UAB
Universitat Autònoma de Barcelona

Deep learning applied to brain MRI data for patient stratification and prediction of multiple sclerosis disease course

Llucia Coll Benejam

UAB
Universitat Autònoma de Barcelona

PhD Thesis

Deep learning applied to brain MRI data for patient stratification and prediction of multiple sclerosis disease course

Llucia Coll Benejam

Doctoral Program in Bioinformatics

Directors:
Dr. Carmen Tur
Dr. Deborah Pareto
Prof. Xavier Lladó

Academic Tutor: Dr. Albert Ruiz

Vall d'Hebron Institut de Recerca
Universitat Autònoma de Barcelona

2023

Cemcat Centre d'Esclerosi Múltiple de Catalunya

Vall d'Hebron Institut de Recerca

A les incansables,
a les que lluiten,
a les que curen.
A tu,
mamà.

Acknowledgments

Per qui no em coneix, som algú qui sempre cerca escriure ses paraules exactes. Com si existissin. Supòs que un poc són açò, aquests agraïments, molts intents de paraules darrere un únic sentiment: gratitud.

Primer de tot m'agradaria agrair la feina feta pels meus directors, Carmen, Deborah i Xavi, ja que sense vosaltres això no hauria estat possible. Gràcies per l'oportunitat i la confiança que això sortís bé.
To the Stack Overflow, and the whole anonymous community contributing to and believing in open and shared science, thank you for helping me find the light in my bugs (and my aesthetic requirements). I també a tu, Albert, per ajudar-me a trobar-hi solució en moments de desesperació, encara que fos plantejant(-nos) encara més preguntes. Lili, aunque sigas diciendo que a ti de esto no te pregunte, que "no sabes de esto", sabes que seguiré haciéndolo, porque tu apoyo y tus persistentes correcciones han estado siempre ahí. Gracias, amiga.
Al *lab*, por acogerme y hacerme sentir una más de vosotros: entre les bandes de na Mireia, la cosa larga de María, el R de Javi, els ferments de n'Arnau, Rux en física, els simoa de na Lucía, las historietas de Nico y los avisos de comida de Sunny. A las juniors i també, a altra gent polida de collserola. Pero especialmente a mis homólogas, María y Rux, bueno y Javi también, porque sin vosotros esto no habría sido lo mismo.
A tot ViCOROB, per rebre'm com si mai hagués partit. Arnau, merci per no ser només un membre més al camp de futbol. Uma and Valeria, thanks for revitalising the lab and welcoming me there any time.
Al Cemcat, amb el Xavier Montalban al capdavant, per la seva incansable tasca en l'estudi de l'esclerosi múltiple. Especialment, a la Mar, per donar-me el seu punt de vista directe i sincer i, al Pere, per l'interès i entusiasme envers el que

faig. Tampoc puc oblidar els (per jo) anònims contribuïdors d'aquesta tesi, els pacients (i les seves famílies). És el seu consentiment d'utilitzar les seves dades el motor de la nostra recerca.
I would also like to thank my former tutors, Raul, Alan and Xavi, who believed in me before getting to know me better, and gave me the tools to "navigate" in academia. Especialment a tu, Xavi, pels incontables sacs de suport i paciència, que mai podré agrair prou.

Més enllà de tot açò, de codi, de papers, de terminologies i altres històries, hi ha tota sa gent que ha estat i és, present o a distància, en aquest paral·lelisme vital a cavall entre Menorca i Barcelona, i encara no s'ha cansat de jo.
Ets amics de sa uni, que ja no sou es de sa uni, sou es meus amics sense adjectius. Per ets afterworks, ses jams, ses mudances, ses vostres converses zero des meu interès i, un llarg i variat etcètera. Aina, Aleix, Carlos, Elen, Glòria, Mariona, Oli i Víctor, merci per ser-hi. Clàudia i Andrea, a més dels deu anys d'anar fent camí juntes, què hauria fet sense sa mitja cadira de sa cuina de Puigmartí? Merci per ser es millor *squad* de confiança. Anna, no podria haver topat amb una *hamfri* millor amb qui conviure ses alegries i desgràcies d'es *màgic learning*. Es millors inventors de tradicions i missions impossibles, *La Societé*, es meus físics preferits i una neboda amb molta closca, per estar sempre a punt. Tòfol, Maria, Charlie, Joan, Martí, Dario, Adrià i Ànec, merci de tot cor.
Seria mentida dir que no ens uneix res en especific, perquè en cert moment (per separat) sa música ho va fer. Avançar cadascuna per es seu camí però compartint açò de fer-nos grans, quina sort sa meua! Merci, Clara's.
I tornar, sempre tornar. Gràcies a ses àvies per ser-ne motiu: s'àvia Aguedita, s'única que em pot dir en diminutiu es nom que s'àvia Llucia em va deixar (i no ha sigut a temps de veure'm acabar). Gràcies Benejam's. Gràcies Coll's. Gràcies *Dems*. Gràcies a ses cosines de ver i de mentida, Àgueda, Sara i Caterina, per ser casa, sigui on sigui.

Gràcies mamà i papà per confiar en jo, per deixar-me triar es meu propi camí i fer (quasi) sempre el que he volgut. Potser a vegades no m'ha sortit del tot bé, però saber que a n'es final, principi o moment *random* puc comptar amb voltros, és el que em fa mirar endavant.
I a tu, Talina, que havent arribat fins aquí, ja consideres que et pots repartir en tota sa resta de seccions. Gràcies per "obligar-me" a aprendre Latex i, que es teu

final de tesi inspirés es meu interès per inkscape; sense açò aquesta tesi no seria el mateix. De sa mateixa manera que, si no fossis sa meua germana (grossa, pes despistats), sa meua vida, tampoc. Merci, germana preferida.

En definitiva, merci a totes aquelles que en algun moment us heu interessat per com m'anava i, per qui encara creu que som de lletres, aquí teniu es meu primer llibre.

Barcelona, setembre 2023

Abstract

Multiple sclerosis (MS) is a chronic disease of the central nervous system characterised by inflammation, demyelination and neurodegeneration. MS is one of the main non-traumatic causes of irreversible disability in young adults, and its disease course is highly variable among individuals. In the clinical routine, magnetic resonance imaging (MRI) is an essential tool to help with the diagnosis and prognosis of MS. The combination of deep learning-based models with MRI has presented promising results in this field. However, the underlying predictive features of MS progression remain a subject of ongoing research.

In this PhD Thesis, we question whether deep learning models solely applied to structural brain MRI scans can be used to predict different disease course status of patients with MS and their potential applicability in clinical practice. To achieve this goal, we present two approaches for stratifying MS patients in cross-sectional studies and one prognostic study. For each proposal we include interpretability strategies to make the deep learning models more explainable and trustworthy to clinicians. An in-house cohort of MS patients prospectively followed over time after their first demyelinating attack has been the unifying thread of this PhD Thesis.

In the first study, we proposed a deep learning approach to perform a binary stratification of patients with MS based on their disability score assessed by a neurologist, building a model that only takes as input the whole brain MRI at a single time-point. Assessing the interpretability of the results obtained through attention maps, we observed the importance of the frontotemporal cortex and the cerebellum for the development of disability accumulation. In the subsequent study, we analysed the use of the same deep learning approach applied to different predefined regional inputs to compare them with the whole brain approach. These regions included white matter and grey matter tissues,

lateral ventricles, brain stem and cerebellum structures, and subcortical grey matter structures. Even though the grey matter regional model performed the best, when evaluating the model in an external validation dataset, the results suggested that the global approach offered the best and most robust performance for generalisation. Finally, we proposed a discrete-time survival model to predict the survival function of patients with a first demyelinating attack who will develop a first event of progression independent of relapse activity (PIRA), only using a brain MRI scan acquired at symptom onset as input. With the obtained results, we were able to improve the predictive power of a classical survival model built with the strongest PIRA predictor, the age at the first attack. Furthermore, we extracted attention maps that revealed the frontoparietal cortex as the most important anatomical region for the decisions made by the deep learning survival model.

In this PhD Thesis, we have successfully developed precise and automated deep learning models supported by explainability algorithms trying to reveal the focuses leading the output decisions. This work marks the initial steps of potential new predictive models aimed to improve the management of patients with MS, paving the path toward potential integration into routine clinical practice in the future.

Resum

L'esclerosi múltiple és una malaltia crònica del sistema nerviós central caracteritzada per la inflamació, la desmielinització i la neurodegeneració. L'esclerosi múltiple és una de les principals causes no traumàtiques de discapacitat irreversible en adults joves, la qual presenta un curs altament variable entre individus. En la pràctica clínica, l'ús d'imatges de ressonància magnètica ha esdevingut una eina essencial en la diagnosi i prognosi de l'esclerosi múltiple. En els últims anys, la combinació de models basats en l'aprenentatge profund (deep learning) amb imatges de ressonància magnètica ha presentat resultats prometedors en l'estudi de l'esclerosi múltiple. No obstant això, les característiques predictives subjacents de la progressió de l'esclerosi múltiple encara són desconegudes.

En aquesta tesi doctoral, qüestionem si els mètodes d'aprenentatge profund aplicats únicament a l'anàlisi de imatges cerebrals de ressonància magnètica poden ser utilitzats per predir diferents estats del curs de la malaltia en pacients amb esclerosi múltiple i, com podrien ser aplicats en la pràctica clínica. Per aconseguir aquest objectiu, presentem dues aproximacions per a l'estratificació de pacients amb esclerosi múltiple: dos estudis d'anàlisis transversals i un estudi de pronòstic. Per a cada proposta, incloem estratègies d'interpretació per a que els models d'aprenentatge profund siguin més explicatius i fiables des d'un punt de vista clínic. Tots els experiments que composen aquesta tesi doctoral han estat realitzats amb una cohort interna de pacients amb esclerosi múltiple seguits prospectivament en el temps després del seu primer brot desmielinitzant.

En el primer estudi, vam proposar un model d'aprenentatge profund per a l'estratificació binària de pacients amb esclerosi múltiple en funció de la seva puntuació de discapacitat avaluada per un neuròleg. El model presentat, únicament pren com a entrada imatges de ressonància magnètica cerebral en

un únic punt temporal en qualsevol estat del curs de la malaltia. En avaluar la interpretació dels resultats obtinguts mitjançant mapes d'atenció, vàrem observar una major importància en el còrtex frontotemporal i el cerebel en el desenvolupament de l'acumulació de discapacitat.

En el següent estudi, vàrem analitzar l'ús del mateix model d'aprenentatge profund utilitzant diferents dades d'entrada regionals, prèviament definides, per a comparar-les amb l'entrada global de tot el cervell. Aquestes regions incloïen els teixits de matèria blanca i matèria grisa, els ventricles laterals, el cerebel i tronc cerebral i les estructures subcorticals de matèria grisa. Tot i que el model regional de matèria grisa va obtenir el millor rendiment, en avaluar els models en un conjunt de dades d'una cohort externa, es va suggerir que l'enfocament global oferia un millor rendiment i més robust per a la generalització.

Finalment, vam proposar un model de supervivència en temps discret per predir la probabilitat de desenvolupar una primera progressió independent de l'activitat de brots (PIRA, per les sigles en anglès) a partir de la primera ressonància magnètica del cervell després del primer brot desmielinitzant. Amb els resultats obtinguts, vam ser capaços de millorar el poder predictiu d'un model de supervivència clàssic construït amb el predictor de PIRA més rellevant, l'edat en el primer brot.

En aquesta tesi doctoral, hem desenvolupat amb èxit models d'aprenentatge profund precisos i automatitzats amb el suport d'algoritmes d'interpretació que intenten revelar quines regions en el cervell dirigeixen les decisions obtingudes a partir del model. Aquest treball marca els passos inicials de potencials models predictius destinats a millorar la gestió dels pacients amb esclerosi múltiple, aplanant el camí cap a la seva potencial integració en la pràctica clínica en el futur.

Contents

List of Figures

List of Tables

Acronyms

2D/3D	Two/Three-dimensional
AUC	Area Under the Curve
BSC	Brain stem and cerebellum structures
CAM	Class Activation Mapping
CDA	Confirmed Disability Accumulation
CI	Confidence Interval
CIS	Clinically Isolated Syndrome
CNN	Convolutional Neural Network
CNS	Central Nervous System
CPH	Cox Proportional Hazard
CSF	Cerebrospinal Fluid
$\mathbf{c}^{td}$	Time-dependent Concordance Index
EDSS	Expanded Disability Status Score
FLAIR	Fluid Attenuated Inversion Recovery
FN	False Negative
FP	False Positive
FWHM	Full Width at Half Maximum
GAN	Generative Adversarial Network
GAP	Global adaptive pooling
GM	Grey matter
Grad-CAM	Gradient-weighted Class Activation Mapping
HR	Hazard Ratio
IBS	Integrated Brier Score
LRP	Layer-wise Relevance Propagation
LST	Lesion Segmentation Tool

MBConv	MoBile inverted bottlenecks
MNI	Montreal Imaging Initiative
MPRAGE	Magnetization-Prepared Rapid Acquistion with Gradient Echo
MRI	Magnetic Resnoance Imaging
MS	Multiple Sclerosis
MS PATHS	Multiple Sclerosis Partners Advancing Technology and Health Solutions
NMOSD	Neuromyelitits optica spectrum disorder
OB	Oligoclonal Bands
PDDS	Patient Determined Disability Score
PIRA	Progression Independent of Relapse Activity
PPMS	Primary Progressive Multiple Sclerosis
RAW	Relapse-Associated Worsening
ReLU	Rectified Linear Unit
ROC	Receiver Operative Characteristic Curve
ROI	Region-Of-Interest
RRMS	Relapsing Remitting Multiple Sclerosis
SDMT	Symbol Digit Modalities Test
SE	Squeeze-and-Excitation
SHAP	SHapley Additive exPlanations
SPMS	Secondary Progressive Multiple Sclerosis
SVM	Support Vector Machine
T1-w	T1-weighted
T2-w	T2-weighted
TE	Echo Time
TI	Inversion Time
TN	True negative
TP	True Positive
TR	Repetition Time
VHUH	Vall d'Hebron University Hospital
WM	White Matter

Chapter 1

Introduction

1.1 Multiple sclerosis

Multiple sclerosis (MS) is a chronic disease of the central nervous system (CNS) characterised by inflammation, demyelination and neurodegeneration, being one of the main non-traumatic causes of irreversible disability in young adults. The exact cause of MS and the pathological mechanisms ultimately leading to an irreversible accumulation of disability are still unknown. Furthermore, its disease course can be highly variable among individuals [1].

In people with MS, inflammation of the CNS and spinal cord leads to the development of lesions or plaques that harm the protective myelin sheaths surrounding nerve cells. These myelin sheaths play a vital role in the nervous system, acting as insulating layers around nerve fibers (axons) and facilitating the swift and efficient transmission of nerve impulses along the axon. The demyelination process disrupts axonal transmissions, resulting in cognitive decline and physical disability. Our knowledge about the neurodegenerative nature of MS is based on observations that axonal loss and neurodegeneration, which contribute to irreversible disability, occur early in the disease course and become predominant as the disease progresses, forming the underlying pathogenetic mechanisms [2, 3].

1.1.1 Demographic factors from a global perspective

MS is a complex disease with both genetic and environmental factors involved in disease susceptibility [4, 5]. The main environmental risk factors associated with MS in observational studies include obesity, vitamin D deficiency, Epstein-Barr virus infection, and smoking [4]. MS is not a directly inherited disease, however, it has an inherited component, meaning that genes play a role in the development of the disease. The largest genome-wide association study of MS discovered 233 genetic signals associated with MS, collectively explaining around 50% of MS heritability [6]. MS typically manifests between the ages of 20 and 50 years. However, a small percentage, approximately 0.5%, of adults with this disease experience symptom onset at age 60 or older, and up to 5% of patients with MS show first symptoms in childhood [1]. The median time to death from disease onset is around 30 years, representing a reduction in life expectancy of 5 to 10 years compared to unaffected people. Additionally, similar to many autoimmune disorders, MS is more prevalent in women (3:1).

In 2023, MS affects over 2.8 million people worldwide and its prevalence varies between countries, ranging from 1 up to >200 per 100,000 residents [7]. The disease global distribution generally increases with increasing distance from the equator, for its relation with vitamin D levels, although there are exceptions. Nowadays, Canada and United States are the ones with the highest prevalence of cases, followed by the Northern European countries. In Spain the most recent calculation of prevalence raises to 123 cases per 100,000 people (data consulted on July 2023).

1.1.2 MS clinical course

Diagnosis

A definitive diagnosis of MS cannot rely on a single clinical feature or test. Instead, a combination of clinical, imaging, and laboratory findings is necessary [8]. Despite this approach, misdiagnosis rates can still be as high as 10%, due to the confounding conditions that mimic MS, showing similar symptoms or other clinical outcomes [9].

Diagnosing MS requires objective evidence of CNS lesions disseminated in both time and space, meaning that multiple lesions can occur in different regions of the brain at different times. This dissemination in space refers to the presence

of MS lesions in different regions of the CNS, as outlined in the McDonald diagnostic criteria for MS [1], including periventricular, cortical or juxtacortical, infratentorial, and spinal cord locations (see Figure 1.1(b)). The evidence of dissemination in time involves the development of new MS lesions over time, either by the simultaneous presence of gadolinium-enhancing and non-enhancing lesions at any time or by the appearance of a new lesion on follow-up MRI compared to a baseline scan. Magnetic resonance imaging (MRI) plays a crucial role in this process, helping to exclude conditions that mimic MS. In 2001, the McDonald criteria introduced MRI evidence of CNS lesions disseminated in time and space, becoming the gold standard for MS diagnosis [10]. These criteria have undergone revisions over the years, offering increased certainty and enabling earlier and more reliable diagnosis, as outlined in the 2017 revision [1]. Additionally, the presence of oligoclonal bands (OBs) in the cerebrospinal fluid (CSF) together with absence of such OBs in serum may also be used as indicating of dissemination in space [1].

MS phenotypes

Impairment of nerve function in the brain or spinal cord occurs due to damage or loss of myelin, resulting in the disease's symptoms. MS symptoms can vary significantly among patients, depending on which areas of the CNS are primarily affected. MS can manifest mainly as a (i) relapsing-remitting disease, or as a (ii) progressive disease. Relapsing-remitting MS (RRMS) are typically initiated with a first acute demyelinating attack of the CNS, i.e., clinically isolated syndrome (CIS) and characterised by the presence of acute episodes of neurological dysfunction or relapses. Otherwise, the progressive forms are characterised by the gradual accumulation of disability which typically occurs independent of relapse activity. Traditionally, these forms include primary progressive MS (PPMS) and secondary progressive MS (SPMS), depending on whether progression is the initial manifestation of the disease (PPMS) or instead occurs after a relapsing-remitting phase (SPMS).

The initial presentation of the disease depends on the location of the lesions and the type of symptom onset (relapsing or progressive). Figure 1.1(a) shows some relevant anatomical regions and structures in the brain. Figure 1.1(b) illustrates a schematic diagram of lesion locations in brain and cord, as well as, other MRI prognostic features.

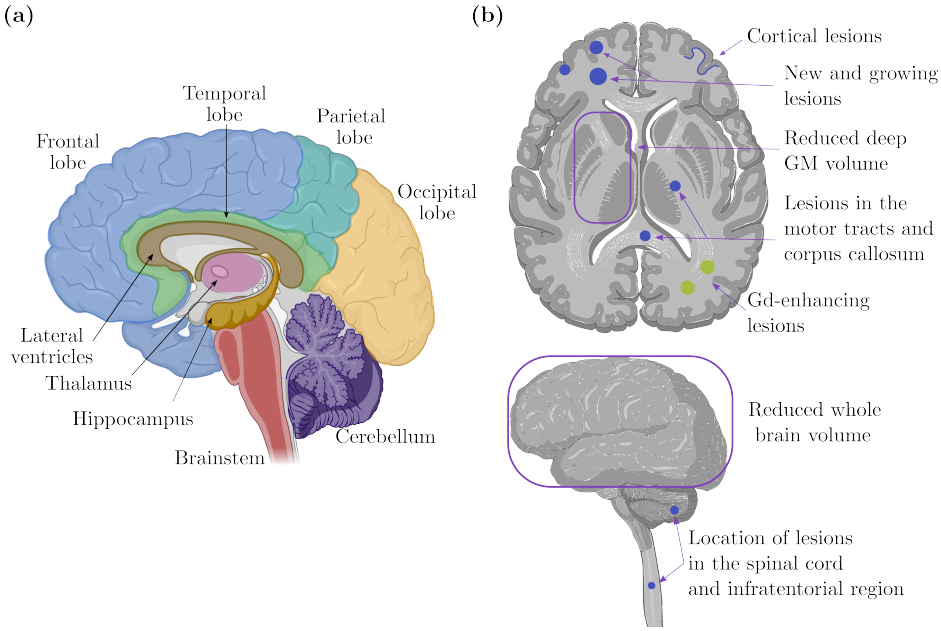

Figure 1.1: (a) Main brain anatomical areas and structures seen from a sagittal view. (b) Lesion locations and their evolution, together with other prognostic features that can be obtained using brain and spinal cord MRI, are a key tool for understanding and following the MS disease course. *Gd*: Gadolinium, *GM*: grey matter.

CIS is the first presentation of neurological symptoms attributed to an acute inflammatory and demyelinating event of the CNS, which typically involves the optic nerve, brainstem, or spinal cord. In general terms, a patient with a CIS fulfills the diagnostic criteria for MS when showing dissemination in space and time. This dissemination in space and time of the inflammatory-demyelinating disease can be clinical, when presenting with new symptoms suggestive of MS, and/or through MRI, when showing new lesions in the brain and cord MRI affecting different anatomical locations [1].

The most common form of MS is RRMS, affecting about 85% of people with MS. RRMS is characterised by relapses (at least two), acute attacks of new or recurrent neurological signs and symptoms, followed by either complete or partial recovery, with periods of stable neurological condition without clinical disease activity in between. It mainly affects young adults, with women being three times more affected than men.

In some cases, a patient may exhibit brain lesions on MRI scans resembling those observed in MS, yet they do not display evident clinical symptoms of the disease. Such cases are referred to as radiologically isolated syndrome. Around one-third of radiologically isolated syndrome patients may develop clinical symptoms of MS within 5 years of follow-up, either in the form of a relapse or progressive symptoms [11].

MS progression mechanisms

The most standardised way to monitor the increase in disability in MS in clinical practice is by scoring patients on the Expanded Disability Status Scale (EDSS) over time. To record disability worsening or progression, sustained increases in EDSS must be confirmed at 3 to 6 months or beyond [12]. This event of disability worsening confirmed at 3 or 6 months is also known as confirmed disability accumulation (CDA).

The main disability worsening mechanisms in MS are progression independent of relapse activity (PIRA), i.e., disability worsening in the absence of relapses, and relapse-associated worsening (RAW), i.e., disability worsening which occurs as a consequence of a relapse with incomplete recovery (illustrated in Figure 1.2). Typically, to define these events of PIRA and RAW, disability worsening needs to be confirmed 6 months after the first detected increase in disability [12, 13, 14]. It has been demonstrated by different authors that PIRA is the main cause of irreversible disability accumulation in MS [15, 16, 17].

1.1.3 Neuroimaging in MS

MRI is the preferred imaging modality for assessing brain and spinal cord damage caused by MS. It excels at detecting CNS demyelination, which is the deterioration of the myelin sheath. Over the last decades, MRI has become an essential tool in diagnosing MS, predicting its prognosis, and routinely assessing disease activity and treatment efficacy through successive MRI analysis over time. MRI-based biomarkers, such as the number of brain lesions and their evolution over time [1, 18], as well as quantifying brain parenchymal fraction or changes between two time-points [19, 20], have proven to be reliable indicators for determining patients' prognosis. Additionally, some of these biomarkers have been associated with the accumulation of patient disability [21, 22].

