thust, John Ashburner, Ian Law, and Koen Van Leemput. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Medical Image Analysis*, 54:220–237, 2019.

[247] Shahab Aslani, Michael Dayan, Loredana Storelli, Massimo Filippi, Vittorio Murino, Maria A Rocca, and Diego Sona. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196:1–15, 2019.

[248] Hans E Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M Ellingsen. SegAE: unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder. *NeuroImage: Clinical*, 24:102085, 2019.

[249] Siqi Bao, Pei Wang, Tony CW Mok, and Albert CS Chung. 3D randomized connection network with graph-based label inference. *IEEE Transactions on Image Processing*, 27(8):3883–3892, 2018.

[250] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. DRINet for medical image segmentation. *IEEE Transactions on Medical Imaging*, 37(11):2453–2462, 2018.

[251] Alexander de Brebisson and Giovanni Montana. Deep neural networks for anatomical brain segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 20–28, 2015.

[252] Yang Deng, Yao Sun, Yongpei Zhu, Yue Xu, Qianxi Yang, Shuo Zhang, Zhanyu Wang, Jirang Sun, Weiling Zhao, Xiaobo Zhou, et al. A new framework to reduce doctor's workload for medical image annotation. *IEEE Access*, 7:107097–107104, 2019.

[253] Yi Ding, Chang Li, Qiqi Yang, Zhen Qin, and Zhiguang Qin. How to improve the deep residual network to segment multi-modal brain tumor images. *IEEE Access*, 7:152821–152831, 2019.

[254] Yi Ding, Fujuan Chen, Yang Zhao, Zhixing Wu, Chao Zhang, and Dongyuan Wu. A stacked multi-connection simple reducing net for brain tumor segmentation. *IEEE Access*, 7:104011–104024, 2019.

[255] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2018.

[256] Joseph Enguehard, Peter O'Halloran, and Ali Gholipour. Semi-supervised learning with deep embedded clustering for image classification and segmentation. *IEEE Access*, 7:11093–11104, 2019.

[257] Beibei Hou, Guixia Kang, Ningbo Zhang, and Chuan Hu. Robust 3D convolutional neural network with boundary correction for accurate brain tissue segmentation. *IEEE Access*, 6:75471–75481, 2018.

[258] Kai Hu, Qinghai Gan, Yuan Zhang, Shuhua Deng, Fen Xiao, Wei Huang, Chunhong Cao, and Xieping Gao. Brain tumor segmentation using multi-cascaded convolutional neural networks and conditional random field. *IEEE Access*, 7:92615–92629, 2019.

[259] Yuankai Huo, Zhoubing Xu, Yunxi Xiong, Katherine Aboud, Prasanna Parvathaneni, Shunxing Bao, Camilo Bermudez, Susan M Resnick, Laurie E Cutting, and Bennett A Landman. 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119, 2019.

[260] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[261] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In *International Conference on Information Processing in Medical Imaging*, pages 348–360. Springer, 2017.

[262] Ming Li, Lishan Kuang, Shuhua Xu, and Zhanguo Sha. Brain tumor detection based on multimodal information fusion and convolutional neural network. *IEEE Access*, 7:180134–180146, 2019.

[263] Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, 183:650–665, 2018.

[264] Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötzel, et al. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding*, 164:92–102, 2017.

[265] Pim Moeskops, Jeroen de Bresser, Hugo J Kuijf, Adriënne M Mendrik, Geert Jan Biessels, Josien PW Pluim, and Ivana Išgum. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. *NeuroImage: Clinical*, 17:251–262, 2018.

[266] Melanie A Morrison, Seyedmehdi Payabvash, Yicheng Chen, Sivakami Avadiappan, Mihir Shah, Xiaowei Zou, Christopher P Hess, and Janine M Lupo. A user-guided tool for semi-automated cerebral microbleed detection and volume segmentation: Evaluating vascular injury and data labelling for machine learning. *NeuroImage: Clinical*, 20:498–505, 2018.

[267] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, 2020.

[268] Dong Nie, Li Wang, Ehsan Adeli, Cuijin Lao, Weili Lin, and Dinggang Shen. 3-D fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE Transactions on Cybernetics*, 49(3):1123–1136, 2018.

[269] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, 2016.

[270] Sérgio Pereira, Adriano Pinto, Joana Amorim, Alexandrine Ribeiro, Victor Alves, and Carlos A Silva. Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks. *IEEE Transactions on Medical Imaging*, 38(12):2914–2925, 2019.

[271] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36(2):674–683, 2016.

[272] Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(5):1911–1919, 2018.