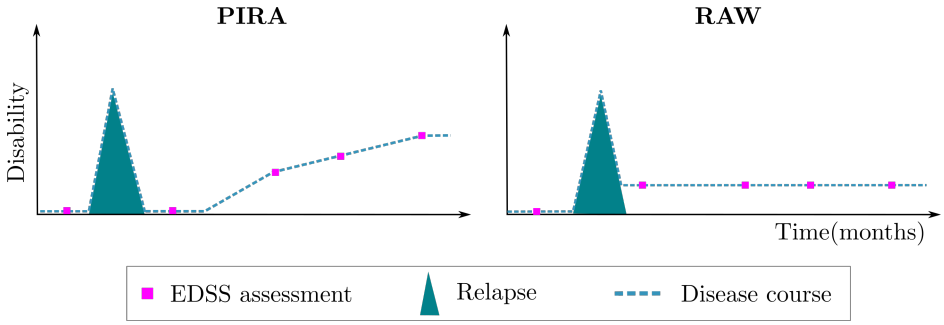

Figure 1.2: Mechanisms of disability worsening in MS. The main disability worsening mechanisms in MS are PIRA, i.e., disability worsening in the absence of relapses; and RAW, disability worsening which occurs as a consequence of a relapse with incomplete recovery. *PIRA*: progression independent of relapse activity, *RAW*: relapse associated worsening.

Even though MRI is a powerful tool to study MS disease, it has its limitations. Inherent to MRI scans, there is a high degree of heterogeneity, largely due to bias, noise in image intensity resulting from patient movement, or magnetic field inhomogeneity. Additionally, various factors, including hardware upgrades and standardisation of acquisition protocols, affect cohort and longitudinal studies. The high-dimensionality of MRI data makes it hard to manually analyse and make interpretations on them, which might also vary among radiologists. In recent years, with the continuous emergence of machine learning and deep learning techniques, significant advancements have been made in various medical imaging fields, including addressing some of the limitations associated with MRI data. Deep learning techniques are not dependent on predefined features; instead, they are capable of automatically extracting relevant information from raw or minimally processed data, improving the quality of MRI scans and their reliability of diagnosis. Moreover, these techniques have the ability to automate repetitive tasks, analyse large volumes of data such as MRI scans more efficiently, and achieve accuracies and reproducibility comparable to clinical experts [23]. In the context of MS, the use of MRI for deep learning-based models has been widely applied for segmentation and detection tasks or quantification of imaging features, which would otherwise be time-consuming [24, 25, 26]. More recently, they have begun to be used for the classification and future prediction of the disease, addressing more complex tasks that could provide valuable support to

clinicians in their day-to-day practice [22].

However, deep learning models are commonly perceived as "black boxes", lacking transparency and explanations for the decisions they make. A current trend in the field emphasises the development of explainable deep learning techniques. Their aim is to elucidate the rationale behind deep learning decisions, thereby providing additional information that can be immensely useful to end-users and enhancing data insights [27, 28].

1.2 Hypothesis

Machine learning and deep learning approaches in medical imaging have shown remarkable progress across a wide range of tasks such as diagnosis, often surpassing the expertise of clinical professionals in various clinical domains. However, the application of these approaches to classify patients based on their prognosis remains an underexplored area and is met with some reservations from physicians due to the perceived black box nature of these models. Concerns have been raised regarding the difficulty of tracing model decisions, attributed to the high number of parameters and highly non-linear interactions [29]. Despite these challenges, there have been attempts to use MRI scans in conjunction with machine learning [30] and deep learning algorithms [31, 32] for individual clinical prognosis of patients with MS, revealing promising capabilities in capturing complex non-linear relationships among data and demonstrating generalisation abilities. However, these approaches have not yet gained widespread acceptance in routine clinical practice.

Our *hypothesis* postulates that the geomorphometric features of CNS tissue damage are strongly associated with the individual disease course and disability accumulation in MS. Therefore, these features hold the potential to predict the clinical prognosis of patients with MS using deep learning techniques.

To test our hypothesis, we will use a large dataset of brain MRI scans derived from the clinical practice of people with MS diagnosed with a CIS. Building upon state-of-the-art deep learning approaches, we aim to develop models that are not only highly accurate but also comprehensible and trusted for experienced neurologists. By predicting short- and long-term disease progression, we seek to demonstrate the effectiveness of deep learning in prognosis prediction. Moreover, our aim extends beyond prediction accuracy; we will attempt to extract interpretable features that underline these prognostic predictions. These

interpretable features will offer valuable insights for clinicians, helping them to understand and embrace these novel methodologies for potential inclusion in routine clinical practice. By addressing the "black box" challenge and providing meaningful explanations for model decisions, we aim to bridge the gap between cutting-edge deep learning techniques and clinical acceptance.

In conclusion, our hypothesis centres on the potential of deep learning techniques to capture CNS tissue damage features from brain MRI scans, to significantly enhance clinical prognosis prediction for people with MS. Through a combination of accuracy, interpretability, and acceptance by medical experts, we hope that our work can contribute to a new era of data-driven precision medicine for MS.

1.3 Objectives

The main goals of this PhD Thesis are: (i) **to predict disease progression in MS, through deep learning approaches, using anatomical MRIs acquired in clinical practice** and, (ii) **to understand which spatiotemporal features of the CNS damage, extracted through a deep learning approach, imply a greater risk of disease progression**.

In order to achieve these goals the specific objectives to accomplish are:

- To curate a neuroimaging and clinical database adjusted for inclusion criteria, also including pre- and post-processing of anatomical brain MRI scans.
- To propose a deep learning image-based approach for cross-sectional stratification based on clinical data, only using as main input brain MRI scans.
- To analyse the suitability of various brain regions as input for a deep learning model in a cross-sectional stratification task.
- To propose an image-based deep learning model to predict long-term clinical outcomes, using anatomical brain images as input data.
- To extract class-specific attention maps for each outcome class, i.e., disabled and non-disabled patients in cross-sectional studies, as well as patients who reach a clinical outcome and those who do not in long-term studies. These

attention maps will provide insights into the most relevant features in the input images for predicting the disease course.

- To quantitatively compare the class-specific attention maps for each one of the clinical outcomes to help understand the pathological mechanisms underlying a worse prognosis.

1.4 Manuscript structure

The rest of this manuscript is organised as follows:

- **Chapter 2. Background.** After this introduction, we present a general background of the main methodological topics of this PhD Thesis: MRI and deep learning. First, we provide the fundamental details of MRI, followed by the main manifestations of MS in brain MRI scans. After that, we explain the basic concepts for building and understanding deep learning models and how they can be interpreted to unveil the underlying decision-making process of such models. Finally, combining the main three topics of this PhD Thesis, MS, MRI, and deep learning, we present a general state of the art of their principal applications.

- **Chapter 3. Database and image preparation.** In this Chapter, we present the main clinical and MRI data, and the databases used throughout this manuscript. Furthermore, we describe and present the main image pre-processing and automatic segmentation tools applied to the preliminary analysis of our input data, brain MRI scans.

- **Chapter 4. MS patients stratification [33].** In this Chapter, we present our first contribution, a deep learning pipeline for the stratification of patients with MS at any time-point solely using brain structural MRI data as input. This pipeline is supported by the use of an interpretability algorithm to decipher which are the main regions that contributed the most to the performed classification. Additionally, we compare our proposal with a traditional machine learning model and validate it on an additional unseen dataset.

- **Chapter 5. Regional approaches for MS patients stratification [34].** After analysing how deep learning models perform

on whole brain images, our second contribution is a comparison of different input regional models against the global approach. We define five different brain regions that have a strong relationship with MS prognostic factors and extract the different sampling patches using traditional image processing procedures. Additionally, we compare the performance of each single model with the ensemble of all of them and evaluate all the results with an external dataset.

- **Chapter 6. MS patients prognosis prediction.** In this Chapter, we present an image-based deep learning pipeline to estimate the survival function of patients with MS who will experience a first PIRA, using only brain MRI data at the time of the first demyelinating attack. Furthermore, we investigate which brain regions are most relevant to make these predictions, thus providing important insights into the pathological processes underlying PIRA.

- **Chapter 7. Conclusions.** Lastly, we provide a comprehensive discussion of the results obtained, as well as the main conclusions based on the contributions of this PhD Thesis. Based on these conclusions, we also point out different future investigations to improve and extend the work carried out for this PhD Thesis.

Chapter 2

Background

"Don't do it."
Now, a Doctor.

This Chapter aims to provide a theoretical background, including necessary concepts and definitions for the main techniques involved in this PhD Thesis: (i) the input data, MRI scans, (ii) the methods, deep learning algorithms, and (iii) the end-user output, the explainability behind deep learning models. Furthermore, it presents a contextualisation of the state of the art on deep learning models for MS using MRI data as input.

2.1 Magnetic Resonance Imaging

MRI is a non-invasive medical imaging technique used to create detailed cross-sectional images of the body's organs and tissues. In medical practice, MRI has become a standard tool for disease diagnosis and monitoring, including the evaluation of treatment response and the assessment of the development of brain damage over time [35, 36, 37, 38].

MRI scanners use strong magnetic fields (typically 1.5 or 3 T) and radio waves to generate three-dimensional (3D) images without ionising radiation. This imaging principle relies on the magnetic properties of hydrogen atoms within

water molecules in the body. When exposed to a magnetic field, hydrogen protons absorb and re-emit electromagnetic radiation. An excitation radio frequency pulse creates transverse magnetisation, evolving subsequently with the local Larmor frequency associated with the spatial position. Based on the energy that the atoms emit, a greyscale image is produced, corresponding to different energy levels that the MRI machine has read, correlated with different tissue types in the output image.

Protons in different types of tissue realign at different speeds, producing distinct signals measurable by receivers in the scanner machine, which are then transformed into an image. The contrast between different tissues is determined by the rate at which excited atoms return to the equilibrium state. The faster the protons realign, the brighter the image. Manipulating the contrast in an MR image is achieved by varying (i) the repetition time (TR), which is the time between the radio frequency pulses sent by the MR machine, and (ii) the echo time (TE), which is the time from the radio frequency pulse until the energy emitted is read (echo peak). Different pulse sequences can be set depending on the length of these times, resulting in different weighted images.

In brain MRI (see Figure 2.1 for an example), T1-weighted (T1-w) sequences, characterised by short TR and TE, produce images with dark CSF, light white matter (WM) and grey grey matter (GM). In contrast, T2-weighted (T2-w) sequences, with longer TR and TE, produce images with bright CSF, dark-grey WM, and light GM. Similar to T2-w, the Fluid Attenuated Inversion Recovery (FLAIR) sequence has a much longer TR and TE, with liquids being suppressed from the image.

An MRI scan produces a 3D volume, typically studied from the three orthogonal views: axial, sagittal and coronal. One important factor that determines the quality of the scan is the voxel spacing, which refers to the distance between pixels in a 2D plane and the thickness of the slices between planes. The voxel spacing affects how sharp the details appear in the final image. Smaller voxel spacing provides higher spatial resolution but it takes longer to acquire the data. Radiologists follow specific protocols to balance acquisition time and image quality. In brain MRI, common voxel spacings are $1 \times 1 \times 1mm^3$ for T1-w and $1 \times 1 \times 3mm^3$ for T2-FLAIR scans.

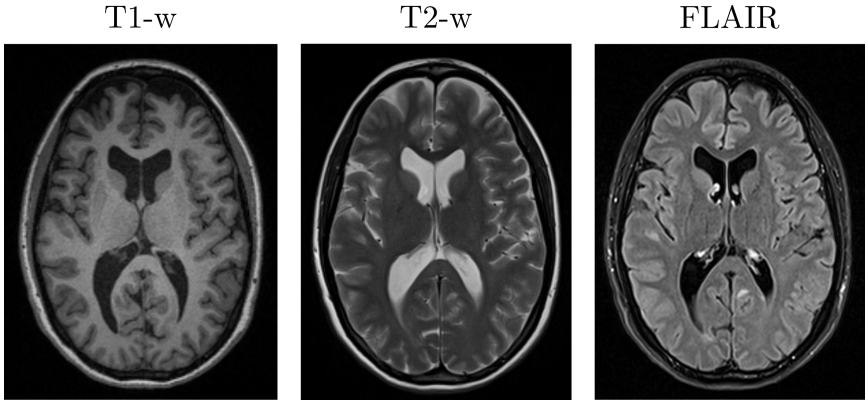

Figure 2.1: Axial view slice on different structural MRI image modalities: T1-weighted, T2-weighted and FLAIR sequences (from left to right).

2.2 Brain MRI in MS

Brain MRI principles have been explained in Section 2.1. In this Section, we will provide a more detailed description of how anatomical brain MRI scans of patients with MS are representative and discuss the main studied biomarkers extracted from them.

In clinical practice, the most frequently performed anatomical sequences are T1-w and T2-FLAIR scans. T1-w scans offer highly contrasted anatomical delineation of the different brain structures due to the high signal intensity from fat, which appears white (see Figure 2.1) and low signal from the CSF, which appears dark. This provides good tissue contrast. The main brain structures that can be easily identified on these type of images are the WM and GM, the lateral ventricles, the infratentorial area and the non-brain tissues, constituted mostly for the brain meninges.

2.2.1 MS lesions

The most representative sign of MS disease in brain MRI scans is the presence of MS lesions, which are areas of abnormal tissue as a consequence of demyelination. These lesions are typically small (<1 cm in diameter) and are located in characteristic regions of the brain (see Figure 2.2). Their appearance varies depending on the used imaging sequence used. In most cases, these brain

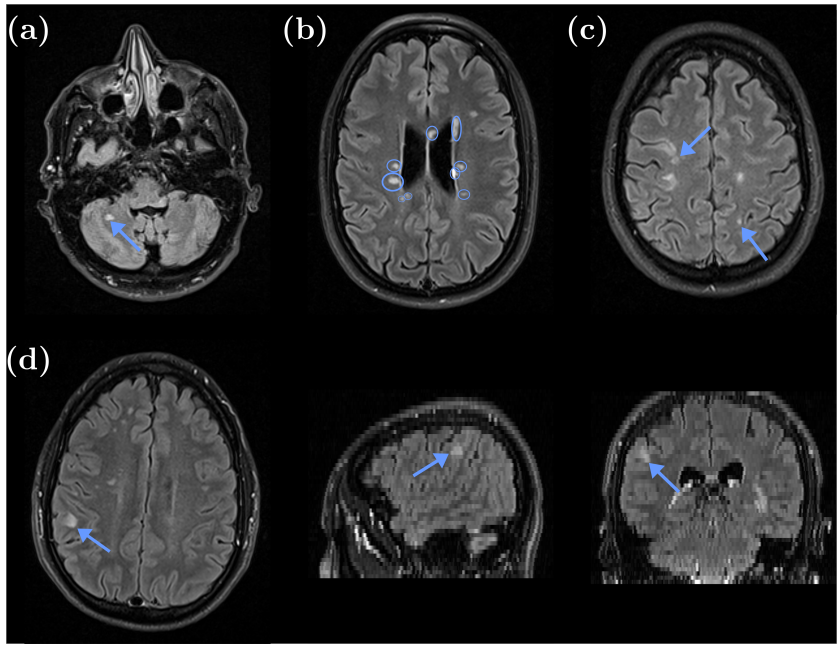

Figure 2.2: Brain WM lesions characteristic locations (T2-FLAIR sequence): (a) infratentorial, (b) periventricular and (c, d) juxtacortical lesions, (d) seen from the different views (from left to right) axial, sagittal and coronal.

lesions have T1-w values comparable to those of normal brain tissue, making them insufficiently contrasted to be identified accurately. For this reason, T2-FLAIR sequences are, normally, included in clinical routine imaging, as they are more sensitive to demyelination and inflamation signs. Oppositely to T1-w, in T2-FLAIR sequences, WM lesions are hyperintense (see Figure 2.1) and are typically located in the infratentorial, periventricular and juxtacortical areas. There also exist cortical lesions, which occur within the cortex and are more difficult to detect in structural MRI sequences.

2.2.2 Atrophy

Cerebral atrophy, also known as brain atrophy, refers to the loss of neurons and their connections, resulting in a reduction in the volume of both GM and WM in the brain (see Figure 2.3 for an example). While brain tissue loss is considered a

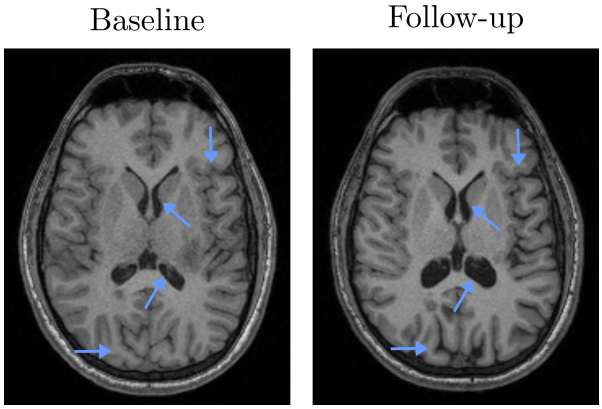

Figure 2.3: Brain atrophy in baseline time-point (left) and 5-years follow-up scan (right). Note the enlargement of the lateral ventricles and reduction in brain parenchymal volume.

natural part of the aging process [39], it is also a common neuroimaging feature of various disorders affecting the brain, including MS [40, 41]. In the context of MS, cerebral atrophy is characterised as diffuse and slow. Brain volume measurements play a pivotal role in MS studies, as they can be calculated for all patients, are associated with clinical risk factors, and can predict disease evolution [42, 43]. Therefore, accurate and reliable methods for quantifying brain atrophy may help the monitoring of disease progression and the assessment of the efficacy of new treatments.

However, the evaluation of brain atrophy quantification methods presents challenges, mainly due to the non-existence of sufficiently accurate manual ground truths for direct performance evaluation [44]. On the other hand, automated methods for brain atrophy quantification rely on indirect evaluation metrics, such as short-interval imaging errors or correlations with known clinical differences between populations. In general, longitudinal brain atrophy quantification methods can be classified into two groups: (i) segmentation-based methods, which measure volume differences between cross-sectional tissues or structures in each of the longitudinal scans [45]; and (ii) registration-based methods, which derive atrophy measures from the observed spatial deformation between two longitudinal scans [46]. Other studies have shown atrophy measures based on Jacobian integration [47]. As illustrated in Figure 2.4, these methods assess

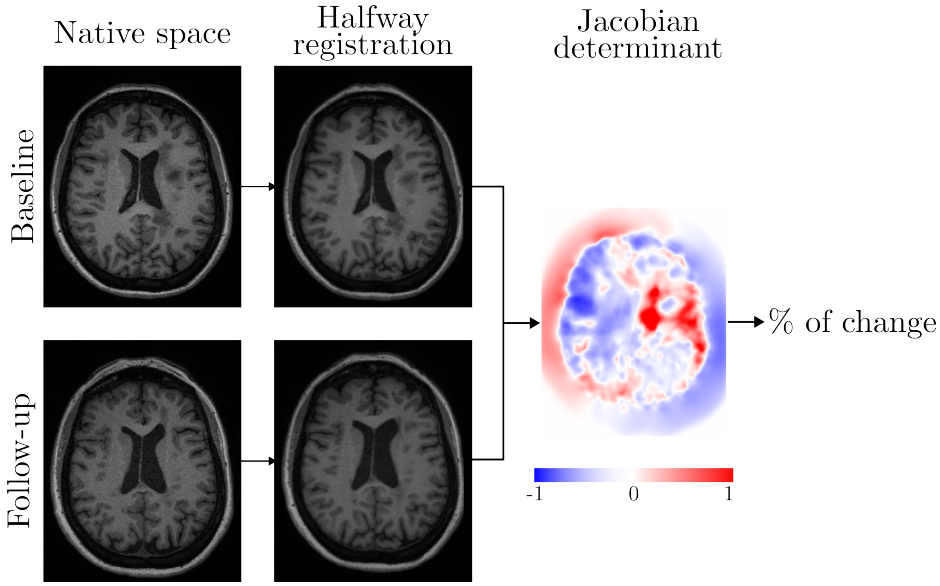

Figure 2.4: Example of brain atrophy quantification using the Jacobian determinant. A halfway registration is calculated from the native space of baseline and follow-up time-points scans to an intermediate space between both [48], which is used to calculate the Jacobian determinant which provides the brain volume change, i.e., the percentage of atrophy.

volume changes by integrating the determinant of the Jacobian of a non-linear transformation between two longitudinal scans. The region for integration is typically obtained from a cross-sectional segmentation; hence, cross-sectional tissue segmentation remains a necessary step.

Furthermore, to quantify cerebral atrophy in cross-sectional studies (i.e., without having longitudinal scans), the Jacobian determinant has also been employed to measure deformation from a scan in native space to a template based on a healthy cohort [49]. This extraction process will be elaborated upon later (see Section 3.3.3) since it will be used in the image processing pipeline of one of the conducted experiments.

2.3 Deep learning

Traditional machine learning refers to the development and study of algorithms and mathematical models that enable computer systems to make data-driven predictions. Traditional machine learning methods rely on hand-crafted or pre-extracted features to train models. For instance, in classification tasks, methods like support vector machine (SVM) or random forests have been commonly employed [50]. However, when working with images, additional procedures are required to extract these features for model training. Deep learning, a subset of machine learning and part of artificial intelligence, offers a more direct approach by simultaneously extracting relevant features and making predictions. This makes it particularly well-suited for handling high complexity data such as images [51]. Over the past few decades, deep learning has made a significant impact across various application fields, encompassing both industrial and day-to-day life applications. In the medical field, deep learning-based models have shown great promise, delivering accurate results in tasks such as diagnosis, prediction, and segmentation. In some cases, these models have even outperformed human experts.

2.3.1 Convolutional neural networks

Convolutional neural networks (CNNs) are a popular type of deep learning network for image analysis. They extract features from data using layers of learned filters, called **convolutional layers**. These filters, often referred to as kernels, are optimised during training to extract the most relevant features for a given task. These convolutional layers are arranged sequentially, where the output of one layer serves as input to the next one. This sequential arrangement enables the network to model complex non-linear relationships and extract increasingly sophisticated features from the input data. Each layer typically consists of multiple kernels, allowing the network to capture various features at different levels. The user-defined kernel size is usually much smaller than the input data, which reduces the number of connections and decreases computational costs. The kernels are effective at detecting features in different spatial locations within the image, but they do not account for scaling or rotation deformations [52]. Early layers in a CNN capture simple image features, often referred as low-level features, such as edges and blobs. Deeper layers represent more complex composite features. This architectural design makes CNNs particularly well-suited for the

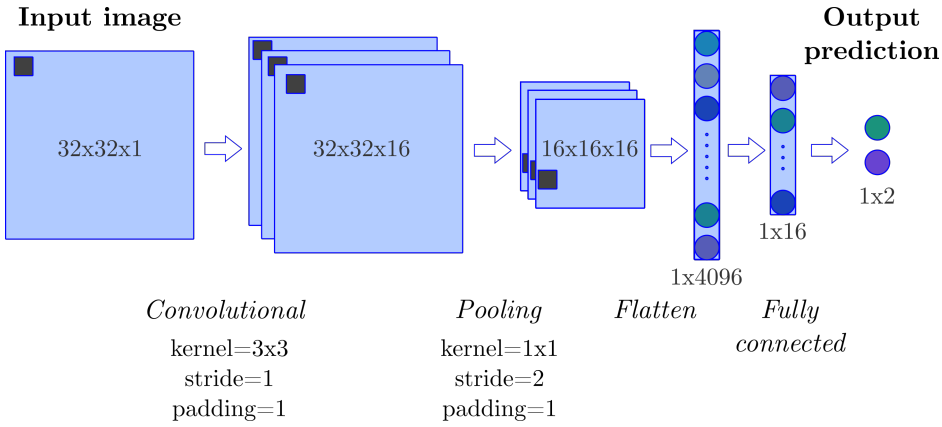

Figure 2.5: Example of a generic 2D CNN with one unique hidden layer. As input we have a 2D image with one channel and the 16 initial number of filters/kernels. Dimensions are expressed by width × height × channels after each layer operation is computed.

analysis of images and other high-dimensional data. A typical CNN architecture is represented in Figure 2.5.

Following a convolutional layer, a **pooling layer** is commonly applied to condense activation maps into a smaller representation, by performing operations such as max-pooling or average pooling. Pooling layers help reduce parameters and, sometimes, are replaced by convolutional layers with stride size of two, which also serves to reduce dimensionality [53].

Another common layer for feature extraction in CNNs is the **fully connected layer**, which applies linear transformations to the input vector through a weighted matrix. Fully connected layers are often employed after convolutional blocks in a CNN and can be responsible for the final classification or prediction. Additionally, they can be sequentially connected to each other and produce a dense network.

Activation functions are used to apply non-linear transformations to the outputs of convolutional and fully connected layers. Different types of activation functions exist, such as the *Sigmoid* function, which scales its input between 0 and 1, the *Softmax* function, which scales the input values (logits) into probabilities, or *Rectified Linear Unit (ReLU)* [54], which trims all negative input values to 0 ($ReLU(x) = max(0,x)$). Sigmoid and Softmax are commonly used as output

activation functions to normalise the network's output. On the other hand, ReLU activation, or its variants such as Leaky ReLU, are applied after every hidden layers or CNN block.

Flattening layers are used to reshape from the 2D/3D activation map at lower levels of the network into a one-dimensional representation that can be further processed by a fully connected layer. Flattening layers can also be replaced by a **global adaptive pooling layer** (GAP), which, instead of merely reshaping the input, computes the mean (or another operation, such as maximum) for each input feature map and conveys these values to the subsequent layers.

Finally, in classification tasks, a fully connected layer, or a relative convolutional layer with kernel size of 1, is employed as the **output layer**. The number of neurons in this layer correspond to the number of classes defined by the problem.

All these layers constitute the fundamental blocks of CNNs, and the combination of these blocks forms the basis for state-of-the-art architectures, with new developments continually being constructed upon this foundation [55, 56, 57, 58, 59].

Training process

In addition to the layers presented earlier, there are other crucial components that enable a neural network to operate effectively, specifically during training. The **loss function**, also referred as cost function, is responsible for quantifying the difference between the predicted output and the target output. It is continuously evaluated and optimised throughout the training process, typically with the aim of minimising its value. The choice of an appropriate loss function depends on the specific task at hand. Common loss functions include the *mean squared error*, often used for regression tasks, *categorical* or *binary cross entropy*, commonly used for classification tasks, or the *Dice Similarity Coefficient* loss, which is prevalent in segmentation tasks.

The **training process** of a neural network involves several key steps. Initially, input data is fed into the network and processed through each layer to generate an output. This phase is known as **forward propagation**. Subsequently, the network's output is compared to the desired outcome using the loss function to assess the network's performance. Following this, the error is **back-propagated**

through the network, which allows adjustments to the model's parameters based on their contribution to the error [60] (see Figure 2.6(a) for a graphical representation of these two processes). This process is repeated iteratively with the objective of optimising the network's weights to extract the most relevant features for the given task. **Optimisers** are algorithms used to update the network's weights and minimise the loss, while **regularisation** techniques can help prevent overfitting and enhance the model's generalisation [61].

2.4 Explainability

The adoption of deep learning-based models in clinical practice has been limited because of their black box nature. For these models to be truly valuable to clinicians, they must possess not only high accuracy but also transparency and trustworthiness. In response to this need, the field of explainability or interpretability methods, often referred to as explainable artificial intelligence, has emerged. The primary aim of explainable artificial intelligence is to render the decision-making process of deep learning models more interpretable [27, 28, 62, 63]. Although the field is still in its early stages, interpretability is defined by the capacity to elucidate how a model arrived at a specific prediction.

Explainability approaches may be classified based on several criteria, including (i) local vs global, (ii) model-specific vs model-agnostic, and (iii) post-hoc vs intrinsic explanations. At the end, an explainability approach can be defined by one of these categories or combinations of them.

- **Local** explanations focus on identifying the features within a specific image that influenced the model's output, while **global** explanations identify crucial features or characteristic attributes within a specific class as determined by the model.

- **Model-specific** explanations are based on finding the attributes characterising specific network architectures or pipelines. On the other hand, **model-agnostic** explanations concentrate on elucidating the relationship between the input and the output, independently of the network architecture used.

- Model-based or **intrinsic** explanations pertain to models, such as linear regression, that are conceptually simple and comprehensible while

maintaining a strong alignment between input and output. In contrast, deep learning models, due to their inherent complexity, demand **post-hoc** explanations. Post-hoc explanations delve into the learned features and their interactions to generate visual or textual explanations of the input-output relationship.

Beyond this categorisation, explainability approaches can be further subdivided based on the representation provided as an explanation. In the domain of medical image analysis, different explainability techniques adapted from the computer vision methods are prevalent [28]. These can be categorised into three main groups: visualisation, textual and example-based methods (see Figure 2.6(b)).

- Visualisation methods are the most extended and frequently used. These methods will be detailed later.

- Textual methods provide textual descriptions of the image's content, encompassing simple characteristics and, in some cases, complete medical reports. Techniques such as image captioning [64, 65], which employs language generation models like LSTM [66], fall into this category. See an example in Figure 2.6(b).

- Example-based explanation methods are closely aligned with human reasoning [67, 68]. These methods associate relevant examples from prior experiences with the presently analysed data. They produce an output that highlights the parts of the image on which the model based its decision. An example is illustrated in Figure 2.6(b).

2.4.1 Visual explanation

Visualisation methods aim to identify the regions of interest or attribution in input features that are primarily associated with a specific model decision. In the context of images, these explanations are typically represented as heatmaps or saliency maps superimposed on the input image. We can distinguish two main types of these methods (i) back-propagation-based and (ii) perturbation-based methods.

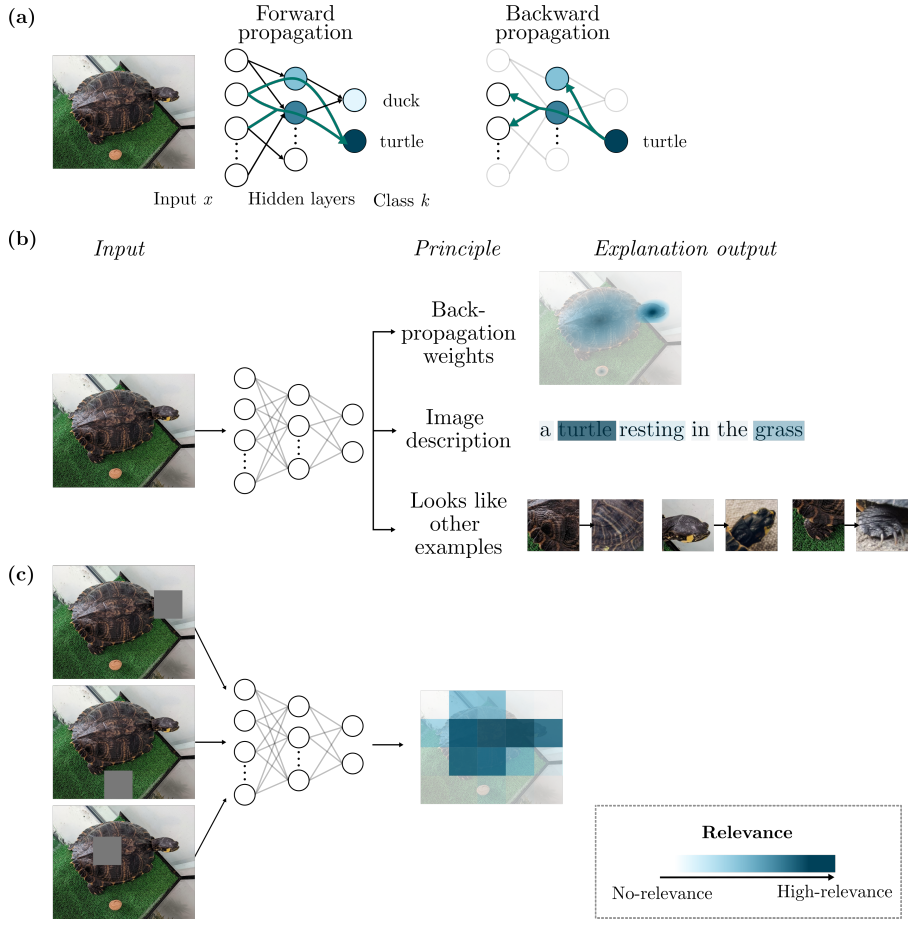

Figure 2.6: Explainability methods used in image classification tasks. (a) Gradient-based methods rely on a forward and backward propagation. Given an image x, a class k is maximally activated through forward passing throughout all layers of the network. All positive forward activations are recorded for later use during the backward pass. (b) Examples of different explainability methods used in image analysis. Taking an image as input the explanation output might be visual, textual or example-based depending on the principle that the method is based on. (c) An illustration of how perturbation methods, a subtype of visualisation methods, work.

Back-propagation-based

Back-propagation-based methods use the gradients of the trained model to identify the relevant parts of the input image. Some of the earliest techniques in this category visualised pixel relevance either examining partial derivatives, deconvolution, or guided back-propagation of the network output or intermediate network layers [53, 69, 70]. After that, more refined algorithms, such as Class Activation Mapping (CAM) [71] and Grad-CAM (Gradient-weighted Class Activation Mapping) [72], emerged as widely used methods [28]. These techniques generate attention maps that represent a weighted linear sum of different patterns captured by convolutional filters at various spatial locations. The layer-wise relevance propagation (LRP) method [73] produces attention maps that reflect pixel-specific relevance to the classification output. Additionally, the Deep SHapley Additive exPlanations (SHAP) method [74], grounded in game theory, calculates the contribution, known as Shapley values, of each feature (which in the case of images may refer to pixels or regions) to the network's output. More recently, trainable attention modules have been presented to be included into the network [75]. These modules amplify relevant areas and suppress irrelevant ones during training, based on the attention payed by the network during this process.

Perturbation-based

Perturbation-based methods compute marginal relevance. The relevance is attributed by comparing the output obtained from the original input and a modified version of it, removing or altering some of the input features [70, 76, 77, 78]. Thus, the application of these methods require multiple inferences to identify which parts of the image have a stronger effect on the output. It is important to note, that altering the image might not always have clear clinical significance. Therefore, there is a need for meaningful perturbations that are interpretable from a clinical standpoint [77, 79, 80]. Figure 2.6(c) illustrates how this method operates using a natural image example.

2.4.2 Method selection

In medical imaging analysis, most of these techniques have been explored mostly for classification and detection tasks [28]. The choice to use one or another technique depends on several aspects [62]. On one hand, simplicity

is preferred, with post-hoc agnostic methods that are not dependent on the model implementation itself and are available as open-source implementations. From a computational standpoint, cost is a critical factor to consider for future implementations, including resource allocation. It is important to consider that a single pass back through in back-propagation-based methods is not equivalent to multiple passes in perturbation-based methods. Lastly, validating the acquired explanations is as essential as assessing the model's outcomes. There is a need for checking the correctness of the explanations obtained or at least checking that they are following a reasonable path (when dealing with the non-existence of ground truth, e.g., discovery of new biomarkers). For that, clinicians, as end-users and experts on the studied problem, are required to proof the findings and their explanations.

2.5 Deep learning models for MS

In this Section a summarised state of the art of deep learning models for MS using structural MRI as main input is presented [22, 25, 81]. We have categorised them into four distinct groups based on their applications. Firstly, we find deep learning frameworks designed to (i) enhance image processing pipelines when dealing with MRI scans from patients with MS. In addition, the categories that could be directly implemented in clinical practice comprise tools that aid clinicians in (ii) detecting MS signs, particularly by segmenting MS lesions, (iii) diagnosing the disease, which includes phenotype classification, clinical profiles, and the exclusion of MS-mimics diseases, and (iv) predicting the disease's progression.

As follows, we will explain and present these four categories and provide a more detailed description of those studies that are more directly related to the topics covered in this PhD Thesis, that is, classification of clinical profiles and prognosis studies.

2.5.1 Image processing enhancement

The different brain volumes that can be quantified from MRI scans offer promising biomarkers for monitoring and prognosis of patients with MS. To obtain these volumes, segmentation of brain tissue and anatomical structures is necessary. Segmentation involves the process of extracting a region of interest (ROI) mask. However, most standard segmentation tools do not account for the presence of

WM lesions, which are characteristic in scans of patients with MS. To address this limitation, two distinct types of deep learning approaches have been proposed: (i) segmentation of brain tissue and structures considering the presence of WM lesions, and (ii) lesion filling or inpainting.

For **brain structures segmentation**, studies are based on the joint segmentation of brain structures and WM lesions at the same time, as in McKinley *et al.* [82], where the authors trained state-of-the-art CNNs with annotated lesions with and without anatomical labels. Billot *et al.* [83] proposed an unsupervised segmentation pipeline which included the generation of synthetic scans using Gaussian mixture models conditioned on label maps. In a similar manner, some **brain tissue segmentation** studies combine the segmentation of tissue and lesions on the same pipeline [84]. Clèrigues *et al.* [85] presented a lesion inpainting and posterior tissue segmentation pipeline, while reducing the effect of lesions when performing tissue volume quantification.

Other studies just centred on the **lesion inpainting** task. In some studies the lesions are filled with morphological and textual information of normal-appearing tissues [86], or taking into account the whole-image intensities [87]. Manjón *et al.* [88] proposed a blind inpainting which reconstructs a "corrupted image" into a "clean" one, with a two-step pipeline which first generates a synthetic training dataset (non-blinded) for the blind inpainting network.

Synthesis

One of the emerging proposals to overcome the lack of available data to train deep learning models has been the generation of synthetic MR images. Generative adversarial networks (GANs) were introduced in [89] achieving impressive results of realistic-looking images from an implicit distribution that follows the real data distribution [90]. In MS, there are two primary objectives for synthesising MR images. Both aim to enhance other tasks related to MS, like detection or diagnosis, particularly in the context of longitudinal studies.

New T2-lesions. For instance, Salem *et al.* [91] proposed the generation of new T2-lesions on healthy controls fused with lesion masks from patients with MS by using a U-Net [92]. With the same purpose, Basaran *et al.* [93] used a GAN to further used the generated synthetic images of healthy controls with lesions for being used in the training set for a segmentation pipeline, showing better performance than when training with real ones. From another point of

view, Kamraoui *et al.* [94] employed a U-Net model, using an MS patient's scan as source to generate a longitudinal patient-specific scan. This scan included new lesions either extracted from the given scan time-point or with the lesions suppressed.

MR image modalities. A target MRI sequence modality is estimated from a source sequence or from multiple sequences. Wei *et al.* [95] proposed a 3D-CNN to map multi-sequence source sequences into their corresponding (missing) FLAIR sequence. In a more recent study, Valencia *et al.* [96] proposed the use of partial volumes from a source modality (T2-FLAIR) to generate a different target modality (T1-w). This study, along with the one by Basaran *et al.* [93], reported improved segmentation performance for new T2 lesions by incorporating synthetic training images.

2.5.2 Detection

Most of the studies of MS involving deep learning and MRI data have as main goal the detection of biomarkers that will help with the diagnosis and the prediction of the prognosis of the disease. These main biomarkers are the T2-lesions [81].

T2-lesions

Manual detection of MS lesions is time consuming and prone to errors of inter-rater variability [97]. The search of a domain invariant, accurate and robust algorithm to perform this detection and precise segmentation of MS lesions rapidly from MRI is highly needed [26]. In the last years, a wide range of automatic segmentation algorithms based on deep learning methods have been proposed [98]. We can find CNN implementations with all different dimensions, 2D [99, 100], 2.5D, i.e., combining the different 2D views [101, 102] and, 3D [103, 104, 105]. Encoder-decoder, feature extraction and up-sampling reconstruction paths, and their variations such as the well-known U-Net [92] and the recent nnU-net [106] proposed by Isensee *et al.* are the most commonly used CNN architectures used for recent MS lesion segmentation. Additionally, these have been the best scoring algorithms in MS segmentation challenges [97, 107, 108], which provided datasets further exploited in a wide range of MS studies, allowing comparisons between proposed methods [26].

Nonetheless, despite the endeavours to address the challenge of domain adaptation for scans obtained from various vendor machines or different centres

lacking a standardised protocol [109], these efforts may limit the suitability of these implementations for clinical use, as demonstrated by previous studies [110, 111, 112].

2.5.3 Diagnosis

Using deep learning models in combination with MRI scans as input data to diagnose or identify different MS clinical profiles has not emerged as a prevailing focus within the MS research. However, with the introduction of interpretability models and their ability to provide additional insights into the anatomical regions in the brain potentially linked to the disease, the focus of interest has significantly increased. Table 2.1 summarises the main studies that have been published regarding classification tasks in MS including some of the disease subgroups.

MS patients vs healthy controls

One of the first fundamental tasks in the analysis of normal-appearing MRI scans of MS patients is the differentiation between MS patients and healthy individuals. Yoo *et al.* [118] presented the first deep learning model designed to extract information from normal-appearing patches from patients with MS and distinguish them from those of healthy controls, despite having a relatively small sample size. Using a much larger dataset, Siar *et al.* [119] proposed the use of a multi-class 2D-CNN for a classification task including patients with MS, with tumors and healthy controls. More interestingly, Eitel *et al.* [115] took a step beyond disease diagnosis. In addition to distinguishing between patients with MS and healthy controls, they introduced the first saliency maps revealing the specific brain areas on which the neural network focused to make its final predictions. Their approach involved the pre-training of a 3D CNN on a more extensive MRI dataset, specifically from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Subsequently, they fine-tuned the model using data from MS patients and healthy controls. During inference, they used the LRP [73] to uncover the decision-making process.

MS-mimics

Fundamentally, to diagnose a patient with MS, other pathologies must be excluded first, especially those that present similarities with MS. [1, 120]. For

Table 2.1: Subgroups classification in MS state-of-the-art.

Method	**MS subgroups**	**Strategy**	**Data**	**N** (train/test)	**XAI**
Karaca *et al.*, 2017 [113]	RRMS, PPMS, SPMS	1D-DNN	lesions diameter	120 / 50	-
Marzullo *et al.*, 2019 [114]	CIS, RRMS, SPMS, PPMS, HC	Graph-CNN	DTI	602 (cross-val)	-
Eitel *et al.*, 2019 [115]	MS, HC	2D-CNN	MRI (T1-w, FLAIR)	921(AD) / 147	LRP
Eshaghi *et al.*, 2021 [116]	CIS, RRMS, PPMS, SPMS	unsupervised machine learning	pre-extracted MRI features	6322 / 3068	Proposal model-based
Zhang *et al.*, 2021 [98]	RRMS, SPMS, HC	2D-ResNet & 2D-VGG	MRI (T1-w, T2-w, FLAIR)	25 / 8	Grad-CAM
Cruciacni *et al.*, 2021 [117]	RRMS, PPMS	3D-VGG	T1-w MRI, DTI, 3D-SHORE, RIFs	91 (cross-val)	LRP

XAI: explainable artificial intelligence, *CIS*: clinically isolated syndrome, *RRMS*: relapsing remitting multiple sclerosis, *PPMS*: primary progressive multiple sclerosis, *SPMS*: secondary progressive multiple sclerosis, *HC*: healthy control, *FU*: follow-up, *DTI*: diffusion tensor imaging, *SHORE*: simple harmonics oscillator based reconstruction and estimation, *RIFs*: rotation invariant features, *DNN*: dense neural network, *LRP*: Layer wise relevance propagation, *AD*: Alzheimer's disease.

example, there have been studies focused on distinguishing patients with MS from those with neuromyelitis optica spectrum disorder (NMOSD). Kim *et al.* [121] used a 3D CNN only using FLAIR scans and clinical data as input. Similarly, Wang *et al.* [122] proposed a two-view 2D CNN that took lesion masks as input for this classification task. Rocca *et al.* [123] proposed a deep learning-based model to discriminate MS from NMOSD, as well as two other pathologies that mimic MS, i.e., CNS vasculitis and migraine. Their model demonstrated levels of accuracy that surpassed those achieved by two expert neuroradiologists. Moreover, Maggi *et al.* [124] proposed a pipeline to automatically assess the central vein sign in WM lesions, allowing for the differentiation of MS from its mimics (grouping different pathologies).

Identification of clinical profiles

As has been introduced in Chapter 1.1.2, patients with MS can manifest different clinical profiles or phenotypes throughout the course of the disease. Being able to determine a patient's clinical profile at a specific cross-sectional time-point holds potential as a valuable tool for screening and large-scale therapeutic interventions. However, studies of this nature that utilise MRI data as the primary input are rarely found in the literature.

Zhang *et al.* [125] presented a bi-comparison of classification architecture models and interpretability algorithms. They only used 19 patients with MS (10 RRMS and 9 SPMS), and a group of healthy controls. The authors compared the performance of two well-known 2D classification architectures, ResNet [56] and VGG [55]. They investigated variants of these architectures with fully connected or GAP layers at the end of the feature extractor and evaluated their performance using either pre-trained ImageNet weights or training from scratch. As interpretability algorithms, they compared three CAMs: CAM, Grad-CAM and Grad-CAM++. Input data consisted of 2D-axial slices of T1-w, T2-w and FLAIR sequences. The optimal performance was achieved using a pre-trained VGG model with a GAP layer (accuracy of 95.42%), while the Grad-CAM revealed to be the best localising patterns in the output heatmap. This best performance revealed that the frontal, temporal and parietal brain areas held the greatest significance in differentiating RRMS from SPMS.

Aiming to compare the same types of populations, RRMS and PPMS, Cruciani *et al.* [117] used a larger dataset consisting of 91 patients with MS (RRMS=46 and PPMS=45). They proposed the use of a 3D VGG [55] that was independently trained on data of 4 different image modalities, including T1-w MRI, diffusion tensor imaging, 3D simple harmonics oscillator based reconstruction and estimation, and rotation invariant features. The T1-CNN model employed the GM tissue probability map as input, rather than the T1-w image itself. In terms of interpretability, they used the LRP [73]. Results unveiled that the model trained with the MRI-extracted representation outperformed the other modalities (accuracy of 0.84 ± 0.1 with a 5-fold cross validation strategy). Notably, all models reported the same LRP behavior for both RRMS and PPMS classes, with the T1-CNN model yielding the lowest relevance values and selectively highlighting the temporal pole.

As can be seen, these studies represent a shift beyond pure classification and

introduce model interpretability elements into the evaluation of clinical profiles.

2.5.4 Prognosis

In the literature, deep learning models focused on the prognosis evaluation of MS remain limited. Prognostic tasks require large cohorts with multiple follow-up time-points, which are rarely found in the study of neurodegenerative diseases departing from clinical data. In the context of MS, existing studies are often constrained by relatively small databases, usually encompassing fewer than 500 patients (including training and testing sets), with exceptions. These studies primarily target the prediction of short-term disease progression, typically spanning a 1-2 year follow-up period, and predominantly involve patients with relapsing-remitting MS (RRMS) or primary progressive MS (PPMS) within a certain range of disease duration. Table 2.2 summarises the main studies that have been published in the recent years regarding prognostic tasks in MS.