[273] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186:713–727, 2019.

[274] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 36(11):2319–2330, 2017.

[275] Markus D Schirmer, Adrian V Dalca, Ramesh Sridharan, Anne-Katrin Giese, Kathleen L Donahue, Marco J Nardin, Steven JT Mocking, Elissa C McIntosh, Petrea Frid, Johan Wasselius, et al. White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts–the MRI-GENIE study. *NeuroImage: Clinical*, 23:101884, 2019.

[276] Mahsa Shakeri, Stavros Tsogkas, Enzo Ferrante, Sarah Lippe, Samuel Kadoury, Nikos Paragios, and Iasonas Kokkinos. Sub-cortical brain structure segmentation using F-CNN's. In *IEEE International Symposium on Biomedical Imaging*, pages 269–272. IEEE, 2016.

[277] Muhan Shao, Shuo Han, Aaron Carass, Xiang Li, Ari M Blitz, Jaehoon Shin, Jerry L Prince, and Lotta M Ellingsen. Brain ventricle parcellation using a deep neural network: Application to patients with ventriculomegaly. *NeuroImage: Clinical*, 23:101871, 2019.

[278] Feng Shi, Qi Yang, Xiuhai Guo, Touseef Ahmad Qureshi, Zixiao Tian, Huijuan Miao, Damini Dey, Debiao Li, and Zhaoyang Fan. Intracranial vessel wall segmentation using convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 66(10):2840–2847, 2019.

[279] Marijn F Stollenga, Wonmin Byeon, Marcus Liwicki, and Juergen Schmidhuber. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In *Advances in Neural Information Processing Systems*, pages 2998–3006, 2015.

[280] Benjamin Thyreau, Kazunori Sato, Hiroshi Fukuda, and Yasuyuki Taki. Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. *Medical Image Analysis*, 43:214–228, 2018.

[281] Liansheng Wang, Cong Xie, and Nianyin Zeng. RP-Net: a 3D convolutional neural network for brain segmentation from magnetic resonance imaging. *IEEE Access*, 7:39670–39679, 2019.

[282] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2018.

[283] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, 2018.

[284] Yunzhe Xue, Fadi G Farhat, Olga Boukrina, AM Barrett, Jeffrey R Binder, Usman W Roshan, and William W Graves. A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images. *NeuroImage: Clinical*, 25:102118, 2020.

[285] Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Konstantinos Kamnitsas, Athanasios Vlontzos, Ben Glocker, Bernhard Kainz, et al. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical Image Analysis*, 53:156–164, 2019.

[286] Abol Basher, Kyu Yeong Choi, Jang Jae Lee, Bumshik Lee, Byeong C Kim, Kun Ho Lee, and Ho Yub Jung. Hippocampus localization using a two-stage ensemble hough convolutional neural network. *IEEE Access*, 7:73436–73447, 2019.

[287] Laurent Chauvin, Kuldeep Kumar, Christian Wachinger, Marc Vangel, Jacques de Guise, Christian Desrosiers, William Wells, Matthew Toews, Alzheimer's Disease Neuroimaging Initiative, et al. Neuroimage signature from salient keypoints is highly specific to individuals and shared by close relatives. *NeuroImage*, 204:116208, 2020.

[288] Saifeng Liu, David Utriainen, Chao Chai, Yongsheng Chen, Lin Wang, Sean K Sethi, Shuang Xia, and E Mark Haacke. Cerebral microbleed detection using susceptibility weighted imaging and deep learning. *NeuroImage*, 198:271–282, 2019.

[289] Richard McKinley, Rik Wepfer, Lorenz Grunder, Fabian Aschwanden, Tim Fischer, Christoph Friedli, Raphaela Muri, Christian Rummel, Rajeev Verma,

Christian Weisstanner, et al. Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *NeuroImage: Clinical*, 25:102104, 2020.

[290] Jun Zhang, Mingxia Liu, and Dinggang Shen. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Transactions on Image Processing*, 26(10):4753–4764, 2017.

[291] Steffen Bollmann, Kasper Gade Bøtker Rasmussen, Mads Kristensen, Rasmus Guldhammer Blendal, Lasse Riis Østergaard, Maciej Plocharski, Kieran O'Brien, Christian Langkammer, Andrew Janke, and Markus Barth. DeepQSM-using deep learning to solve the dipole inversion for quantitative susceptibility mapping. *NeuroImage*, 195:373–383, 2019.