Tousignant *et al.* [128] presented the first end-to-end deep learning approach to predict disease progression at one-year follow-up in 465 RRMS patients with MRI scans of a multi-scanner and multi-centre cohort. They implemented a 3D-CNN model with convolutional blocks composed of 4-parallel pathways at different image resolution sizes. As input, they used 5 different MRI modalities with and without lesion masks, and also compared the model's performance with that of a VGG architecture [55]. The best results were obtained when using all sequences in conjunction with pre-extracted lesion masks (area under the ROC curve (AUC) of 0.701 ± 0.024 with a 4-fold cross validation strategy).

In 2019, a challenge focusing on MS was organised as part of the *Journées français de radiologie* (the annual meeting of the French Society of Radiology) with the mission to predict the EDSS score of patients with MS at two-years follow-up [132]. The challenge organisers provided a total of 971 MRI scans (multi-centre and multi-scanner) for training and validation, along with an additional test set of 475 MRI scans. The winners of the challenge, Roca *et al.* [31], proposed a combination of complementary predictors. First, as deep learning predictor, they used a 3D-CNN using FLAIR sequences and their corresponding lesion mask as input, and the addition of the demographic variable, age, at the last fully connected layer. The team also employed several classical machine learning algorithms, including random forest regression and manifold learning, which were trained on different volumetric and lesion location features

Table 2.2: Prognostic tasks in MS state-of-the-art.

Method	**Prognostic task**	**Strategy**	**Data**	**N** (train/test)	**XAI**
Wottschel *et al.*, 2015 [30]	1- and 3-years FU from CIS	SVM	pre-extracted MRI features, clinical	74 and 70 (cross-val)	-
Bendfeldt *et al.*, 2019 [126]	2-years FU from CIS	SVM	pre-extracted MRI features, clinical	364 (cross-val)	-
Wottschel *et al.*, 2019 [127]	1-year FU from CIS	SVM	pre-extracted MRI features	400 (cross-val)	-
Tousignant *et al.*, 2019 [128]	1-year FU of RRMS	3D-CNN	MRI (T1p, T1c, T2-w, FLAIR, PDw), lesion mask	465 (cross-val)	-
Yoo *et al.*, 2019 [129]	CIS to MS conversion	3D-CNN	lesion masks and user-defined measurements	140 (cross-val)	-
Roca *et al.*, 2020 [31]	EDSS score at 2-years FU	3D-CNN, machine learning algorithms	MRI (FLAIR), lesion mask, age	971 / 475	-
Storelli *et al.*, 2022 [32]	2-years FU of RRMS/PMS	3D-CNN	MRI (T1-w, T2-w)	325 / 48	LRP
Durso-Finley *et al.*, 2022 [130]	treatment recommendation	ResNet	MRI (T1p, T1c, T2-w, FLAIR, PD-w), lesion masks, clinical	1872 (cross-val)	-
Falet *et al.*, 2022 [131]	treatment effect estimation	MLPs	pre-extracted MRI features, clinical	3830 (cross-val)	-

XAI: explainable artificial intelligence, *CIS*: clinically isolated syndrome, *RRMS*: relapsing remitting multiple sclerosis, *PMS*: progressive multiple sclerosis, *HC*: healthy control, *FU*: follow-up, *SVM*: super vector machine, *MLPs*: multilayer perceptrons, *T1p*: T1-w pre-contrast, *T1c*: T1-w post-contrast, *LRP*: Layer wise relevance propagation.

within WM tracts. They compared the different individual predictors and obtained the best performance with an aggregated model using all of them (mean square error of 3 in the test set).

More recently, similarly to [128], Storelli *et al.* [32] proposed a 3D-CNN for predicting disease progression at 2-years follow-up. They used 325 (262 RRMS and 63 PPMS) patients for training and 48 (26 RRMS and 22 PPMS) for testing. The outcome-target disability was based on changes on clinical evaluation (based on EDSS) and cognitive evolution (based on symbol digit modalities test [SDMT]) scores between baseline and follow-up. They compared the performance of a model built with one (EDSS) or the other (SDMT) score, or both (EDSS & SDMT). The best performance was obtained using both EDSS and SDMT information (accuracy of 87.5%) which exceeded the one performed by 2 expert physicians. They also offered examples of visual explanation of the decision of the network.

Additionally, without taking MRI scans as input, Yoo *et al.* [129], proposed a 3D-CNN for predicting conversion to MS from CIS. The dataset consisted of 140 early MS patients (or CIS), which 80 of them converted to MS within two years. They used the Euclidean distance transform obtained from pre-extracted lesion masks as main input to the network and additional user-defined measurements. They reported an accuracy of 75% (SD=11.3%) with a 7-fold cross validation strategy.

Treatment monitoring

Recently, the first deep learning-based models introducing the task of evaluating treatment response have been presented.

Durso-Finley *et al.* [130] presented the first image-based deep learning model for estimating treatment recommendations for MS. This model goes beyond prognosis prediction and incorporates an evaluation of treatment-associated risks. They used baseline multimodal MRI scans and clinical data from RRMS patients to learn shared latent features using a well-known ResNet [56] encoder as feature extractor, followed by treatment-specific multilayer perceptrons for outcome prediction. One notable aspect of their work was the exploration of treatment efficacy estimation, particularly concerning indicators of new disease activity, such as the appearance of new or enlarging T2-lesions, as measured by their suppression. After that, Falet *et al.* [131] introduced a deep learning framework

centred on pre-extracted MRI metrics, primarily related to lesions and other brain volumetric measures. Together with clinical data, their approach aimed to estimate the effects of treatment on disability progression. This model relied on an ensemble of different treatment-specific multilayer perceptrons, which were pre-trained using data from RRMS patients and fine-tuned for PPMS patients. Additionally, they conducted survival analyses, providing risk stratification based on treatment effects for patients participating in clinical trials.

Chapter 3

Database and image preparation

"The best thing about image processing is that you can apply it to any field you're into."

A Doctor.

This Chapter aims to provide the reader with the required concepts and details about all the data used in this PhD Thesis. First, we explain the main clinical variables commonly collected in MS clinical routine, followed by the explanation on the main type of data used for this PhD Thesis, i.e., the brain MRI scans. Then, we present the main dataset from our centre, Vall d'Hebron University Hospital (VHUH), along with the Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATHS) [133] database which was used as a validation cohort for the first part of this PhD Thesis (Chapters 4 and 5). Finally, the different image processing steps conducted on all the scans are introduced at the end of this Chapter.

3.1 Clinical data

Clinical data was collected at baseline and during the disease course of a patient. A wide variety of data can be included, from patient's symptoms, medical history, or diagnostic test results to treatment outcomes. This data and its follow-up are important for understanding the disease course and offering a more personalised treatment plan. In this Section, we will detail the main clinical variables that have been widely used in MS studies.

3.1.1 Demographics

Demographic data refers to information about patients' personal characteristics. This data is collected at baseline and during the disease course of a patient's medical treatment, and can be used to understand patterns and trends in the disease within a specific population. In MS, the most studied demographic variables are [134]:

- **Sex** refers to the biological and physiological characteristics that define women and men.
- **Age at CIS** refers to the age at which a person experiences their first episode of neurological symptoms. **Age at MS diagnosis** is the age at which a person fulfills the diagnostic criteria.
- **Race and ethnicity** refers to a person's self-identification with one or more social groups.
- **Socioeconomic status** includes environmental factors like educational level or occupation among others.

3.1.2 Disease related data

The main clinical data related to MS that we have used to describe our cohorts are the following ones.

CIS topography

CIS topography refers to the location in the CNS of the acute damage supposedly associated with the first attack, i.e., the first episode of neurological symptoms.

The most common CIS topographies are: brainstem, optic nerve, spinal cord, hemispheres, or polyregional, i.e., when more than one CNS region is involved.

Presence of oligoclonal bands

OBs are a type of protein, i.e., immunoglobulins (specifically immunoglobulin G), which may be present in the CSF in people with MS. In this condition, the OBs are typically detected in the CSF and not in the blood, indicating inflammation in the CNS. The presence of OBs is tested in the laboratory and plays an important diagnostic and prognostic role. That is, the presence of OB has been demonstrated to be a characteristic feature of MS and is associated with increased risk of disability accumulation as well as an increased risk of CNS tissue damage [18, 135].

Disease duration

We refer to disease duration as the elapsed time from the first demyelinating attack.

Radiology report

In clinical practice, clinicians have access to radiology reports issued by experienced neuroradiologists. From these reports, the more interesting entries are the number of lesions in the brain and, if available, the number of lesions in the spinal cord. In our studies, these are presented as categorical variables, with multiple ranges: 0, 1-3, 4-8 or >9 for the brain, and 0, 1, 2-3 or >3 for the spinal cord.

Treatment

MS is a chronic neurodegenerative disease with no cure. However, there are treatments that can help control the inflammatory component and, possibly through their anti-inflammatory mechanism, delay the accumulation of irreversible disability. However, accounting for treatment exposure in observational studies, such as those presented in this PhD Thesis, is always complex due to their non-randomised nature. This is even more challenging when building deep learning models solely based on imaging data as input data to the models.

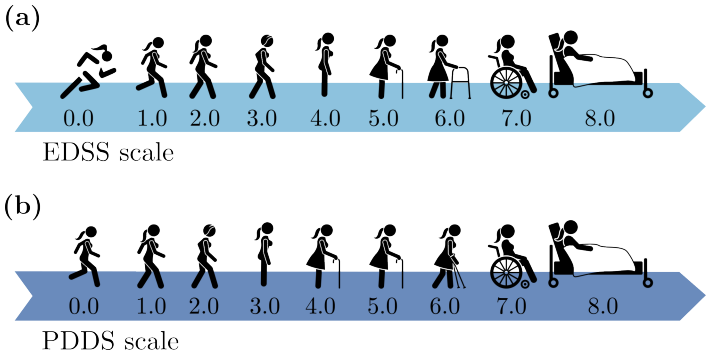

Figure 3.1: The main clinical outcome measure scales in MS are (a) the Expanded Disability Status Scale (EDSS) and (b) the Patient Determined Disease Scale (PDDS).

3.1.3 Clinical outcome measures

Accurate clinical evaluation of patients with MS is key for detecting MS progression and assessing response to treatment in clinical practice. In this Section, we present two of the most widely used disability scales for characterising the clinical disease status: the EDSS and the patient determined disability status (PDDS). Figure 3.1 shows a graphical representation of these scales.

EDSS

The EDSS is the most widely used clinical scale to assess the MS-related disability [136]. EDSS is an ordinal scale that ranges from 0 (normal neurological exam) to 10 (death due to MS), with 0.5-point increments interval after an EDSS of 1.0. The definitions of each EDSS score are based on walking capacity but also on other functional system affectations (see Figure 3.1(a)). The EDSS score is assigned by a neurologist as a result of a complete neurological exam.

In longitudinal studies, we can characterise disease progression by an increase in the EDSS score of 1.5-point from an EDSS of 0.0, 1.0-point for an EDSS score between 1.0 and 5.0, or 0.5-point from an EDSS$\geq$6.0. Other significant thresholds or events in this scale include: (i) reaching an EDSS$\geq$3.0, commonly recognised as the boundary between mild or no-disability and moderate disability status, and (ii) an EDSS$\geq$6.0, revealing signs of severe disability [18, 137, 138]. Any EDSS increase should be confirmed at least 3 or 6 months later; when this occurs, we refer to it as CDA.

PDDS

The PDDS is a patient-reported outcome based on a self-administered questionnaire, which, in a standardised way, attempts to measure the level of disability without the direct intervention of the clinician [139]. The PDDS is an ordinal scale ranging from 0 (normal) to 8 (bedridden) with 1.0-point increments (see Figure 3.1(b)).

The PDDS score has been proven to have a strong correlation with the EDSS score [140, 141]. However, despite both scales being non-linear, the EDSS is obtained by a neurologist, after performing an anamnesis and a neurological examination, while the PDDS is reported by the patient, making it inherently more subjective. This difference likely accounts for the lack of a "perfect" correlation (ρ=1).

3.2 Datasets

3.2.1 VHUH cohort

Our main cohort of study to carry out all the experiments of this PhD Thesis is the Barcelona CIS cohort [18]. This is an ongoing prospective cohort of patients followed up over time after their first demyelinating attack (or CIS) at the Multiple Sclerosis Centre of Catalonia (Cemcat), VHUH. This cohort started in 1995 and is still in course.

The inclusion criteria for this cohort encompass patients younger than 50 years who experienced a first demyelinating attack of the CNS, which could not be attributed to other diseases. These patients were assessed at Cemcat within 3 months of the first demyelinating attack and provided written informed consent [18].

The subsets of this cohort used for our experiments ranged between 2009 and 2020, depending on the specific goals of each study, which will be detailed in their respective Chapters.

All patients included in our cohort underwent clinically evaluated by a neurologist in our centre, and they had brain MRI scans within the first 5-6 months after symptom onset. These scans were routinely repeated every year to monitor the disease course. At baseline, demographic data and previous clinical history were collected, along with other clinical data such as CIS topography and

Table 3.1: MRI sequence acquisition parameters for each scanner used in the VHUH cohort.

	Tim Trio	Symphony Tim	Symphony	Avanto Fit	Avanto
Field Strength, T	3	1.5	1.5	1.5	1.5
MPRAGE					
TR, ms	2300	1980	2700	2300	1980
TE, ms	2.98	3.08	4.8	3.05	3.1
TI, ms	900	1100	850	900	1100
Voxel size, mm	$1 \times 1 \times 1.2$	$1 \times 1 \times 1$	$1 \times 1 \times 1.2$	$1 \times 1 \times 1$	$1 \times 1 \times 1$
Plane	Sagittal	Sagittal	Sagittal	Sagittal	Sagittal
FLAIR					
TR, ms	9000	8500	9000	8500	8500
TE, ms	87	95	114	99	92
TI, ms	2500	2440	2500	2440	2439
Voxel size, mm	$0.5 \times 0.5 \times 3$ $1 \times 1 \times 1^a$	$1 \times 1 \times 3$	$0.5 \times 0.5 \times 3$	$1 \times 1 \times 3$	$1 \times 1 \times 3$
Plane	Axial	Axial	Axial	Axial	Axial

[a] With the Tim Trio scanner FLAIR sequences are acquired either with one or other voxel size (2D and 3D).
TR: repetition time, *TE*: echo time, *TI*: inversion time.

the first EDSS score. Subsequent visits recorded EDSS scores and documented any relapses or other disease-related symptoms [18].

MRI acquisition parameters

Brain MRI data were acquired as part of clinical practice in our centre. There are five different scanner machines from the same vendor, Siemens, and patients may have been scanned by any of them at any time of their disease course. A standardised acquisition protocol, including MPRAGE and T2-FLAIR sequences, was employed. The acquisition protocol for these different scanners is summarised in Table 3.1.

3.2.2 MS PATHS cohort

MS PATHS is a learning health system in MS initiated in 2016, comprising a collaborative network of 10 healthcare centres that provide standardised routinely-acquired clinical and MRI data [133]. Among the clinical outcome measures collected in this large cohort, we used the PDDS [139] for our studies.

All MRI data for MS PATHS were obtained from Siemens scanner machines. The standard acquisition protocol included 3D T1 MPRAGE (TR=2300 ms, TE=2.96 ms, TI=900 ms, voxel size=$1\times1\times1mm^3$) and 3D T2-FLAIR (TR=5000 ms, TE=392 ms, TI=1800 ms, voxel size=$1 \times 1 \times 1mm^3$) sequences [133].

MS PATHS is a substantial cohort with over 20,000 MRI scans of people with MS. As we will see in the following Chapters, our purpose was to use patients from this external cohort to validate our experiments performed with the VHUH cohort, wherever feasible. To strike a balance between utilising our in-house dataset and the validation cohort, we deliberately selected a subset of 440 patients. It is important to note that, given our centre's involvement in this network, any subset extracted from the MS PATHS cohort consistently excluded data from our in-house cohort.

3.3 Image processing

Image processing is one of the most important steps in medical image analysis, especially when working with images acquired using different machines or in different centres and protocols. The application of distinct pre-processing steps helps us correct possible image artifacts and homogenise the dataset. Additional processing steps are specific to the particular target or task at hand. For brain MRI scans, common processing steps mainly centre around the segmentation of the different structures of interest. Below, we introduce the image processing steps applied to all the studied datasets.

3.3.1 Image pre-processing

The datasets used in this PhD Thesis are part of prospective cohorts that are continuously updated over time. The data were selected at the time of the Thesis study, and all image analyses were performed uniformly.

Bias field correction

Bias field signal, a low-frequency and smooth signal, can corrupt MR images by affecting the intensities of homogeneous tissue regions. Bias, inherent to MRI scans, primarily results from improper image acquisition processes.

Among the available strategies to perform bias field correction [142], we employed the well-known N4 algorithm [143], a non-parametric non-uniform normalisation approach. The N4 algorithm iteratively refines the bias field, capturing its smoothness and variations from the image data itself, resulting in a more robust correction. Bias corrected T1-w and FLAIR sequences can be observed in Figure 3.2.

Skull-stripping

Brain MRI scans contain non-brain tissue parts of the head, such as eyes, neck, or the skull. Skull-stripping, also known as brain extraction or whole brain segmentation, involves removing the skull, dura and scalp from the brain scan.

In our pre-processing pipeline, we used the HD-BET, a deep learning-based skull stripping method [144]. HD-BET demonstrated a higher capability to learn a better general representation of the brain tissues and the capacity to reduce the intrinsic reproducibility errors in comparison to past state-of-the-art methods [145]. The skull stripped sequences can be observed in Figure 3.2. We also used BET [146] to obtain the removed skull, since, so far, it remains as the only method providing this second output. The extracted skull will be later used for registration purposes.

Registration

In neuroimaging, registration involves aligning two or more images into a common space, ensuring spatial anatomical correspondence. It is a fundamental step for both intra- and inter- subject analysis. Intra-subject analysis aligns different sequences from the same subject (modalities co-registration), while inter-subject registration is necessary when the source and target spaces differ or multiple spaces exist in a dataset (see Figure 3.2). All subjects are moved to the same target space, which may be based on atlases or templates. The Montreal Neurological Institute (MNI) template is one of the most recognised neuro templates. For both objectives the same strategy is followed: deform the source image to match the target, as much as possible, by an optimisation process of

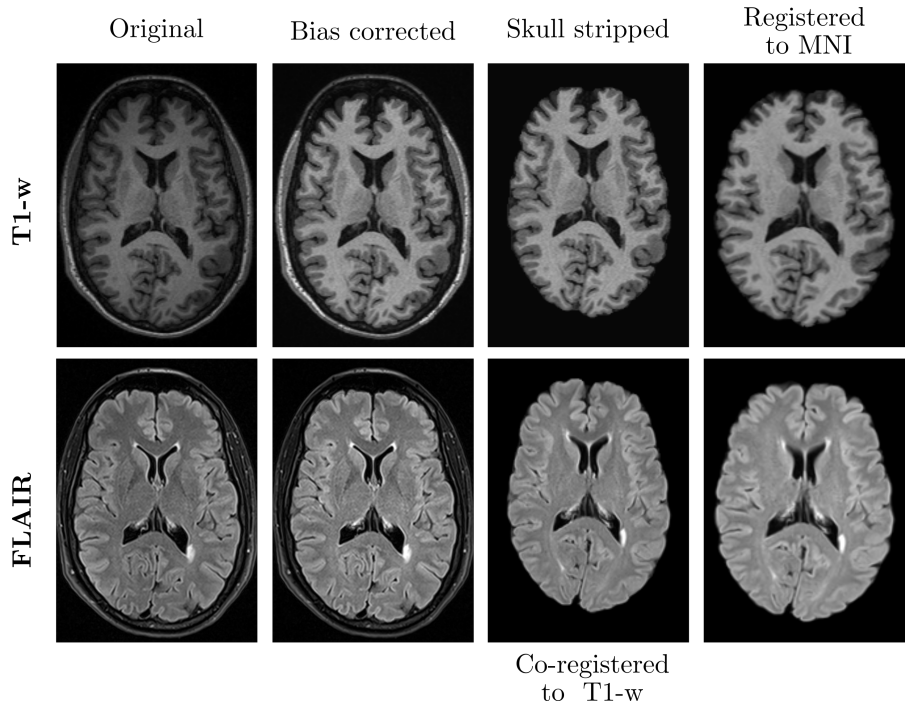

Figure 3.2: Pre-processing pipeline of T1-w and FLAIR scans. First, the original scans, T1-w and FLAIR, are bias corrected and skull stripped. The FLAIR sequences are co-registered to the T1-w, and both are registered to the MNI space.

different transformations. There are different registration techniques but mainly are based on two approaches:

- Rigid (or linear) and affine registration. It involves aligning two images by applying only rotation and translation transformations (on the x, y and z coordinate axes) without changing the shape of the images. The rigid and affine methods differ in the number of parametric transformations performed, since the affine registration method also includes scaling and shearing.

- Non-rigid (or non-linear) registration. It allows more complex transformations to account for differences in shape between images, such as scaling, shearing, and local deformations.

Linear registration is a common pre-processing approach to homogenise datasets without affecting local details. On the other hand, non-linear registration is an essential tool for morphological comparisons in longitudinal studies (intra-subject) or simply, to account for the local deformations suffered when registering the native space to a template.

In our work, we employed the *pairreg* function from FSL library [147] to linearly co-register all T1-w scans to their paired T2-FLAIR scans and co-register all T1-w scans in their native space to the MNI152 space. FSL is a library composed of analytic tools for different brain imaging data, including MRI [147]. It has been widely used for all kinds of image processing steps since its creation and up to date. The *pairreg* function allows for consistent skull scaling between scans [47]. Additionally, we performed non-linear registration of the T1-w to MNI for future use in tissue modulation (see Section 3.3.3).

Intensity normalisation

For the purpose of homogenisation, intensity normalisation plays an important role in all type of image analysis. Variability in acquisition scanners or protocols can lead to a wide spectrum of maximum intensity values in images. As a result, it is desirable to normalise these intensities. Histogram matching, Z-score or min-max are some of the most widely normalisation methods used for brain image analysis [148, 149]. Although depending on the final application or dataset, one method may be more appealing than another for our objectives. We consistently employed min-max normalisation approach at 95% confidence interval (CI) for all the scans to ensure that they share the same scale without allowing hyperintensities to dominate over other intensity values.

3.3.2 Automatic segmentation

Manual segmentation, specially manual lesion segmentation, is a challenging and highly time-consuming task, subject to the inter- and intra-rater variability. However, in clinical practice, it still stands as the *gold standard* procedure. Over the past decades, automatic and semi-automatic methods for lesion and tissue segmentation have emerged with the aim of facilitating this task, reducing time, and mitigating the inherent variability of manual annotations.

Here, we provide a brief definition and discuss the utility of the main automatic segmentation tasks commonly employed in MS studies. We also introduce the

specific algorithms used in our processing pipeline. In most cases, these processing steps have been performed with the purposes to improve the input data for the main pipeline.