[292] James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

[293] Florian Dubost, Pinar Yilmaz, Hieab Adams, Gerda Bortsova, M Arfan Ikram, Wiro Niessen, Meike Vernooij, and Marleen de Bruijne. Enlarged perivascular spaces in brain MRI: Automated quantification in four regions. *NeuroImage*, 185:534–544, 2019.

[294] Zhiwei Li, Ting Gong, Zhichao Lin, Hongjian He, Qiqi Tong, Chen Li, Yi Sun, Feng Yu, and Jianhui Zhong. Fast and robust diffusion kurtosis parametric mapping using a three-dimensional convolutional neural network. *IEEE Access*, 7:71398–71411, 2019.

[295] Gustav Mårtensson, Daniel Ferreira, Lena Cavallin, J-Sebastian Muehlboeck, Lars-Olof Wahlund, Chunliang Wang, Eric Westman, Alzheimer's Disease Neuroimaging Initiative, et al. AVRA: automatic visual ratings of atrophy from MRI images using recurrent convolutional neural networks. *NeuroImage: Clinical*, 23:101872, 2019.

[296] Seyed Sadegh Mohseni Salehi, Shadab Khan, Deniz Erdogmus, and Ali Gholipour. Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. *IEEE Transactions on Medical Imaging*, 38(2):470–481, 2018.

[297] Hongjiang Wei, Steven Cao, Yuyao Zhang, Xiaojun Guan, Fuhua Yan, Kristen W Yeom, and Chunlei Liu. Learning-based single-step quantitative susceptibility mapping reconstruction without brain extraction. *NeuroImage*, 202:116064, 2019.

[298] Jaeyeon Yoon, Enhao Gong, Itthi Chatnuntawech, Berkin Bilgic, Jingu Lee, Woojin Jung, Jingyu Ko, Hosan Jung, Kawin Setsompop, Greg Zaharchuk, et al. Quantitative susceptibility mapping using deep neural network: QSM-net. *NeuroImage*, 179:199–206, 2018.

[299] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[300] Changhee Han, Leonardo Rundo, Ryosuke Araki, Yudai Nagano, Yujiro Furukawa, Giancarlo Mauri, Hideki Nakayama, and Hideaki Hayashi. Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. *IEEE Access*, 7:156966–156977, 2019.

[301] Lei Xiang, Qian Wang, Dong Nie, Lichi Zhang, Xiyao Jin, Yu Qiao, and Dinggang Shen. Deep embedding convolutional neural network for synthesizing CT image from T1-weighted MR image. *Medical Image Analysis*, 47:31–44, 2018.

[302] Yoonmi Hong, Jaeil Kim, Geng Chen, Weili Lin, Pew-Thian Yap, and Dinggang Shen. Longitudinal prediction of infant diffusion MRI data via graph convolutional adversarial networks. *IEEE Transactions on Medical Imaging*, 38(12):2717–2725, 2019.

[303] Jia Liu, Fang Chen, Changcun Pan, Mingyu Zhu, Xinran Zhang, Liwei Zhang, and Hongen Liao. A cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas. *IEEE Transactions on Biomedical Engineering*, 65(9):1943–1952, 2018.

[304] Euijin Jung, Philip Chikontwe, Xiaopeng Zong, Weili Lin, Dinggang Shen, and Sang Hyun Park. Enhancement of perivascular spaces using densely connected deep convolutional neural network. *IEEE Access*, 7:18382–18391, 2019.

[305] Maosong Ran, Jinrong Hu, Yang Chen, Hu Chen, Huaiqiang Sun, Jiliu Zhou, and Yi Zhang. Denoising of 3D magnetic resonance images using a residual encoder–decoder Wasserstein generative adversarial network. *Medical Image Analysis*, 55:165–180, 2019.

[306] Jinglong Du, Lulu Wang, Yulu Liu, Zexun Zhou, Zhongshi He, and Yuanyuan Jia. Brain MRI super-resolution using 3D dilated convolutional encoder–decoder network. *IEEE Access*, 8:18938–18950, 2020.

[307] Jiaqi Gu, Zeju Li, Yuanyuan Wang, Haowei Yang, Zhongwei Qiao, and Jinhua Yu. Deep generative adversarial networks for thin-section infant MR image reconstruction. *IEEE Access*, 7:68290–68304, 2019.

[308] Zar Nawab Khan Swati, Qinghua Zhao, Muhammad Kabir, Farman Ali, Zakir Ali, Saeed Ahmed, and Jianfeng Lu. Content-based brain tumor retrieval for MR images using transfer learning. *IEEE Access*, 7:17809–17822, 2019.

[309] Hákon Gudbjartsson and Samuel Patz. The Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995.

[310] David Moratal, A Vallés-Luch, Luis Martí-Bonmatí, and Marijn E Brummer. k-Space tutorial: an MRI educational tool for a better understanding of k-space. *Biomedical Imaging and Intervention Journal*, 4(1), 2008.

[311] Maxim Zaitsev, Julian Maclaren, and Michael Herbst. Motion artifacts in MRI: a complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 42(4):887–901, 2015.

[312] Richard Shaw, Carole Sudre, Sebastien Ourselin, and M Jorge Cardoso. MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty. In *International Conference on Medical Imaging with Deep Learning*, pages 427–436, 2019.

[313] K Sommer, A Saalbach, T Brosch, C Hall, NM Cross, and JB Andre. Correction of motion artifacts using a multiscale fully convolutional neural network. *American Journal of Neuroradiology*, 2020.

[314] Daiki Tamada, Marie-Luise Kromrey, Shintaro Ichikawa, Hiroshi Onishi, and Utaroh Motosugi. Motion artifact reduction using a convolutional neural network for dynamic contrast enhanced mr imaging of the liver. *Magnetic Resonance in Medical Sciences*, 19(1):64–76, 2020.

[315] José V Manjón and Pierrick Coupe. MRI denoising using deep learning. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 12–19. Springer, 2018.

[316] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6428–6436, 2017.

[317] Xin Yi and Paul Babyn. Sharpness-aware low-dose CT denoising using conditional generative adversarial network. *Journal of Digital Imaging*, 31(5):655–669, 2018.

[318] Johannes Rieke, Fabian Eitel, Martin Weygandt, John-Dylan Haynes, and Kerstin Ritter. Visualizing convolutional networks for MRI-based diagnosis of

Alzheimer's disease. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 24–31. Springer, 2018.

[319] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[320] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[321] Albert Clèrigues, Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, and Xavier Lladó. SUNet: a deep learning architecture for acute stroke lesion segmentation and outcome prediction in multimodal MRI. *coRR*, abs/1810.13304, 2018.

[322] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jia-Wei Chen, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: The iseg-2017 challenge. *IEEE transactions on medical imaging*, 38(9):2219–2230, 2019.

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

[324] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[325] M.-P. Dubuisson and A. K. Jain. A modified Hausdorff distance for object matching. In *IAPR International Conference on Pattern Recognition*, volume 1, pages 566–568. IEEE, 1994.

[326] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[327] Pedro O Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *International Conference on International Conference on Machine Learning*, volume 32, pages 82–90. JMLR.org, 2014.

[328] Hongsheng Li, Rui Zhao, and Xiaogang Wang. Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification. *coRR*, abs/1412.4526, 2014.

[329] Jared Hamwood, David Alonso-Caneiro, Scott A Read, Stephen J Vincent, and Michael J Collins. Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers. *Biomedical optics express*, 9(7):3049–3066, 2018.

[330] Jose Bernal, Mostafa Salem, Kaisar Kushibar, Albert Clèrigues, Sergi Valverde, Mariano Cabezas, Sandra Gonzáles-Villa, Joaquim W Salvi, Arnau Oliver, and Xavier Lladó. MR brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors. `http://mrbrains18.isi.uu.nl/wp-content/uploads/2018/11/nic_vicorob.pdf`, 2018. Accessed: 26-02-2020.

[331] Li Wang, Feng Shi, Yaozong Gao, Gang Li, Weili Lin, and Dinggang Shen. Isointense infant brain segmentation by stacked kernel canonical correlation analysis. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 28–36. Springer, 2015.

[332] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623, 2012.

[333] Jose Bernal, Mostafa Salem, Kaisar Kushibar, Albert Clerigues, Sergi Valverde, Mariano Cabezas, Sandra Gonzáles-Villa, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. MR brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors. `http://mrbrains18.isi.uu.nl/wp-content/uploads/2018/11/nic_vicorob.pdf`. Accessed: 2019-09-01.

[334] Renske de Boer, Henri A Vrooman, M Arfan Ikram, Meike W Vernooij, Monique MB Breteler, Aad van der Lugt, and Wiro J Niessen. Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *NeuroImage*, 51(3):1047–1056, 2010.

[335] Bilge Karaçali and Christos Davatzikos. Simulation of tissue atrophy using a topology preserving transformation model. *IEEE Transactions on Medical Imaging*, 25(5):649–652, 2006.