Lesion segmentation

Over the last few decades, automatic lesion segmentation in MS studies has been a prominent area of investigation with the aim of surpassing the gold standard of manual segmentation, while also enhancing accuracy and efficiency. Despite of all the efforts, at this moment, none of the automatic algorithms has been taken as gold standard. Instead, a wide range of automated WM lesion segmentation techniques are available, with some incorporated into clinical routines [150]. We can find supervised approaches which use prior information, and unsupervised ones, which operate without prior knowledge. As discussed in Section 2.5, the introduction of deep learning models to perform such tasks, segmentations closer to human expert inter-rater variability. However, challenges related to domain adaptation for different protocols and acquisition machines persist in many cases. We used LST (Lesion Segmentation Tool) [151] an unsupervised algorithm based on thresholding hyperintensities from T1-w and FLAIR images to generate a lesion probability map.

Lesion filling

Lesion filling is a crucial intermediate step. Our primary focus is not the analysis of a brain image in the absence of its lesions. Nevertheless, when it comes to tissue and structure segmentation, the majority of algorithms are developed or built upon datasets comprising healthy brains, which typically do not account for hypointense WM lesions (in T1-w sequence). Lesion filling addresses this issue by inpainting lesions on the T1-w scan with signal intensities from normal-appearing WM before conducting tissue or structure segmentation. Thus, for most algorithms at least WM probability maps and lesion masks are required to compute this inpainting. Several lesion filling approaches have been proposed in recent years [85, 152, 153], resulting in a significant reduction in the errors associated with WM lesion tissue volume measurements [154].

We used the lesion filling algorithm presented in [153], which is based on a non-local patch match strategy, a widely adopted method in this field [155, 156] and demonstrates improved results compared to other publicly available methods.

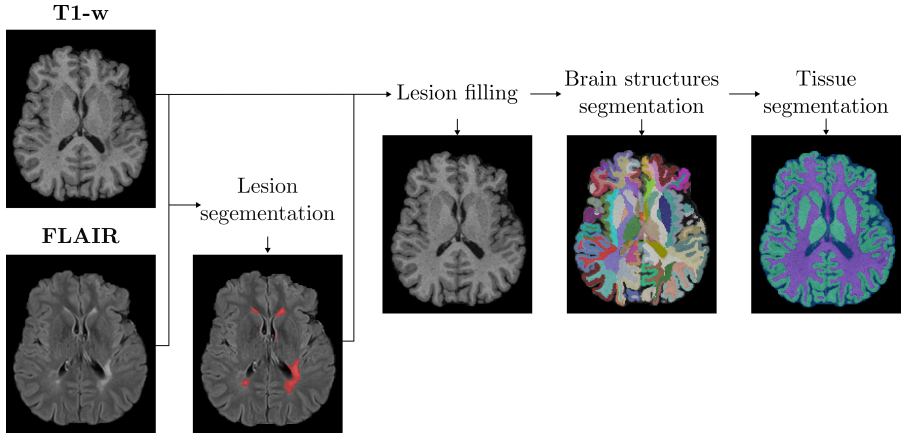

Figure 3.3: Automatic segmentation in MS. T1-w and T2-FLAIR scans are used to obtain a lesion mask, using LST. Afterwards, the lesion mask is used to perform lesion filling on the T1-w scan, which is used to obtain the structure segmentation, using FastSurfer and, joining the relative structures obtain the tissue segmentation.

An example of this resultant image can be observed in Figure 3.3.

Tissue segmentation

Tissue segmentation refers to the process of partitioning the brain into its three main tissues, i.e., WM, GM and CSF (see Figure 3.3). In practice, tissue segmentation is crucial for quantifying the volume of each of these regions and assessing their changes over time. In MS and other neurodegenerative diseases, brain atrophy measurement is of particular interest (see Section 2.2.2). The most commonly used algorithms take the T1-w scans as the input modality, given the clear contrast observed between these tissues [157].

As most of these algorithms do not account for the presence of lesions, T1-w scans must undergo lesion filling before tissue segmentation. We used FAST [158] as a widely accepted and efficient solution for tissue segmentation. This FAST segmentation was only used as input to the lesion filling algorithm [153]. While the masks used for volume calculations and other analyses were obtained by combining different structures representing WM and GM tissue obtained from structure segmentation (as discussed in the next Section).

Structure segmentation

Tissue segmentation is widely used in medical practice but only divides the brain into its three main tissues. This may be insufficient for conducting more comprehensive disease evolution analyses. Brain structure segmentation algorithms, on the other hand, provide a more detailed division of the brain. In this procedure, the three main brain tissues (WM, GM, and CSF) are subdivided into their substructures at various levels of detail, depending on the used segmentation method [159].

We used the FastSurfer pipeline, a deep learning-based model trained on atlases information, which segments the whole brain into 95 classes, including cortical and subcortical segmentation, as well as WM parcellations (see Figure 3.3) [160]. FastSurfer offers improved accuracy and speed compared to the well-known FreeSurfer software [161], which is based on the same atlas for brain segmentation.

3.3.3 Tissue modulation

We refer to tissue modulation as the process to represent actual local tissue volumes, GM or WM, in the original space after being transformed to a different space by means of the Jacobian modulation (see 2.2.2). The steps involved in GM tissue modulation are illustrated in Figure 3.4. The GM modulation was used in one of our experiments (see Chapter 5). As part of the pre-processing steps, the T1-w scan in its native space is first linearly registered to the MNI space (Figure 3.4(a)). Subsequently, the scan in its original space and the linearly registered one are non-linearly registered to the MNI space (Figure 3.4(b)). The linearly registered scan is used to obtain the GM probability map, i.e., tissue segmentation (Figure 3.4(c)). Meanwhile, from the non-linearly registered scan, we can extract the Jacobian determinant (Figure 3.4(d)), which characterises the deformation field experienced during the transformation. GM modulation is the result of multiplying these two maps (as shown in Figure 3.4(e)), i.e., the GM probability map and the Jacobian determinant. Finally, this product is smoothed using a Gaussian kernel with a full width at half maximum (FWHM) of 4.7 mm, equivalent to a standard deviation (σ) of 2, which is the minimum standard sigma value in the field of MRI data analysis (Figure 3.4(f)) [162]. This smoothing helps in detecting activation, influencing results interpretation, and enhancing the signal-to-noise ratio by reducing resolution.

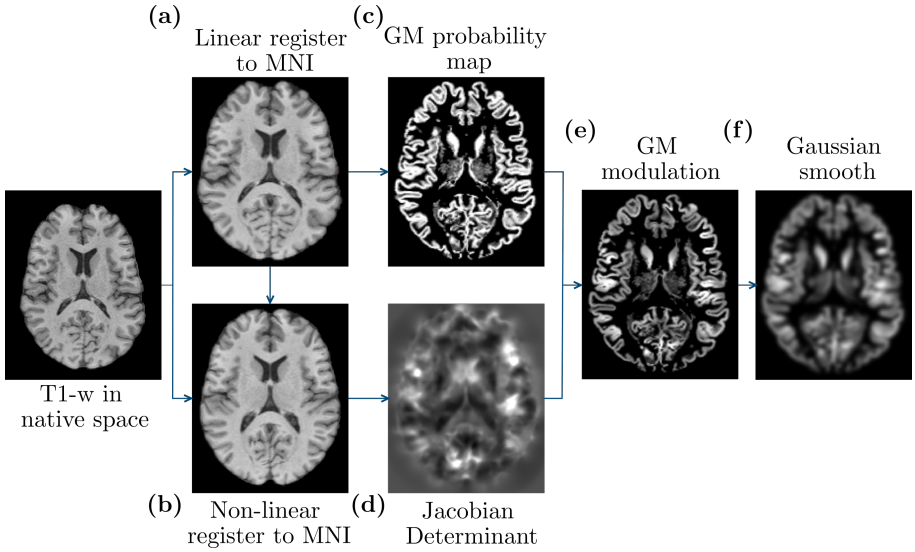

Figure 3.4: GM modulation steps. Registration to the MNI space (a) linearly and (b) non-linearly. From the linearly registered scan the GM probability map (c) is extracted and, from the non-linearly, the Jacobian determinant (d). The GM modulation (e) is the product of (c) and (d), smoothed with a Gaussian kernel (f). *GM*: grey matter.

Chapter 4

MS patients stratification

"Doing a PhD is difficult, not on a technical level, not on a scientific level, but on a human level."
A Doctor.

4.1 Introduction

As introduced in Chapter 2, the implementation of tools combining MRI scans and deep learning models has been gaining popularity in recent years, thanks to their ability to solve complex classification tasks. In classification tasks, the objective is to assign examples from the problem domain to specific class labels. To achieve this, it is crucial to extract relevant features from the provided examples in order to accurately assign them to the appropriate class.

In MS, and other medical fields, some classification tasks are hard to handle due to the complexity of images with, sometimes subjective other hard to delineate, labels, that are aimed to be solve with insufficient data cohorts. In clinical practice there are plenty of challenges that can be seen as a classification task, from diagnosis to patient prognosis, along with clinical phenotyping, e.g., for clinical trials. However, to be useful to the end-users, i.e., clinicians, more than a "simple" classification output probability is needed to justify the black

box character of the deep learning models. For this reason, studies including interpretability of proposed networks in their pipelines are preferred [32, 115, 117].

A cross-sectional study aims to analyse specific variables across a sample population at a single time-point, that might be the same or not over all studied data. In the specific case of a cross-sectional stratification study with people with MS, the aim is to assess if a single-data point is enough to describe the current status of each population individual. From a clinical point of view, a cross-sectional stratification study of patients with MS with different disability status presents two main interests. On one hand, this tool may be helpful to plan therapeutic or prevention interventions at a hospital level and screening purposes in clinical trials. On the other hand, the attention maps supporting the classification decision, revealing the model's focus, might help to interpret the mechanisms responsible for the accumulation of disability in MS.

In this Chapter, we propose to use a deep learning-based model to classify patients with MS based on their disability level at any cross-sectional time-point. We attempt to stratify patients with an EDSS$<$3.0, considered to have mild or no-disability, vs EDSS$\geq$3.0, recognised as moderate disability status [18, 137, 138]. Our goal is to do so only using brain MRI data as input to a deep learning-based model, aiming to investigate the potential contributions of these techniques within the context of the mentioned stratification. For that we only used as input to the network scans from two MRI sequences, T1-w and T2-FLAIR. To support the numerical performance and evaluate the robustness of the presented network, we (i) analysed the use of an interpretability algorithm to understand the reasons behind the decisions taken by the presented network; (ii) performed inference on an external independent dataset; and (iii) compared the deep learning results with the ones of a traditional machine learning algorithm, a logistic regression model.

4.1.1 State of the art

In the Section 2.5.3, we have introduced the studies with patients with MS using structural MRI as main input to a deep learning model. In addition to these studies, we can include other deep learning implementations that employed different input data rather than MRI, or initially pre-extracted image features before training the model (see Table 2.1). This was the case of Karaca *et al.* [113] with the first deep learning study performing MS subgroups

classification (RRMS=76, PPMS=6 and SPMS=38) using pre-extracted MRI features, specifically the lesions diameter (1D network). They demonstrated a higher performance using a deep learning model than with a SVM model. Other studies simply departed from a different type of data. By instance, Marzullo *et al.* [114] introduced a graph-CNN to classify patients with MS on four clinical profiles (12 CIS, 30 RRMS, 28 SPMS, 20 PPMS and 24 healthy controls) using diffusion-tensor imaging. Different 3-cross validation strategies were performed to binary classify the different phenotypes or the association of some of them.

Nowadays, the new deep learning approaches to perform phenotype classification are still coexisting with traditional machine learning models. Using pre-extracted MRI features, we highlight the work of Eshaghi *et al.* [116] with a large dataset composed of RRMS, SPMS and PPMS patients (6322 MS patients for training and an independent set of 3068 for testing), that were classified into different subtypes (cortex-led, normal appearing WM-led and lesion-led) not only based on cross-sectional information by to distinct temporal progression patterns. For that, they used SuStaIn [163], an unsupervised machine learning method that combines disease progression modelling and clustering models. Additionally, they showed the evolution of MRI abnormalities identified in each of the MRI-subtypes. In another study, De Meo *et al.* [164] presented a more standard stratification study of 1212 MS patients with MS and 196 healthy controls. They used linear regression and mixed-effects models to define the clinical and MRI features (only included for 172 MS patients and 50 healthy controls) of each cognitive phenotype.

4.2 Dataset

For training and testing this study we used 319 patients with MS from the VHUH dataset (see Section 3.2.1), with a total of 382 scans i.e., for 33 patients we used one or more MRI scans, performed at different time-points after the first attack, and which were considered as independent subjects. The selected subjects' acquisition dates run from 2010 to 2020 without any further inclusion criteria than the ones established to belong to that specific cohort, patients that had experienced a CIS before the age of 50. Each scan was matched with the first EDSS score obtained within the following six months after the MRI acquisition. The main demographic, clinical and brain MRI characteristics of these patients are summarised in Table 4.1.

Table 4.1: Demographic, clinical history and brain MRI characteristics of patients included in the analysis.

	Full cohort $N_{PAC/SCAN}$= 319/382	**EDSS<3.0** N=215/215	**EDSS≥3.0** N=104/167	***p*-value**
Female, n(%)	207 (65)	147 (64)	60 (58)	0.08
Age at CIS, years, mean[range]	32.3 [14-50]	32.4 [14-49]	32.2 [14-50]	0.78
Confirmed diagnosis, n(%)	260 (82)	160 (74)	100 (96)	<0.001
Age at diagnosis, years, mean[range]	33.2 [14-59]	33.5 [16-59]	32.7 [14-55]	0.43
CIS topography, n(%)				<0.001
Brainstem	83 (26)	59 (27)	24 (23)	
Optic nerve	98 (31)	74 (35)	24 (23)	
Spinal Cord	98 (31)	60 (28)	38 (37)	
Other	40 (12)	22 (10)	18 (17)	
MS topography, n(%)				<0.001
CIS	123 (39)	106 (49.5)	17 (16)	
SP	41 (13)	1 (0.5)	40 (39)	
RR	155 (48)	108 (50)	47 (45)	
Presence of OB, n(%)				<0.001
Positive	194 (61)	115 (53)	79 (75)	
Negative	75 (23)	64 (30)	11 (12)	
Unknown	50 (16)	36 (17)	14 (13)	
DD, years, mean(SD)	10.4 (7.0)	7.6 (6.6)	14.0 (5.6)	<0.001
EDSS, median[range]	2.0 [0.0-9.0]	1.5 [0.0-2.5]	5.0 [3.0-9.0]	<0.001
Lesion load, mL, mean(SD)	27.5 (39.9)	10.4 (13.1)	49.6 (50.7)	<0.001
Ventricles vol, mL, mean(SD)	29.7 (19.8)	21.5 (9.6)	40.2 (25.3)	<0.001
WM vol, mL, mean(SD)	694.8 (68.0)	707.5 (54.3)	678.3 (79.6)	<0.001
GM vol, mL, mean(SD)	787.9 (64.0)	813.1 (48.3)	755.4 (67.2)	<0.001

Brain vol, mL, mean(SD)	1542.7 (103.7)	1572.0 (83.7)	1504.8 (114.4)	<0.001
Scanner, n(%)				<0.001
Avanto	64 (17)	19 (9)	45 (27)	
Avanto Fit	64 (17)	43 (20)	21 (13)	
Symphony	10 (3)	7 (3)	3 (2)	
Symphony Tim	51 (13)	13 (6)	38 (23)	
Tim Trio	193 (50)	133 (62)	60 (36)	

[a] MS confirmed diagnosis by McDonald 2017 criteria. Two patients were confirmed before their first demyelinating attack.
PAC: patients, *EDSS*: expanded disability status scale, *CIS*: clinically isolated syndrome, *OB*: oligoclonal bands, *DD*: disease duration, *WM*: white matter, *GM*: grey matter

Additionally, a subset from the MS PATHS database (see Section 3.2.2) was used to independently test the performance of the proposed model. From this large database, we randomly selected a subset of 440 patients (440 scans), with representation of all grades of disability. Each scan was matched with the corresponding cross-sectional PDDS score, within an acceptation margins of ±6 months. The main demographic, clinical and MRI characteristics of this subset of patients is summarised in Table 4.2. All scans from both datasets were pre-processed with the different steps and proposed algorithms defined in the previous Chapter (see Section 3.3.1).

4.3 Proposed model

We proposed the use of an end-to-end deep learning pipeline to stratify patients with MS based on their disability status at a cross-sectional time-point. Rather than trying to allocate patients depending on their phenotypic clinical profile, we attempted to associate the clinical outcome score (EDSS/PDDS) with the closest acquisition MRI scan. For that, we implemented a model to solve a binary classification task aiming to stratify patients with an EDSS$<$ or $\geq$3.0.

The full deep learning pipeline is represented in Figure 4.1. T1-w and T2-FLAIR scans, previously pre-processed, containing whole brain information are used as unique input to train and test the network. During inference, the accuracy on stratifying patients with MS has been reported, based on the network

Table 4.2: Demographic, clinical and brain MRI characteristics of patients from MS PATHS included in the analysis.

	Full cohort N=440	**PDDS<3.0** N=220	**PDSS≥3.0** N=220	***p*-value**
Female, n(%)	310 (70)	170 (77)	140 (64)	0.007
Age at diagnosis, years, mean[range]	36.8 [19-69]	36.1 [19-62]	37.6 [19-69]	0.24
DD, years, mean(SD)	11.5 (9.1)	8.5 (7.6)	14.9 (9.5)	<0.001
PDDS, median[range]	2.5 [0.0-7.0]	0.5 [0.0-2.0]	5.0 [3.0-7.0]	<0.001
Lesion load, mL, mean(SD)	13.2 (24.7)	7.9 (11.7)	18.5 (32.0)	<0.001
Ventricles vol, mL, mean(SD)	35.9 (24.0)	28.7 (17.4)	43.1 (27.4)	<0.001
WM vol, mL, mean(SD)	706.1 (53.4)	718.6 (47.3)	693.6 (56.3)	<0.001
GM vol, mL, mean(SD)	804.6 (73.5)	828.4 (64.6)	779.7 (74.0)	<0.001
Brain vol, mL, mean(SD)	1496.4 (92.8)	1526.2 (79.7)	1466.5 (95.7)	<0.001

PDDS: Patient Determined Disease Steps, *DD*: disease duration, *WM*: white matter, *GM*: grey matter, *vol*: volume.

output probabilities, thresholded at 0.5. Besides that, the LRP explainability algorithm [73] is used to capture the most relevant regions in the brain that lead to the output prediction.

The main pipeline is also evaluated in two different ways: (i) calculating the performance of the model using an external independent dataset, and (ii) comparing the results obtained with the proposed deep learning model with the ones obtained using a traditional machine learning model such as logistic regression.

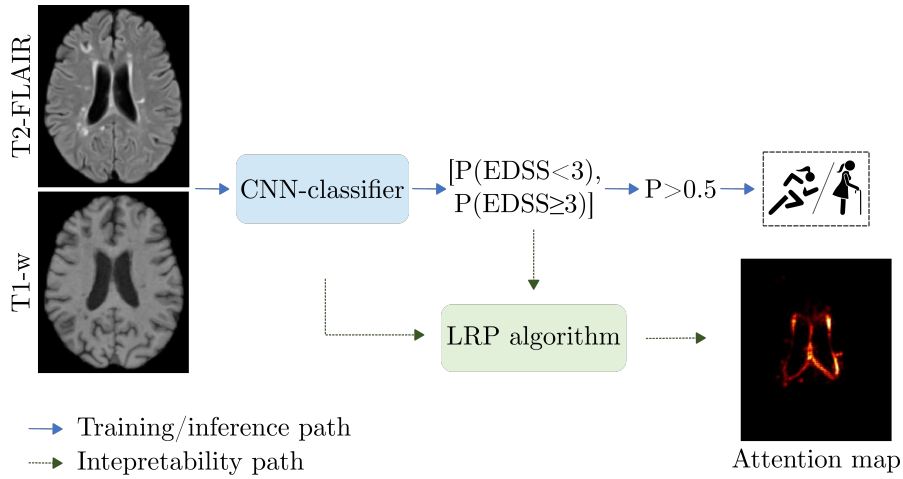

Figure 4.1: Proposed deep learning pipeline. Training and inference paths are drown along with the interpretability path based on the use of the back-propagation LRP algorithm. *LRP*: layer-wise relevance propagation, *P*: probability, *EDSS*: expanded disability status scale, *CNN*: convolutional neural network.

4.3.1 Network architecture

The proposed CNN is based on a modified ResNet architecture [56], built with 3D layers. The included modifications are mainly based on reducing the final number of model parameters, thus reducing model's complexity. Each residual block is based on 3D convolutional layers that produce $3 \times 3 \times 3$ and $1 \times 1 \times 1$ kernel convolutional layers, obtaining two different feature maps. Both maps are aggregated and the resultant map is normalised with batch normalisation and activated with a LeakyReLU. As shown in Figure 4.2, the architecture is composed of four residual blocks with an increasing number of kernels k (16, 32, 64 and 128), followed by a $2 \times 2 \times 2$ downscale max pooling operation. Afterwards, the extracted feature map is projected in a GAP layer to reduce feature dimensionality and to allow independence of the input size. The final classification layer, fully connected in the standard ResNet, is replaced by three successive $1 \times 1 \times 1$ 3D convolutional layers, with k=128, 64, 2, where the first two are activated with a ReLU and the last one with a Softmax, obtaining as output the probability to belong to one or the other class.

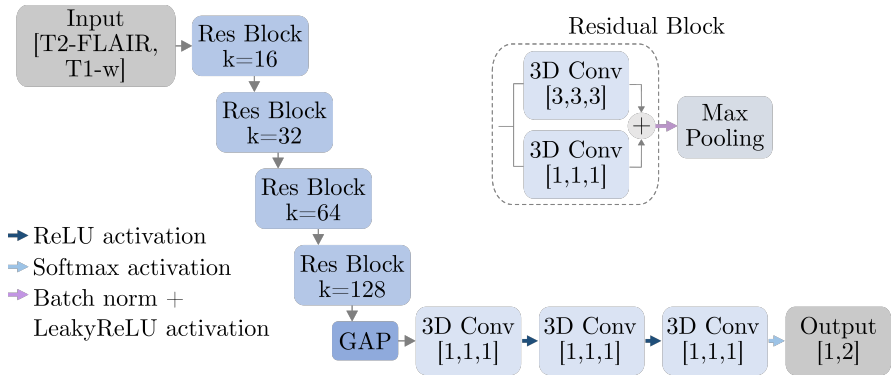

Figure 4.2: 3D-CNN based on the ResNet architecture. *Res block*: residual block, *Conv*: convolutional layer, k: number of kernels, *GAP*: global adaptive max pooling, $ReLU$: rectified linear unit.

4

4.3.2 Training and inference procedures

The model was trained only using the VHUH dataset. Due to the size of the dataset, a 7-fold patient cross-validation strategy was used to train and test the model (see Figure 4.3(a) for a graphic representation of this procedure). We sampled the folds to keep the same class distribution in each one, while following the distribution present along the dataset (dependent on the associated cross-sectional EDSS of each scan). Finally, in each iteration five folds were used for training, one fold for validation and the last one for inference.

The input network's patch size was $2 \times 144 \times 184 \times 152$ (*channels* $\times$ *height* $\times$ *width* $\times$ *depth*). To mitigate the size of the data available for training –and considering the class imbalance– we used data augmentation. The data augmentation was performed on-the-fly, when batch calling, and it consisted on the application of an axial flip to all scans with an EDSS$\geq$3.0 (i.e., doubling the samples) and to 75% of the already larger class, EDSS$<$3.0, considering the difference in class-size. Additionally, a random Gaussian noise (σ=0.02) was applied on both input channels (T1-w and T2-FLAIR) to create intensity variation.