[336] Konstantin Ens, Fabian Wenzel, Stewart Young, Jan Modersitzki, and Bernd Fischer. Design of a synthetic database for the validation of non-linear registration and segmentation of magnetic resonance brain images. In *Medical Imaging 2009: Image Processing*, volume 7259, page 725933. International Society for Optics and Photonics, 2009.

[337] Snehashis Roy, Aaron Carass, and Jerry L Prince. Magnetic resonance image example-based contrast synthesis. *IEEE Transactions on Medical Imaging*, 32(12):2348–2363, 2013.

[338] Swati Sharma, François Rousseau, Fabrice Heitz, Lucien Rumbach, and Jean-Paul Armspach. On the estimation and correction of bias in local atrophy estimations using example atrophy simulations. *Computerized Medical Imaging and Graphics*, 37(7-8):538–551, 2013.

[339] Bishesh Khanal, Nicholas Ayache, and Xavier Pennec. Simulating longitudinal brain MRIs with known volume changes and realistic variations in image intensity. *Frontiers in Neuroscience*, 11:132, 2017.

[340] Wen Wei, Emilie Poirion, Benedetta Bodini, Stanley Durrleman, Olivier Colliot, Bruno Stankoff, and Nicholas Ayache. FLAIR MR image synthesis by using 3D fully convolutional networks for multiple sclerosis. In *ISMRM-ESMRMB 2018 - Joint Annual Meeting*, 2018.

[341] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, 2017.

[342] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[343] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[344] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE Transactions on Medical Imaging*, 38(9):2165–2176, 2019.

[345] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000.

[346] Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration aka Spatial normalisation FMRIB Technial Report TR07JA2. *FMRIB Analysis Group of the University of Oxford*, 2007.

[347] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

[348] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

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

[350] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[351] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

[352] Guoying Li. Robust regression. *Exploring data tables, trends, and shapes*, 281:U340, 1985.

[353] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.

[354] Hyunwoo Lee, Kunio Nakamura, Sridar Narayanan, Robert A Brown, Douglas L Arnold, Alzheimer's Disease Neuroimaging Initiative, et al. Estimating and accounting for the effect of MRI scanner changes on longitudinal whole-brain volume change measurements. *NeuroImage*, 184:555–565, 2019.

[355] Kunio Nakamura, Stephen Jones, Wim Van Hecke, Douglas Arnold, Carl de Moor, Carrie Wager, Dominique Jennings, Nancy Richert, Richard Rudick, Jeffrey Cohen, et al. Comparison of brain atrophy measurement techniques in a longitudinal study of multiple sclerosis patients with frequent MRIs (p4. 376), 2017.

[356] Nicolas Guizard, Vladimir S Fonov, Daniel García-Lorenzo, Kunio Nakamura, Bérengère Aubert-Broche, and D Louis Collins. Spatio-temporal regularization for longitudinal registration to subject-specific 3D template. *PloS one*, 10(8), 2015.

[357] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, 57:226–236, 2019.

[358] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[359] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[360] Stephen M Smith, Anil Rao, Nicola De Stefano, Mark Jenkinson, Jonathan M Schott, Paul M Matthews, and Nick C Fox. Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: cross-validation of BSI, SIENA and SIENAX. *NeuroImage*, 36(4):1200–1206, 2007.

[361] Kunio Nakamura, Robert A Brown, David Araujo, Sridar Narayanan, and Douglas L Arnold. Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: implications for monitoring atrophy in clinical studies. *NeuroImage: Clinical*, 6:166–170, 2014.

[362] Michael G Dwyer, Diego Silva, Niels Bergsland, Dana Horakova, Deepa Ramasamy, Jaqueline Durfee, Manuela Vaneckova, Eva Havrdova, and Robert Zivadinov. Neurological software tool for reliable atrophy measurement (NeuroSTREAM) of the lateral ventricles on clinical-quality T2-FLAIR MRI scans in multiple sclerosis. *NeuroImage: Clinical*, 15:769–779, 2017.

[363] Niels Bergsland, Dana Horakova, Michael G Dwyer, Tomas Uher, Manuela Vaneckova, Michaela Tyblova, Zdenek Seidl, Jan Krasensky, Eva Havrdova, and Robert Zivadinov. Gray matter atrophy patterns in multiple sclerosis: A 10-year source-based morphometry study. *NeuroImage: Clinical*, 17:444–451, 2018.

[364] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[365] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9252–9260, 2018.

[366] Richard Shaw, Carole H Sudre, Thomas Varsavsky, Sébastien Ourselin, and M Jorge Cardoso. A k-space model of movement artefacts: Application to segmentation augmentation and artefact removal. *IEEE Transactions on Medical Imaging*, 2020.