We trained the model for a maximum of 200 epochs, with a fixed batch size of two, due to the large input size. We used an early stopping strategy based on the validation loss behaviour to prevent overfitting, i.e., to avoid training until a point that the network is memorising instead of learning features. The model

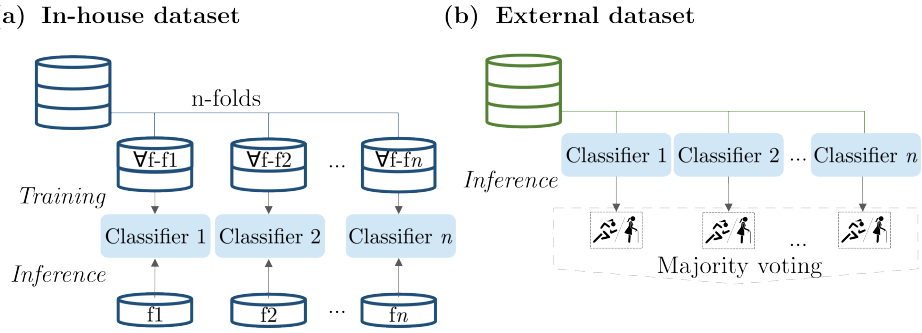

Figure 4.3: Training and testing procedure. (a) The training is only performed on the VHUH dataset with a 7-fold cross validation strategy, obtaining 7 different models. Afterwards, these models are evaluated on the left fold and, (b) on the MS PATHS validation dataset, where the final classification is obtained with the majority voting of all the tested models. f:fold.

was optimised with Adam [165] with a learning rate decay also dependent on the validation loss and the current epoch, $lr \times 0.99^{epoch}$, with an initial learning rate of 10^{-5}. The model was trained by minimising a weighted cross-entropy as loss function. The weights were also proportional to the class-size, [1.0, 1.5] for EDSS<3.0 and EDSS≥3.0, respectively.

Inference was performed in each iteration on the scans composing the left fold not used for training either validation. With the provided probabilities of belonging to one class or the other at the network's output, a final classification is obtained by thresholding them at a fixed 0.5.

Without any retraining or fine-tuning, as shown in Figure 4.3(b), the whole external validation subset from MS PATHS was evaluated on each of the VHUH trained models. The final reported results were obtained using a majority voting of the seven different models.

4.3.3 Evaluation

To evaluate the performance of the different models, deep learning and logistic regression, we used the following metrics:

- Sensitivity of correctly classified subjects with EDSS≥3.0 thresholded at 0.5, Sensitivity = $\frac{TP}{TP+FN}$

- Specificity of correctly classified subjects with EDSS<3.0 thresholded at 0.5, Specificity= $\frac{TN}{TN+FP}$
- Balanced accuracy is adjusted to perform better in imbalanced data, by calculating the average accuracy for each class, instead of combining them as in the standard accuracy. We define balanced accuracy (from this point on we will refer to it as accuracy) in correctly classifying each patient by means of their EDSS, Accuracy = $\frac{Sensitivity+Specificity}{2}$

where true positives (TP) are the number of correctly classified patients with EDSS≥3.0, true negatives (TN) are the number of correctly classified patients with EDSS<3.0, false positives (FP) are the number of patients with EDSS<3.0 classified as they have EDSS≥3.0, and false negatives (FN) are the number of patients with EDSS≥3.0 classified as they have EDSS<3.0. The results are reported in terms of mean and standard deviation for the 7-fold cross validation computed on the VHUH dataset and as majority voting of the 7 models for the external validation dataset.

We also performed statistical analyses on the descriptive characteristics of patients between the two studied groups, EDSS/PDDS<3.0 and EDSS/ PDDS≥3.0. Chi-square or mixed-effect linear regression models accounting for repeated measures were used, as appropriate.

4.4 Proposed interpretability

To support the quantitative performance results we propose two different interpretability implementations. On one hand, we propose to use a traditional machine learning algorithm, the logistic regression, with pre-extracted MRI volumetries, to compare its performance with the one of our deep learning proposal. On the other one, our deep learning model is submitted to its own interpretability, using a back-propagation method to explain its decisions, specifically using the LRP method [73].

4.4.1 Logistic regression

A logistic regression-based model is a simple but powerful traditional machine learning model. Additionally, it is considered an intrinsic explainable model since it is self-interpretable, i.e., the features importance can be intuited from the

coefficients of the logistic model and the exact way a model reaches its conclusions is clear enough by its predictions result of a weighted sum of input features. For that, we built a logistic regression model to assess whether our CNN proposal is superior to a self-interpretable machine learning model.

We considered the disability class (EDSS/PDDS<3.0 and EDSS/PDDS≥3.0) as dependent variable (output) and pre-extracted volumetric measures from the MRI scans as the explanatory variables (input). The explanatory variables consisted of different anatomical regions in the brain closely related to atrophy measures and presence of demyelination. Those included the WM, GM, lateral ventricles and total intracranial volumes extracted by using automatic structural segmentation as explained in a previous Chapter (see 3.3.2), as well as, the brain lesion volume (or lesion load), calculated from the automatically extracted masks.

4.4.2 Attention maps: LRP

As has been introduced in a previous Chapter (see Section 2.4), LRP is an interpretability post-hoc model-specific framework, giving local explanations [73]. The post-hoc character is implicit in a deep learning model, and the model specific is caused by the back-propagation performed by decoding the resulting classification output through the network, propagating the relevance layer by layer, obtaining a heatmap on the input space with each voxel contribution. Obtaining the relevance voxel by voxel is what defines the local property of this algorithm. The LRP follows the principle of conservative relevance, i.e., the total relevance is conserved per layer.

The implementation of this method was inspired in *Pytorch-LRP* [166], adapted to our specific CNN. In this implementation, they are following the β-rule [167]. In this rule, both the positive and negative contributions of each node in the previous layer to the node in the current layer are taken into account. The β parameter allows to adjust how much weight is put on positive contributions relative to inhibitory contributions. We decided to use β=0, i.e., only reflecting the positive contributions that led the classification to the winning class. The decision of choosing only the positive contributions was made based on previous literature findings [166] and corroborated with preliminary analyses on our own data that showed that the negative contributions did not add additional information.

For this study, we extracted the individual LRP heatmaps for each scan of the

training VHUH dataset, during inference. The resultant heatmaps were evaluated (i) individually and (ii) as a class-average prediction.

Individual attention map analyses

Individually, the attention maps were assessed visually and qualitatively showing which voxels contributed the most to the classification given by the model inference. Additionally, semi-quantitative analyses of the individual attention maps were carried out, multiplying such maps by a parcellation map. This allowed us to classify the relevant voxels for the CNN decision into the different anatomical areas. For this purpose, we set a threshold at the 95% percentile of positive relevance, to capture the most relevant areas in each case, although other thresholds have been studied, without presenting relevant variations.

Class-average attention map analyses

Class-average attention maps were also built for visual inspection, through averaging the individual values of the attention maps across the subjects of each one of the prediction results: TP, FP, FN, and TN. In addition, each class-average map was multiplied by the parcellation map, and a mean value per anatomical area was obtained. This allowed us to identify the anatomical areas with greatest relevance in each disability status.

Voxel-wise regression analyses

We then carried out a quantitative analysis based on the LRP heatmaps aiming at investigating to what extent the variability within each voxel (across independent subjects) of the attention map could be explained by the presence of lesions and atrophy, the best-known contributors to disability in MS, at the voxel level.

For that, we first grouped our subjects by the predicted vs real class (i.e., TP, FP, FN, TN). We then computed voxel-wise regression models, one per each prediction group (i.e., TP, FN, FP, TN), where the LRP value at each voxel (considering all individual LRP heatmaps of the same group) was the dependent variable. As explanatory variables of these voxel-wise regression models we included: (i) voxel-wise binary indicator of lesion (obtained from individual T2-FLAIR scans), and (ii) voxel-wise value describing the deformation suffered by the T1-w scan when moving it to a common space (MNI), calculated with the Jacobian determinant [168]. The latter variable (ii) was used as proxy for

atrophy, being aware that the common template is based on healthy controls. After this step we obtained four voxel-wise maps of R-squared values, where each voxel indicated the proportion of the variability of the LRP value that could be explained by the presence of lesions and native-to-MNI deformations. Additionally, we estimated the standardised beta coefficient for each one of the explanatory variables, which indicated the relative importance of each one of these for the prediction of the dependent variable, being the two standardised beta maps comparable (since they were in the same scale).

4.5 Results

4.5.1 Evaluation on VHUH

The descriptive statistical analysis, comparing the different demographics, clinical and MRI characteristics between both analysed groups (shown in Table 4.1), revealed that most of them were significantly different (p-value$<$0.05). Patients with an EDSS$\geq$3.0 had a longer disease duration than patients with EDSS$<$3.0, with a similar age at CIS with a mean of approximately 30 years old (SD=8). Compared with patients with EDSS$<$3.0, patients with EDSS$\geq$3.0 had lower tissue volumes, GM, WM and intracranial volume, and higher ventricle volume and lesion load.

Performing inference on the whole dataset using the 7-fold cross validation, we obtained an average accuracy of 79% (SD=4%), with a sensitivity of 77% (SD=5%) identifying patients with EDSS$\geq$3.0, and a specificity of 81% (SD=9%) of patients with EDSS$<$3.0. Figure 4.4(a,b) shows examples of incorrectly classified patients on this dataset.

4.5.2 Evaluation on MS PAHTS

Similarly to the results obtained in the VHUH dataset, MS PATHS dataset descriptive analysis (see Table 4.2) showed that the characteristics intrinsically related to the disease were significantly different between the two groups. For instance, patients with greater disability (PDDS$\geq$3.0) exhibited longer disease duration, higher lesion load, and other MRI-quantified tissue volumes. In contrast, age at CIS was not significantly different between groups (p=0.24), presenting a similar mean of around 36 years old.

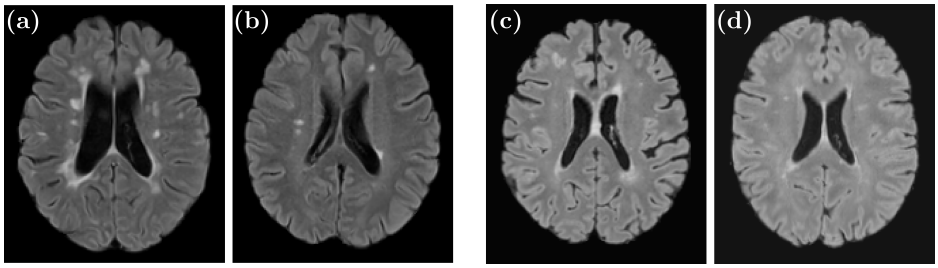

Figure 4.4: Example of incorrectly classified patients. From VHUH dataset (a) with an EDSS=2.0 was predicted as EDSS≥3.0 and, (b) with an EDSS=4.0 was predicted as EDSS<3.0. From the MS PATHS dataset (c) with a PDDS=2.0 was predicted as PDDS≥3.0 and, (d) with a PDDS=4.0 was predicted as PDDS<3.0.

Regarding the model's inference performance, on the models only trained on the VHUH dataset, the majority voting across the different fold-models showed an accuracy of 71%, a sensitivity of 68% and a specificity of 75%. Figure 4.4(c,d) shows examples of incorrectly classified patients on this dataset.

4.5.3 Comparison with a logistic regression

The logistic regression model built with the five brain volumes corresponding to WM, GM, ventricles, intracranial volume and lesion load, achieved an accuracy of 77% (SD=7%), with a sensitivity of 68% (SD=10%) –when classifying patients with EDSS≥3.0– and a specificity of 86% (SD=6%). Therefore, the logistic regression model trained with MRI pre-extracted features showed a 10% lower sensitivity than deep learning-based models. When considering other input combinations with fewer features, the logistic regression model obtained always inferior results than the CNN model and, when the logistic regression models only had a single feature, the accuracies dropped to 50-55%. These low accuracies, close to 50%, were also found when performing inference on the same logistic regression models, trained with VHUH data, with the MS PATHS dataset.

4.5.4 Attention maps analysis

Individual attention maps

Individual attention maps showed that, in both groups, the most relevant voxels that led the classification decisions were mainly located in the periventricular WM

regions, which often contained demyelinating lesions, and frontal and temporal cortical areas (see example Figure 4.5(a)). Figure 4.5(b) shows a case-example of the distribution of relevant voxels across the different anatomical regions, considering all voxels with a relevance above a 95% percentile. Other thresholds were also explored. However, non-significant changes in the distribution were observed.

Class-average attention maps

Class-average attention maps revealed that the most relevant areas were the frontal cortex, cerebellar cortex, periventricular WM, temporal cortex and lateral ventricles, as shown in Figure 4.6. The ranking of these regions slightly varied across the different groups (TP, FP, TN, FN). For the patients classified with EDSS≥3.0 (TP, FP), the relevance of periventricular and frontal WM, where most lesions are located, was particularly high. Instead, for no or mild-disability statuses (TN, FN), the relevance of the cortex, especially frontal and temporal cortical areas and that of the cerebellum, was the highest.

Voxel-wise regression

After carrying out the voxel-wise regressions on the VHUH dataset, we obtained R-squared maps for the four prediction groups (Figure 4.7). In general, the R-squared values were low in all four groups, and mostly ranged between 0 and 0.2. In the correctly classified groups (TP, TN), the R-squared values were even lower (0.0–0.1). That means that in these groups, at most the 80% of the variability of the attention map could not be explained by the presence of lesions (observed in T2-FLAIR scans) or the native-to-MNI deformation (of the T1-w scans), calculated through the Jacobian determinant. When focused on the maps of partial R-squared, we found that values were always greater for the presence of lesions than for the Jacobian determinant (Figure 4.7).

4.6 Discussion

In this Chapter, we have investigated the use of deep learning models to classify patients with MS according to their status while trying to provide an explanation of the decisions taken by the CNN. Our findings showed that a deep learning model using only brain MRI scans, without any guidance, was able to stratify

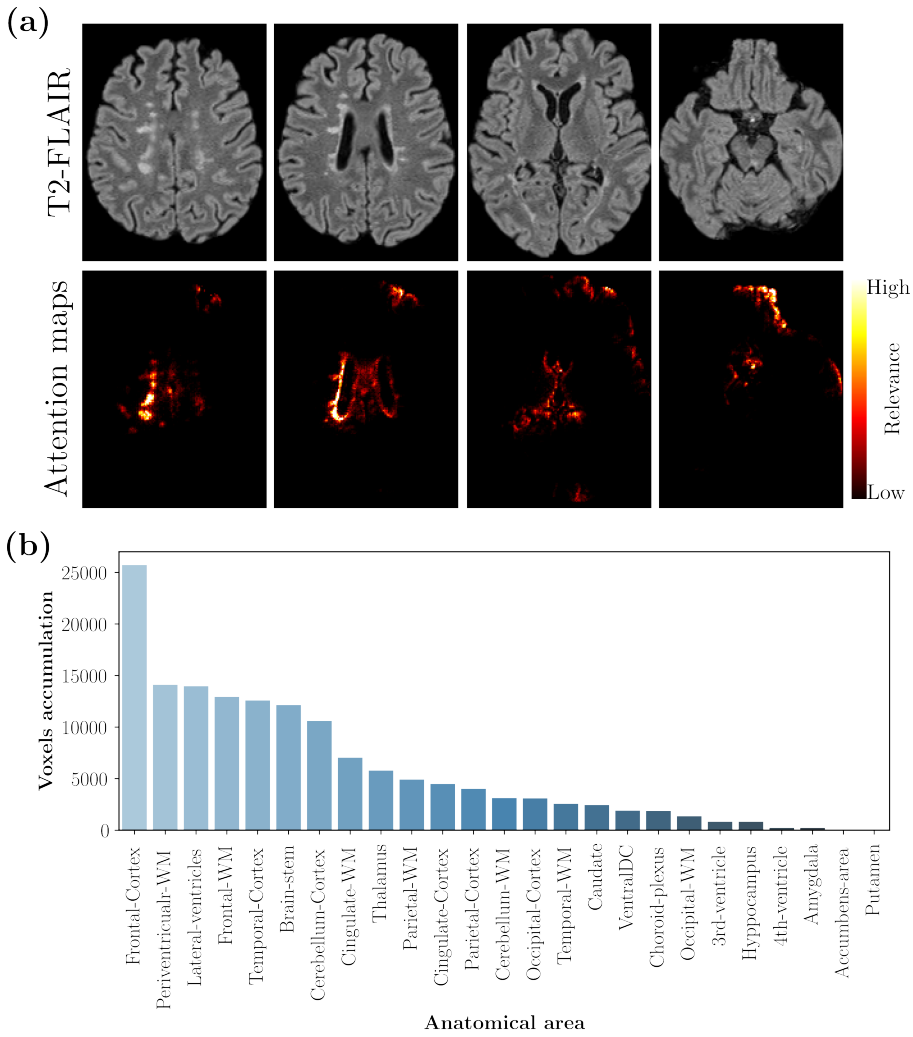

Figure 4.5: Example of individual attention map analysis. This MS patient was correctly classified as moderate disability with an EDSS=6.0. (a) Different T2-FLAIR slices with their corresponding obtained attention map. (b) Relevant-voxel accumulation by anatomical area. In this case, the frontal cortex and periventricular WM were the most relevant areas leading the decision.

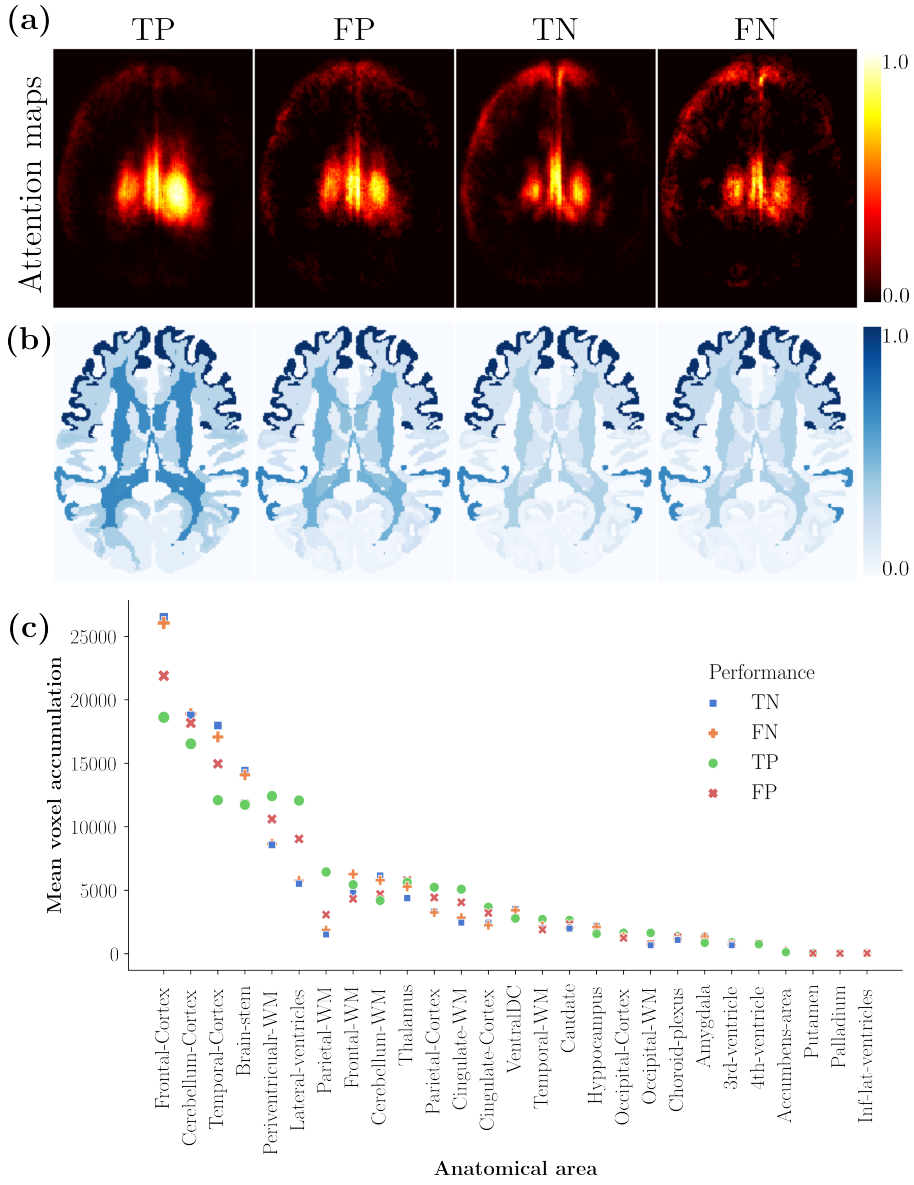

Figure 4.6: Class-average attention map analysis. (a) Class-average attention maps (TP, FP, TN and FN) binarised at 95% percentile. (b) Brain parcellations with the mean attention value (normalised across groups) attributed to each anatomical area. (c) Mean voxel accumulation by each anatomical area on the class-average group (without group normalisation).

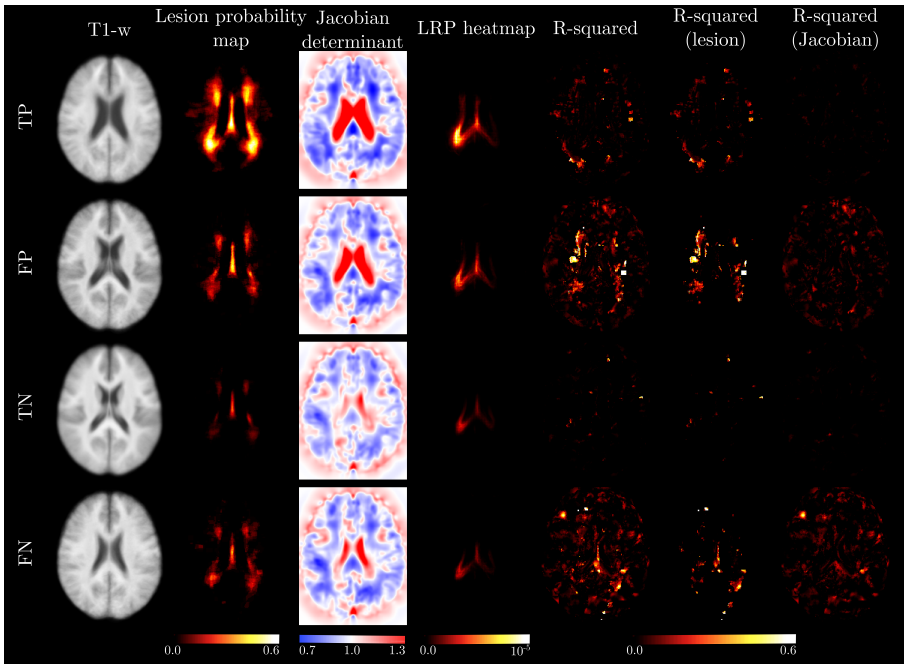

Figure 4.7: Voxel-wise regression analysis. Average map for the TP, FP, TN and FN of the T1-w scan, lesion probability map, Jacobian determinant, LRP heatmap and the R-squared map obtained from the voxel-wise regression model built with the individual attention maps as dependent variable, and the individual lesion masks and individual Jacobian determinants as explanatory ones. The partial R-squared on each separate variable is also represented. The Jacobian determinant is represented in terms of expansion (values>1.0) and compression (values<1.0)

patients with MS with high accuracy (79%) when testing data from the same cohort, and 71% when testing images from an unseen database. Furthermore, the comparison with the logistic regression model brought to light the superiority of the CNN model. Finally, our attention map analyses revealed that the most relevant anatomical areas that the CNN model used to decide the level of patients' disability were the frontotemporal cortex and the cerebellum and did not depend on the mere presence of lesions or atrophy in these locations.

Compared with previously reported studies that used deep learning models to perform predictions of MS progression, our work presents some similarities and some major differences. Despite the differences in the target of the classification,

i.e., cross-sectional vs future predictions, all published studies [32, 128], like ours, used the whole brain as input to the network. However, in our study, we only used routinely-acquired T1-w and T2-FLAIR MRI data, whereas other studies required the use of several MRI modalities and additional masks [128] to achieve similar accuracies. Of note, we validated our CNN model on an external, unseen cohort, where the performance of the classification task was more than acceptable, without needing any recalibration or re-training of the model. Furthermore, having at hand a tool to discriminate patients according to their disability level may be extremely useful in contexts where the EDSS score is not obtained as part of routine practice [169]. Such a discriminating tool may allow the fast identification of patients who already present a certain degree of disability and who, therefore, are at high risk of reaching unfavourable clinical outcomes [170].

Another important remark is that we demonstrated the superiority of our CNN model when compared with a logistic regression model. Notice that the databases used were conformed of different scanner models and magnetic fields, to which the deep learning models were more robust, having less than 10% drop in performance when testing on the external database. On the other hand, logistic regression models are based on obtained volumetric measurements that need to be previously computed using different tools which can be affected by changes on the MRI scanners and image protocols used to acquire the data. Thus, considering that volumetric measures have been frequently used as outcome measures in clinical trials [171], our findings suggest that deep learning-based discriminative tools might also be considered as trial outcomes, given that they seem to explain clinical outcome better than conventional volumetric measures feeding into a logistic regression model. Further research in this regard, evaluating the CNN model's sensitivity to clinical change, is therefore warranted.

Finally, in this study we attempted to understand the reasons behind the decisions made by the CNN through qualitative and quantitative analyses of the individual and class-average attention maps derived from the CNN model. To the best of our knowledge, this is the first large study focusing on giving an explanation to the performance obtained using a CNN when trying to classify patients into different disability categories. Some other studies aiming at solving similar classification problems have also focused on the visualisation of individual attention maps, but without performing further analyses on them [32]. Moreover, other studies have mainly investigated the classification between healthy controls and patients [115, 166, 172], whereas the task of classifying patients into different

disability classes, as in our study, is possibly much more challenging, given the continuous nature of the disease. Thus, we went a step further trying to apply such methods on patients only, i.e., moderately disabled and mildly or non-disabled, with a priori no clear pathological boundaries between them. Making their identification more difficult by providing only brain MRI data to the model. With these analyses we found that the voxels with the highest relevance for the CNN to make the final decisions were in the frontal and temporal cortex, followed by brainstem and cerebellum, and periventricular WM.

The importance of cortical GM for disease progression has been suggested in several studies, although the proposed underlying pathological mechanisms differ across studies [173, 174, 175, 176, 177]. However, no proper head-to-head comparisons across mechanisms have been carried out, which makes the hierarchy or dynamics of such pernicious pathological events difficult to understand. In any case, it is plausible that our CNN has captured at least some of the imaging features related to them. The consistency across individuals in terms of these areas likely reveal genuine differences between subjects, possibly related to atrophy, although not only. Looking at the results of the voxel-wise regression (Figure 4.7), it is possible that other pathological aspects, for example changes in image texture denoting underlying non-obvious demyelination, may be playing a role too. On the other hand, it must be acknowledged that some of the highlighted attentions in the cortex might also be caused by non-pathological aspects such as registration inaccuracies. Although the high accuracy of the model might lead us to think that the potential impact of registration inaccuracies on model performance was not major, this would need to be assessed in further studies. That is, we were unable to discern if all the attentions were related or not to pathological aspects, which is a limitation of the current study. It should also be highlighted the involvement of brainstem and cerebellum, whose role in disability progression in MS has been repeatedly seen in the literature [178, 179]. Future studies should investigate whether those relevant imaging aspects present in the brainstem and cerebellum have the same pathological translation as those of the relevant cortical areas. The relevance of periventricular WM may be related to the presence of demyelinating lesions, which tend to appear in this location [1], but also to lateral and third ventricle enlargement, which has been shown to play a role too [180]. Agreeing with that, cases with moderate disability presented a higher mean lesion load in periventricular WM than in mildly or non-disabled cases. Interestingly, these structures are followed by the thalamus, which also

appears to be very relevant for the CNNs, in line with previous studies showing the importance of deep GM structures for disease progression in MS [181]. All these relevant areas were brought to light through both the individual attention map and the class-average map analyses. To the best of our knowledge, this is the first time that such areas have been identified as relevant for clinical progression in MS in a completely hypothesis-free manner, that is, without forcing the model to pay attention to them. Regarding the voxel-wise regression analysis, the most striking finding was that the attention within a voxel, which indicates the importance of a given voxel to decide the disability class of a given patient, was not only explained by the mere presence of lesions or native-to-MNI deformation (used as proxy for volume change) in that particular voxel. This may suggest that deep learning models pay attention to more general aspects of the image, possibly focusing on complex spatial relationships between voxel-wise information, considering that distributional features of brain lesions might impact on disability progression, or image texture-related information, maybe denoting microscopic processes such as underlying demyelination [173, 182]. Of note, these more general aspects of the image deserve further research and, anyway, cannot be explained by the presence or absence of lesions or atrophy in a given point.

In any case, our findings strongly support the use of deep learning models to perform classification and prediction tasks where the input data is derived from images. This confirms the potential of deep learning models to unveil key aspects of the disease which are uncatchable by the human eye. Future studies focusing on unveiling these aspects are therefore warranted. Potential implications for clinical practice of our findings may include, in the short term, the application of CNN-based models to automatically classify patients according to their disability status using only routinely acquired brain MRI scans. This may be extremely useful in situations where large-scale therapeutic interventions, which may vary depending on the disability status of a given patient population, need to be planned in a relatively rapid manner. Other, mid-term implications include those derived from the development of CNN-based models to predict future disability, which may be extremely powerful to manage patient expectations and design tailored treatment strategies. That is, we should acknowledge that this is a cross-sectional study and, therefore, no strong statements about prediction of disease prognosis based on our specific CNN model can be made. However, we believe that our findings will help build powerful predictive models, possibly focusing on those areas identified as highly relevant through the attention map

analyses. Up to now, most of those CNN-based models for future prediction that have been published so far show a relatively limited accuracy, implying that more research is sorely needed [31, 32, 128].

Among the methodological considerations and possible limitations of our study, it is worth mentioning the relatively small sample size, considering the data needs of deep learning models. For this reason and to make the most of the data available, we did not apply any restriction to the disease duration of our patients and considered each individual scan which could be matched to an EDSS score as an independent piece of information. As future work, we would like to analyse the impact of adding clinical or demographic data, such as disease duration, age at CIS or sex, on model performance. Additionally, we applied data augmentation strategies when possible. As a result of all these strategies, the final database used for training, overall, had enough variability, proving robustness and generalisation of our models to scans from the same vendor. Even though the data from the different MR scanner models used for training were not balanced, they were so in terms of strength field. This was confirmed by the excellent reproducibility of the models when we applied them to the external validation cohort, despite the fact that the disability score was not the same as the one used for training. Another remark relates to the fact that we did not account for potential disease-modifying treatment effects. Future studies accounting for these are therefore warranted. Concerning the attention maps, they are limited by the lack of a ground truth. Surely, other methodological choices could have been made, providing slightly different results, which deserve further research. Moreover, we tried to relate the attention maps with well-known biomarkers of the disease, whereas future studies should investigate the association with new, possibly promising, biomarkers.

In conclusion, the results of this Chapter, show that our CNN model was able to stratify patients with MS based on their disability score solely using a single time-point brain MRI (T1-w and T2-FLAIR sequences) providing a high performance. Furthermore, our findings bring to light the potential of deep learning models to provide key information about the mechanisms responsible for the accumulation of disability in MS, suggesting the relevant role of frontotemporal cortex and cerebellum for the development of irreversible disability. Importantly, these findings may have immediate and especially long-term implications for clinical practice, laying the foundations for building powerful predictive models of future disability.

Chapter 5

Regional approaches for MS patients stratification

"There comes a moment when it goes 'on wheels' (rusty wheels that often deflate, but still, on wheels)."
Now, a Doctor.

5.1 Introduction

As seen in the previous Chapter 4, deep learning models are able to accurately predict relevant features to associate a cross-sectional MRI scan with its corresponding binary disability status. As we noted in previous Chapters, the vast majority of proposed classification and prognostic tasks in MS state-of-the-art use the whole brain volume as input to the network design [32, 115, 123, 130].

Several studies have established associations between the progression of MS and specific biomarkers, such as the analysis of brain regional volumetric changes in GM or WM atrophy [22, 183]. In this Chapter, our objective is to determine whether these specific regions, along with others, play a crucial role in the cross-sectional stratification task of differentiating patients with no-disability or mild disability. To achieve this, we assess the use of different brain regions as

the input for the same CNN architecture introduced in the previous Chapter 4 to perform MS stratification.

The selected regions are mainly related to regional biomarkers present in all subjects, which may play a relevant role in the accumulation of disability in MS: (i) WM and (ii) GM tissues, (iii) subcortical GM, (iv) lateral ventricles, and (v) brainstem and cerebellum structures (BSC from now on). Using the same datasets presented in the previous Chapter, we will study different ways to extract and evaluate these different regions, that ultimately are aimed to be compared with the whole brain input strategy.

5.1.1 State of the art in neuroimaging

As previously seen (see Section 2.5.3), the state of the art regarding stratification for patients with MS with deep learning using raw brain MRI as main input is still limited. For this reason, we will briefly make this problem extensive to other neurodegenerative diseases, such as Alzheimer's Disease, which have been more widely studied [184], also including different regional sampling strategies.

Depending on the extension size of the ROI, the studies can be classified in three categories: whole volume-level, regional-level and patch-level. The whole volume strategy considers as input the whole volume of the analysed structure (the whole brain in our case, as have been seen in the previous Chapter 4). At the regional-level, the sampling is based on pre-segmented ROIs which correspond to structures that have been previously used as biomarkers in a given condition (e.g., hippocampus, ventricles) [185, 186]. This level would also include masked regions of the whole brain, such as tissue segmentation (i.e., GM and WM), tissue probability maps [187, 188] or a variation of these like their modulation by the Jacobian of the deformation field [49, 189]. Finally, there would be the patch-level samplings, where the input consists of several patches, whose size can be reduced as desired, without needing to contain a ROI in its entirety. The patch-level samplings, where patches are commonly randomly selected from an abnormal tissue [190], are more widely used in segmentation tasks, where they were proposed to more effectively capture local structural changes.

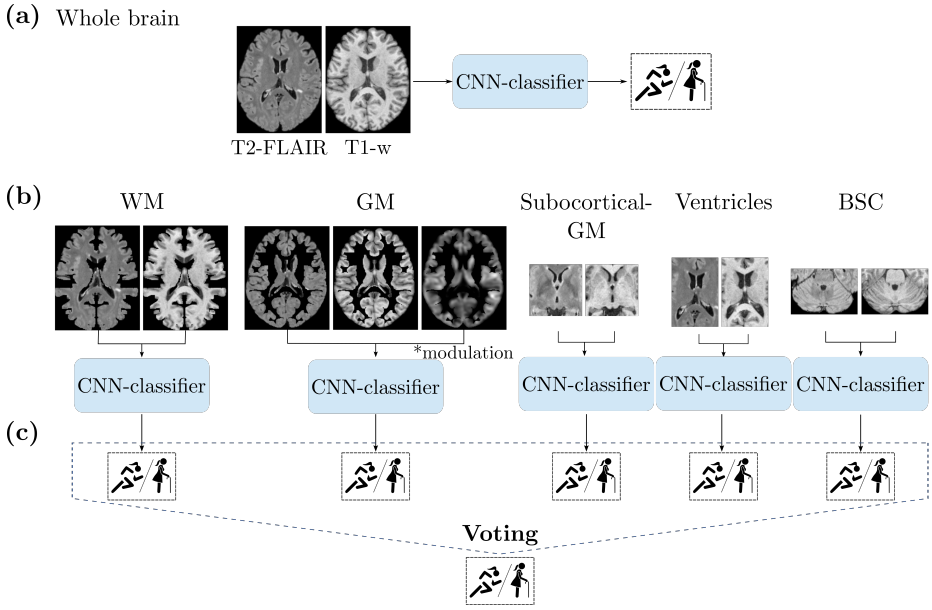

Figure 5.1: Input strategies studied in this work. (a) Whole brain and (b) individual regions evaluated by the classifier model to predict the probability of belonging to EDSS≥3.0 or <3. (c) The single regional model predictions are combined in a voting ensemble. *WM*: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum.

5.2 Dataset

The patients included in this study are the same than the ones used in the previous Chapter 4 for both datasets: 382 patients with MS from the VHUH dataset and 440 patients with MS from the MS PATHS dataset. The same fully automatic image pre-processing pipeline was applied to both datasets, VHUH and MS PATHS, as described in previous Chapters (see Sections 3.2 and 3.3.1).

5.3 Proposed regional models

We proposed an input sampling comparison to evaluate the same CNN implementation to perform a binary stratification task of patients with MS depending on their disability status. We considered a global image-based approach in front of different regional approaches, as summarised in Figure 5.1.

As reference model defining the global approach, we took the whole brain model presented in the previous Chapter 4. As regional approaches, we analysed different ROIs that have proved to be relevant for the prognosis of the disease, which may serve as biomarkers (e.g., localised atrophy measures) or reflect typical locations of WM lesions in MS [22, 183, 191]. Additionally, some of these selected ROIs have already been highlighted during the explainability of the global approach. Thus, we aim to see if each separate region is as able as the whole brain to disentangle this task, or the ensemble of all of them, trained separately, can outperform the single global approach.

The selected regions were distinguished by their volume size, i.e., (i) small regions and (ii) large regions. In the first group, we considered the lateral ventricles, the subcortical GM and the BSC. As large regions, we considered the main brain tissues, i.e., the WM and GM. Depending on the input region that was used for the model, different processing steps were applied to the images to potentiate the input information.

Moreover, as previously done with the global approach in Chapter 4, the obtained performance using the same dataset used for training (VHUH) was compared with the one obtained on the same external validation dataset, from MS PATHS.

5.3.1 Input strategies

In this study, further processing, after pre-processing the scans, was needed to calculate the different input regional strategies: tissue masks and ROI sizes.

All ROIs and masks were obtained automatically and based on the average population of study. First, we performed automatic lesion segmentation [151] to lesion fill the T1-w scans [153]. Afterwards, using the T1-w lesion filled scan, we extracted the whole brain parcellation [160] to obtain the desired regions: subcortical GM structures (thalamus, putamen, caudate and pallidum), lateral ventricles, BSC, as well as, WM and GM tissues. A graphical representation of the different samplings applied may be observed on Figure 5.1(b). For a more detailed explanation of these procedures, as well as, the algorithms used we refer the reader to the Section 3.3.2.

Additionally, for the GM regional input strategy, besides the T1-w and T2-FLAIR scans, a third channel was incorporated as input, the GM modulation. The steps required to extract the final GM modulation are detailed in the

Section 3.3.3 and illustrated in Figure 3.4. The GM modulation was used to preserve the GM volume of the native space, by modulating the GM probability map with the resulting Jacobian determinant from the non-linear registration [168]. Such Jacobian determinant (Figure 3.4(d)) contained the local volume change in each voxel in the common space (MNI in our case).

5.3.2 Training procedure

As in the previous Chapter 4, the same 7-fold patient cross validation strategy was used for training and testing the different models, taking the exact folds of scans, only employing the VHUH dataset for training.

The sampling for each model was decided depending on the type of input region. For small input regions, that is, lateral ventricles, subcortical GM and BSC, we used a square ROI-based patch, delineated from the maximum average map of the region in question from the individual parcellations [185, 186]. The resultant intensity patches had a size of $75 \times 113 \times 41 mm^3$ for the lateral ventricles, $80 \times 72 \times 55 mm^3$ for subcortical GM and $123 \times 85 \times 61 mm^3$ for the BSC, centred at each structure. On the other hand, for the large regions, WM and GM tissues, we took the whole brain patch size ($144 \times 184 \times 152 mm^3$), but only keeping the intensities of the tissue we meant to use as input. This was done by masking the intensity patch with a dilated average mask of the studied tissue. Therefore, only the intensities inside the analysed mask were contemplated [189] (see Figure 5.1(b)).

To mitigate the class imbalance, data augmentation was used. We applied different strategies depending on the input region size. For whole brain patches, the global approach and regional WM and GM approaches, an axial flip was applied to all subjects with EDSS$\geq$3.0 and to a random 75% of the patients with EDSS<3.0, trying to find an equilibrium between balancing the data and not letting the model to learn a non-characteristic feature as the axial flip. For ROI-based models, where the patches were smaller than the whole brain patch, a random voxel displacement in the three dimensions was used to generate additional patches of all subjects, considering the 1:3 proportion of patients with EDSS$\geq$3.0 in the dataset. All proportions were set considering the difference in class-size that we had in the dataset. Each model was trained using T1-w and T2-FLAIR scans (from the in-house dataset) and the corresponding EDSS-based class (EDSS$\geq$ or <3.0). The different training parameters of optimisation, loss function and others were set exactly the same than for the whole brain model

presented in the previous Chapter (see Section 4.3.2).

5.3.3 Inference

For each individual CNN model, the same sampling procedure described during model training was applied. Each predefined patch was used as input through the trained model, providing the output probabilities for belonging to one class or the other. The final classification was determined by the maximum of both probabilities, with a threshold set at 0.5. In addition to the results of each individual model, we computed an ensemble of the regional models with a late fusion strategy [192]. As represented in Figure 5.1(c), the predictions obtained with each trained regional model were aggregated to make a final prediction based on two different voting strategies: (i) maximum and (ii) majority voting. For each patient, the maximum voting strategy was calculated as the prediction of the regional model with the highest probability. While the majority voting approach was calculated as the mode of the different predictions obtained after thresholding each model's probabilities. The calculation of these ensemble models provided individual information of how subjects performed within the different models, presenting either a higher prediction probability (maximum voting) or full agreement across all different models (majority voting).

Without any retraining or fine-tuning of the different models trained with the in-house dataset, inference on the external validation set was also computed as described above. The final prediction per subject was obtained from the majority voting across the 7 different cross-validation models, for each one of the sampling strategies. To evaluate the ensemble of regional models, a maximum voting was computed across the 7-folds of each regional model. Following this, using the winning fold, the specific voting strategy was computed, i.e., majority or maximum voting across regional models.

5.3.4 Evaluation and statistical analysis

Apart from the evaluation metrics already presented in the previous Chapter 4, i.e., accuracy, sensitivity and specificity, we included the ROC (receiver operative characteristic) curve and the AUC measures. The ROC curve shows the performance at all classification thresholds. Thus, AUC is calculated based on all possible pairs of sensitivity and *1-specificity* obtained by changing the thresholds performed on the classification scores.

Paired t-tests were used to compare the performance of each sampling strategy model against the others. T-tests were obtained using the output probabilities from each model to compare the performance on each individual scan-patient in which they were evaluated. To compare the AUCs between models we used DeLong's test [193]. A p-value$<$0.05 was considered statistically significant.

5.4 Results

5.4.1 Evaluation of regional models

Table 5.1 summarises the average performance metrics for the different models tested with the cross-validation evaluation on the VHUH dataset. The global approach, using the whole brain patch as input, achieved a mean balanced accuracy of 79%, for classifying patients with an EDSS$<$ or $\geq$3.0. According to the calculated measures, the best performing individual regional model was the GM model, with a mean accuracy of 81% and the highest sensitivity of 79%. The subcortical GM (a subregion of the GM model) achieved a 78% accuracy, in line with the WM model which had the same accuracy and less variability [72-88]%. Both, subcortical GM and WM models, also achieved similar sensitivity (77% and 75%, respectively) and specificity (79% and 81%). The other two regional models, ventricles and BSC, showed a similar but lower accuracy (76%), but with differentiated sensitivity (76% and 68%) and specificity (76% and 84%) at 0.5 operation point.

Figure 5.2(a) shows the ROC curves and AUC values for all the approaches. As observed with the thresholded values at 0.5 (sensitivity and specificity) the GM regional model had the highest AUC (0.87) and the BSC had the lowest (0.82).

When performing a t-test analysis comparing output probabilities between pairs of models, the BSC and the ventricles regional models had significantly poorer results compared to the other models, which were not significantly different from each other. All the combinations of t-test between models are shown in Table 5.2.

When combining the final performance of the regional models in a voting ensemble approach (see Figure 5.1(c)), we obtained accuracies of 81% and 80% using maximum and majority fusion strategies, respectively, with a lower variability across folds and a higher specificity (88% and 83%, respectively) than

Table 5.1: Model performance across the different folds for the different models assessed using the VHUH dataset.

Model	Accuracy	Sensitivity	Specificity
Whole brain	0.79 ± 0.04	0.77 ± 0.05	0.81 ± 0.09
	[0.70, 0.83]	[0.67, 0.83]	[0.61, 0.90]
WM	0.78 ± 0.05	0.75 ± 0.09	0.81 ± 0.11
	[0.72, 0.88]	[0.63, 0.83]	[0.64, 0.94]
GM	**0.81** ± 0.04	**0.79** ± 0.11	0.83 ± 0.05
	[0.74, 0.87]	[0.63, 0.92]	[0.77, 0.90]
Subcortical GM	0.77 ± 0.04	0.66 ± 0.11	**0.88** ± 0.05
	[0.72, 0.84]	[0.54, 0.83]	[0.81, 0.94]
Ventricles	0.76 ± 0.07	0.76 ± 0.12	0.76 ± 0.08
	[0.66, 0.86]	[0.54, 0.92]	[0.64, 0.84]
BSC	0.76 ± 0.06	0.68 ± 0.18	0.84 ± 0.09
	[0.65, 0.83]	[0.33, 0.88]	[0.74, 0.97]
Majority voting	0.80 ± 0.03	0.75 ± 0.04	0.85 ± 0.07
	[0.75, 0.85]	[0.70, 0.81]	[0.71, 0.94]
Max voting	**0.81** ± 0.03	0.74 ± 0.05	**0.88** ± 0.06
	[0.76, 0.86]	[0.67, 0.81]	[0.80, 0.97]

All terms are specified by mean ± variance [range].
The best results are indicated in bold.
WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum.

the individual models and a similar sensitivity (73% and 77%). In Figure 5.2(a), note that the highest AUC value is obtained with the ensemble model using majority voting (AUC 0.88). In general, the results were slightly better in the voting ensemble approach than in the best individual regional model (GM) in terms of accuracy and specificity. Delong's test showed that most of the paired comparisons with the majority voting ensemble were statistically significant (see Table 5.3).

In the majority voting ensemble, 312 out of the total number of evaluated cases, i.e., 382, were correctly classified: 189 TN (EDSS<3.0) and 123 TP (EDSS≥3.0). For most of patients, all regional models agreed on the same class attribution, being the WM the one most frequently contributing to the vote. In the maximum voting strategy, the models that contributed the most were the GM and BSC,

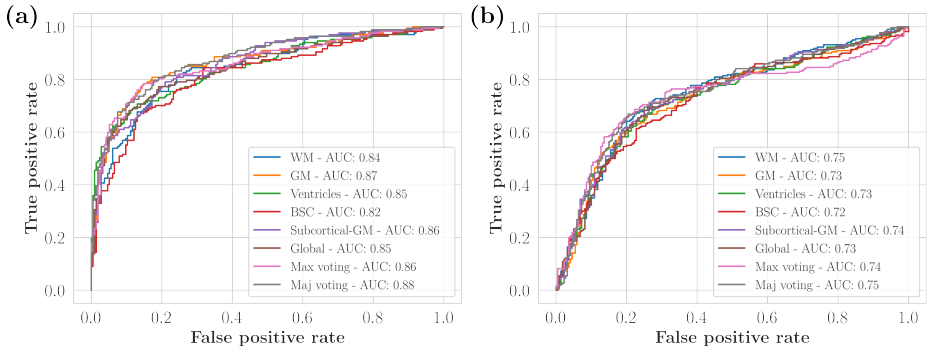

Figure 5.2: ROC curves and AUC values for each regional model, the global approach and the two different ensemble strategies of the regional models for (a) the VHUH dataset and (b) the MS PATHS dataset. In (a) the evaluation of all the cases is represented, unifying the performance of all the folds. In (b) every evaluated subject corresponded to the majority voting result on the 7-folds tested.

reflected in the highest specificity of these models. The relation of all regional models which were 'winning' in the maximum voting approach, i.e., having the highest output probabilities, can be found in Table 5.4.

5.4.2 Validation on an external dataset

Table 5.5 summarises the performance obtained on the independent MS PATHS dataset when directly using the majority voting of the 7-folds models trained with the VHUH dataset. As expected, the overall performance was lower than with the VHUH, obtaining a balanced accuracy of 71%, when using the whole brain approach. The WM regional model obtained the highest individual regional performance with 72% accuracy and the same sensitivity and specificity as the whole brain approach. GM and ventricles regional models presented a similar accuracy (∼70%). The subcortical GM model achieved a similar accuracy (69%) with a much lower sensitivity (59%).

When performing the statistical analysis, there was no evidence of statistically significant differences between these models, as summarised in Table 5.6 with all paired t-test combinations. As in the VHUH cohort, the BSC model performed significantly worse than the other models, with 67% accuracy.

Figure 5.2(b) shows the mean across-folds ROC curve for all the models. As observed with the values obtained at 0.5 operating point, the WM model achieved

Table 5.2: Comparison of models' output probabilities between pairs of regional models in the VHUH dataset: p-values obtained with the paired t-tests.

	Global	WM	GM	Subcortical GM	Ventricles	BSC
Global	-	0.8	0.6	**0.03**	$\mathbf{1.9 \times 10^{-5}}$	$\mathbf{6 \times 10^{-6}}$
WM	-	-	0.15	0.31	**0.04**	**0.001**
GM	-	-	-	0.1	**0.008**	**0.02**
Subcortical GM	-	-	-	-	**0.02**	$\mathbf{6.8 \times 10^{-9}}$
Ventricles	-	-	-	-	-	$\mathbf{1.2 \times 10^{-14}}$
BSC	-	-	-	-	-	-

A p-value<0.05 is considered statistically significant.
Significant p-values are indicated in bold.
WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum.

the highest AUC (0.75) and BSC model had the lowest (0.72). Delong's test showed that none of the AUCs results were statistically significant (see Table 5.7 for all the models' combinations p-values).

When analysing the majority voting strategy, from the total number of evaluated cases, 440, 318 were correctly classified: 184 TN (PDDS<3.0) and 134 TP (PDDS≥3.0). In most cases there was an agreement between all regional models, with the WM model being the most frequent contributor. When analysing the maximum voting strategy, we observed that the GM model provided the highest probabilities followed by the BSC model (see Table 5.8 for the complete contributions in the maximum voting ensemble). As seen in the in-house dataset performance, the majority and maximum voting ensembles showed similar metrics to those of the individual models that best contributed to them (mean balanced accuracy of 73% and 72%, respectively).

5.5 Discussion

In this Chapter, we investigated the ability of different input strategies: global, regional and the combination of these in two different ensembles, to accurately classify patients with MS based on their disability level through deep

Table 5.3: Comparison of models' AUC values between pairs of regional modles in the VHUH dataset: p-values obtained with Delong's tests.

	WM	GM	Subcortical GM	Ventricles	BSC	Maj Vot	Max Vot
Global	0.91	0.29	0.35	0.99	0.19	**0.03**	0.39
WM	-	0.15	0.31	0.91	0.22	$\mathbf{8 \times 10^{-4}}$	0.25
GM	-	-	0.70	0.25	0.06	0.41	0.74
Subcortical GM	-	-	-	0.23	0.05	0.12	0.94
Ventricles	-	-	-	-	0.18	$\mathbf{4.4 \times 10^{-3}}$	0.33
BSC	-	-	-	-	-	$\mathbf{1.4 \times 10^{-3}}$	**0.03**
Maj Vot	-	-	-	-	-	-	0.11
Max Vot	-	-	-	-	-	-	-

A p-value<0.05 is considered statistically significant.
Significant p-values are indicated in bold.
WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum, *Maj Vot*: majority voting, *Max Vot*: maximum voting.

Table 5.4: Percentage of cases in each regional model contributing to the maximum voting ensemble on the VHUH dataset.

	WM	**GM**	**Subcortical GM**	**Ventricles**	**BSC**
N cases (%)	26 (7)	133 (35)	67 (18)	36 (9)	120 (31)
N correct	16 (5)	113 (37)	57 (18)	29 (9)	97 (31)
EDSS<3.0	11 (6)	68 (36)	19 (10)	7 (4)	84 (44)
EDSS≥3.0	5 (4)	45 (37)	38 (31)	22 (18)	13 (10)

WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum, *EDSS*: expanded disability status score.

learning-based CNN models, only using two sequences (T1-w and T2-FLAIR) of a single brain MRI time-point. The study was performed in a large cohort of patients with MS and validated in an external MS cohort.

Our findings showed that, in the VHUH cohort, the best accuracies were

Table 5.5: Performance of the independent set (MS PATHS) on the in-house (VHUH) trained models by means of majority voting across the 7-folds.

Model	Accuracy	Sensitivity	Specificity
Whole brain	0.71 [0.68, 0.72]	**0.68** [0.6, 0.74]	0.75 [0.62, 0.80]
WM	**0.72** [0.69, 0.72]	**0.68** [0.64, 0.70]	0.75 [0.67, 0.78]
GM	0.70 [0.68, 0.71]	0.67 [0.59, 0.71]	0.73 [0.68, 0.79]
Subcortical GM	0.69 [0.68, 0.7]	0.54 [0.53, 0.61]	0.84 [0.75, 0.84]
Ventricles	0.70 [0.69, 0.71]	0.64 [0.56, 0.71]	0.76 [0.62, 0.82]
BSC	0.67 [0.62, 0.68]	0.48 [0.35, 0.62]	**0.85** [0.74, 0.90]
Majority voting	**0.73** [0.72, 0.73]	0.64 [0.63, 0.65]	0.82 [0.79, 0.82]
Max voting	0.72 [0.70, 0.74]	0.61 [0.59, 0.64]	0.83 [0.80, 0.84]

The ranges are calculated using each individual fold-model.
The best results are indicated in bold.
WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum.

achieved with the regional GM approach followed by the whole brain approach, whereas the best performing models in the external dataset (MS PATHS) were from the regional WM approach, followed by the whole brain approach. Thus, the global whole brain approach offered the best trade-off between internal performance and external validation, although some regional models such as GM and WM models showed similar overall performances.

Among the different individual strategies presented, the regional GM model achieved the best overall performance results (with an average accuracy of 81%). This may be explained by the fact that in this approach a third input channel, the GM modulation, was incorporated. The GM modulation represents the deformation suffered by the image when registering the scans to a common space. Thus, it provides information about the native space, accounting for a possible effect of GM atrophy, which is known to be important for the development of future disability in MS [181, 194].

The next models in ranking were the whole brain approach and regional WM model and subcortical GM models. The whole brain and the WM models are those that may have a direct relationship with the WM lesion load. Indeed, a post-hoc analysis showed that there was an association between WM lesion load and model output for the global and WM-regional models. We found that the correctly classified cases ($\geq$ or $<$3.0) had a mean lesion volume of approximately

Table 5.6: Comparison of models' output probabilities between pairs of regional models using the MS PATHS dataset: p-values obtained with the paired t-tests.

	Global	WM	GM	Subcortical GM	Ventricles	BSC
Global	-	**0.007**	0.7	0.9	0.08	$\mathbf{1.4 \times 10^{-6}}$
WM	-	-	**0.02**	**0.01**	0.3	$\mathbf{9.3 \times 10^{-14}}$
GM	-	-	-	0.79	0.12	$\mathbf{8.9 \times 10^{-7}}$
Subcortical GM	-	-	-	-	**0.03**	$\mathbf{1.9 \times 10^{-6}}$
Ventricles	-	-	-	-	-	$\mathbf{7.1 \times 10^{-9}}$
BSC	-	-	-	-	-	-

A p-value<0.05 is considered statistically significant.
Significant p-values are indicated in bold.
WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum.

60 ± 52 mL and 9 ± 10 mL, for TP and TN, respectively. However, the presence of greater lesion volumes may not always be associated with a worse prognosis [195], as reflected in the similar mean lesion load for the FP and FN cases (16 ± 22 mL and 18 ± 21 mL).

The same behaviour for the correctly classified cases was observed when evaluating the performance of the whole brain and WM regional models on the MS PATHS subset: 27.1 ± 38.7 mL for TP and 4.6 ± 7.5 mL for TN, approximately half of the ones in the in-house dataset. On the contrary, the groups with incorrect classifications exhibited lower lesion loads, which were closer to their true values. Specifically, the FN cases had an average lesion load of 21.4 ± 15.9 mL, and the FP cases had an average lesion load of 5.9 ± 7.9 mL. This suggests that the classification may not heavily rely on this particular biomarker, and it implies that factors other than those present in the training dataset, potentially powerful but undisclosed features, have influenced the decision. On the other hand, the subcortical GM model performance was similar to that obtained with the GM tissue model. These results are in line with the strong correlation shown between GM subcortical volumes and disease severity [196].

In this study, the ventricles model performed acceptably well in both datasets. The presence (total or partial) of WM lesions together with the presence or absence of atrophy that can be measured by the ventricles size may have helped

Table 5.7: Comparison of models' AUC values between pairs of regional modles in the MS PATHS dataset: p-values obtained with Delong's tests.

	WM	GM	Subcortical GM	Ventricles	BSC	Maj Vot	Max Vot
Global	0.26	0.73	0.57	0.83	0.48	0.19	0.96
WM	-	0.17	0.66	0.27	0.10	0.81	0.38
GM	-	-	0.31	0.87	0.69	0.03	0.71
Subcortical GM	-	-	-	0.36	0.3	0.47	0.67
Ventricles	-	-	-	-	0.64	0.06	0.84
BSC	-	-	-	-	-	0.12	0.3
Maj Vot	-	-	-	-	-	-	0.28
Max Vot	-	-	-	-	-	-	-

A p-value$<$0.05 is considered statistically significant.
WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum, *Maj Vot*: majority voting, *Max Vot*: maximum voting.

with the model accuracy [1, 180]. However, as for the BSC regional model, not having been given the whole brain image as input may have constrained the model, leading to statistically significant poorer performance than the models with a larger regional input (see Table 5.2). The BSC model did not have the same ability to correctly classify both disability status, presenting the poorest accuracy in both datasets. However, the BSC model was a relevant addition to the maximum voting ensemble model due to its high specificity identifying patients with EDSS$<$3.0. On the other hand, the GM model was the model that contributed the most for the correct classification of patients with EDSS$\geq$3.0, due to its very high sensitivity when evaluated individually.

In general, the performance of the different models on the MS PATHS subset resulted in slightly lower performance than on the in-house cohort. However, considering that there was no additional training or fine-tuning to evaluate the unseen data, the accuracy of all models was satisfactory, supporting the generalisation of our models. However, using a different score (PDDS instead of the EDSS used for training) to classify this external validation dataset may be seen as a limitation of this evaluation, despite the strong correlation between both metrics [140].

Table 5.8: Percentage of cases in each regional model contributing to the maximum voting ensemble on the MS PATHS dataset.

	WM	**GM**	**Subcortical GM**	**Ventricles**	**BSC**
N cases (%)	20 (5)	168 (38)	15 (3)	71 (16)	166 (38)
N correct	16 (5)	122 (38)	11 (3)	55 (17)	114 (36)
PDDS<3.0	4 (2)	56 (30)	1 (1)	16 (9)	107 (58)
PDDS≥3.0	12 (9)	66 (49)	10 (8)	39 (29)	7 (5)

WM: white matter, *GM*: grey matter, *BSC*: brainstem and cerebellum, *PDDS*: patient determined disability score.

Our results highlight the importance of considering whole brain input sampling strategies to promote generalisability of CNN-based stratification models. Although this might be intuitive, this study quantitatively assessed the effect of the type of input that a CNN-based model must have in this MS stratification problem. Building accurate CNN-based models is key to predicting individual patients' disease course in order to achieve a personalised approach [1]. The methodology presented in this study, along with retraining or fine tuning, may have potential for diagnostic or progression prediction tasks.

This study is not without limitations. Apart from the relatively small sample size used for training the models from the in-house dataset, all five MRI scanners present in the study were from the same vendor. This can be seen as a limitation in terms of model generalisation. However, the training cohort did include scans acquired at different strength fields (1.5 T and 3 T), with some variation in protocols between scanners. Approximately half of the patient data was acquired with the same acquisition protocol that was used in the external validation cohort (MS PATHS), where images were also acquired with different MRI scanners from the same vendor, three of which were not included in the in-house dataset. The external validation was also restricted by not using the same clinical score as used in the training set (PDDS instead of EDSS). Of note, although EDSS and PDDS are both non-linear scales, the EDSS is obtained by a neurologist, after performing an anamnesis and a neurological examination, and the PDDS is instead reported by the patient, implying a strong subjective nature. Therefore, they reflect essentially different points of view of the disease. However, they are

highly correlated [140], which is reassuring and suggests that they may be used for similar predictive and monitoring purposes. Indeed, the PDDS is frequently used in those clinical settings where the EDSS is not available. In our study, we considered an EDSS of 3.0 to be equivalent to a PDDS of 3.0. However, other equivalences were indeed possible and should be explored in further studies.

Transfer learning and fine tuning are common techniques used for domain adaptation in medical image analysis [109] when there are different target domains departing from a common source. Another complementary analysis would be to analyse the unexplored patch-based sampling strategy. A way to do that would be by exploring different seeds from the already used ROIs to centre the patches [190].

From the clinical point of view, setting a threshold at a certain EDSS may not involve all the factors that determine disability in patients with MS at a cross-sectional point. Despite this, the EDSS is the most frequently used clinical score to quantify disability in MS clinical practice and clinical trials and reaching an EDSS$\geq$3.0 has been shown to be a relevant outcome when studying the disease progression course [18]. However, in this study, the EDSS$\geq$3.0 was not always confirmed in a follow-up visit, which means we may not have analysed a clinically stable population. Also, we did not account for any disease-modifying treatment. Further studies taking these aspects into account are therefore necessary.

In conclusion, when it comes to patient classification based on their disability level, CNN-based models are able to extract features from different input strategies and lead to a correct classification. The global (whole brain) and large ROI-input models (WM and GM) resulted in the highest classification accuracies. While their similar behavior suggests that the CNN is able to adapt to its inputs, this also indicates that focusing on specific regions, even if a priori important for MS, does not necessarily translate into better performance. Indeed, using global input approaches may result in a better generalisation of such CNN models as it offers the best trade-off between internal performance and external validation.

Chapter 6

MS patients prognosis prediction

"You must be optimistic."
A Professor.

6.1 Introduction

Predicting the prognosis of MS involves assessing the course of the disease and its impact on an individual's health over time. The underlying cause of MS is still unknown, and its disease course varies significantly among individuals [1]. While some experience rapid deterioration in physical and cognitive abilities, others remain relatively stable for years after symptom onset. Therefore, early and accurate predictions of the short and long-term disease course are crucial for patient management.

As introduced in Chapter 1 (see Section 1.1.2), patients with MS acquire disability through two main mechanisms: PIRA and RAW [13]. However, some studies consistently indicate that PIRA is the primary driver of disability accumulation in this disease [15, 16, 17]. Despite this, the pathological basis of PIRA remains unclear.

In recent years, there has been a growing interest in developing deep learning models to predict disease progression, particularly in early disease stages, before irreversible disability reaches high levels. This early stage may represent a unique therapeutic window [1, 18]. For example, various studies have analysed different MS populations to predict disability status at 1- or 2-years follow-ups using classification and regression image-based deep learning models [31, 32, 128]. However, long-term predictive deep learning models for MS progression are still largely unexplored.

In this Chapter, we propose a deep learning-based survival model that uses T1-w and T2-FLAIR brain MRI scans acquired after a patient's first demyelinating attack to predict the survival function of patients with MS until they experience their first PIRA event. To evaluate the performance of the survival function, in addition to providing the predicted cohort survival curve, we assess its goodness-of-fit through risk stratification based on the model's output. Furthermore, as previously performed in Chapter 4, we present two distinct interpretability strategies: one involving the verification of our results using classical statistical methods, and the other using a deep learning interpretability algorithm to elucidate the decision-making process of our deep learning-based model.

6.1.1 Survival analysis

A survival model (or time-to-event model) is a statistical approach that assesses the impact of various covariates on the time elapsed before an event takes place. It characterises the future behavior of a subject by generating a risk score or a time-to-event distribution.

The data used in a survival model consists of three main components: baseline data (x), an event indicator (E), and the observed time-to-event (T). The event indicator (E) is a binary value that indicates whether the event has been observed (E=1) or if there is censoring (E=0). Censoring occurs when we lack enough information to determine if the event will occur within the analysed period (left censoring), or when the event has not happened during the studied period (right censoring). For censoring cases, the last available follow-up time from baseline is considered as the value for T.

The primary objective of a survival model is to estimate the probability of an event occurring at a specific time. This probability is often estimated using

a survival function $S(t)$, which denotes the likelihood of a subject surviving (i.e., not experiencing the event) beyond time t (as shown in Equation 6.1). An alternative representation of the survival function is through the hazard function $h(t)$, expressing the conditional probability of the event occurring at time t, given that it has not occurred prior to that time.

$$S(t) = P(T > t) = \exp\left(-\int_0^t h(s), ds\right) \tag{6.1}$$

Here, $H(t) = \int_0^t h(s), ds$ represents the cumulative hazard function. The risk score, dependent on x, can be computed through an approximation of the cumulative hazard function:

$$r(x) = \exp\left(-\sum_{j=1}^{J} H(t_j, x)\right) \tag{6.2}$$

In this equation, J encompasses the time intervals into which t is divided.

Traditional survival models

Traditional statistical survival models are based on specific assumptions about the data and the relationship between the given covariates. The traditional types of survival models are categorised as parametric and non-parametric. A parametric model assumes a specific distribution for the survival times (e.g. exponential) and estimates the parameters of that distribution to model the survival function. On the other hand, the non-parametric models do not assume the distribution of the data and directly estimate the survival function from the data provided. The most known non-parametric estimator is the Kaplan-Meier [197].

The classical statistical approach for survival analysis is the Cox Proportional Hazards model (CPH) [198]. CPH is a semi-parametric linear model which assumes that the effect of the predictors is a fixed time-independent multiplicative factor on the value of the (baseline) hazard function.

6.1.2 State of the art in MS

In a previous Chapter (see Section 2.5), it was highlighted that the use of deep learning strategies for predicting prognosis in MS is still severely limited. Current studies have mostly focused on short-term future cross-sectional

predictions, addressing questions like *"will this patient progress in a 1- or 2-year follow-up?"* [32, 128]. Another approach involves regression tasks that aim to estimate the EDSS score at specific follow-up time-points [31].

Expanding the methodology to encompass machine learning techniques and more conventional statistical methods reveals studies applying SVMs on pre-extracted baseline demographic, clinical, and brain MRI features [30, 126, 127]. These studies aimed to assess the most influential features that contributed to enhanced predictions.

Regarding our specific focus, the prediction of PIRA, only two studies have so far concentrated on this using clinical, categorical radiology features, and demographic characteristics at the onset of the first attack or early disease stages [16, 17]. Among the features present at the initial attack, advancing age has been associated with a bigger risk of PIRA [16, 17]. Conversely, other crucial characteristics at the onset of the first attack, such as the quantity of lesions in brain or cord MRI, which have been linked to a generally unfavorable MS prognosis, have not been correlated with an elevated PIRA risk [17]. However, once the disease is established and patients are distanced from the initial demyelinating event, brain and cord MRI have proven highly valuable in uncovering some of the presumed neurodegenerative mechanisms underlying PIRA, including brain and cord atrophy [14, 43, 199, 200].

6.1.3 State of the art in medical imaging

In an end-to-end deep learning-based survival prediction model, non-linear features are extracted from the training data, analogous to covariates in traditional survival models. These features are then combined linearly to estimate the log-risk function. Similar to conventional survival models, deep learning models are optimised using a cost function that takes into account censored data, such as the Cox partial likelihood.

The first study using deep learning for survival analysis was marked by a simple model comprising just one hidden layer, establishing the foundation for a non-linear proportional hazard model [201]. However, it failed to surpass the performance of the linear CPH. It was after decades that survival studies using deep learning techniques emerged with the use of more complex neural network architectures applied to a variety of clinical applications with different types of data [202, 203, 204, 205, 206, 207].

A key advantage of deep learning-based models lies in their capability to manage high-dimensional data, including images or genomics. Some image-based approaches incorporate clinical data (e.g., demographics, genomics, radiomics) with the corresponding diagnostic image modality within their architectures [204, 205, 206].

Within the field of image-based deep learning time-to-event or survival models, two distinct approaches can be identified. In one approach, the task is framed as a regression problem, aiming to predict the time an outcome is reached from baseline or a common reference time-point for all subjects. This can be seen in recent examples such as *Pix2Surv* by Uemura *et al.* [208], which used a conditional GAN to predict time-to-prognosis of COVID-19 from baseline chest CT scans. Additionally, *DAAL* by Xu *et al.* [209] employed novel vision transformer networks to forecast hazard risk of brain cancer survival using MRI scans. On the other hand, the task can also be cast as a discrete-time survival model, where the objective is to learn the survival function rather than the precise time of an outcome's occurrence. Notably, the work of Vale-Silva and Rohr with *Multisurv* [205] exemplifies this perspective. They implemented a discrete-time survival model for various cancer types using histopathology images, combined with clinical and genetic data.

6.2 Dataset

For training and testing this study we used patients with MS from the VHUH dataset (see Section 3.2.1). The main inclusion criteria was the availability of a brain MRI scan at the time of the first demyelinating attack and a long enough follow-up clinical evaluation of disease progression (at least 3 EDSS assessments, i.e., approximately 1 year after symptom onset). Additionally, for the censored cases, that is, the ones that did not experience a PIRA, they were required to have a minimum follow-up equal to the median time to the first PIRA event in our population (i.e., 4 years - see Table 6.1). The final selected subjects' acquisition dates run from 2009 to 2018, with a maximum follow-up time of 12 years. In this period of time and with the applied criteria, the resulting dataset consisted of 259 patients of which 58 had at least one PIRA event. Censored and non-censored patients were all considered for the progression analysis of time to reach a first PIRA event. The main demographic, clinical and brain MRI characteristics of these patients are summarised in Table 6.1.

Table 6.1: Demographic, clinical history and brain MRI characteristics of patients at baseline included in this study.

	Full cohort	**PIRA**	**non-PIRA**	***p***
N, (%)	259	58 (22)	201 (78)	
Outcome time, years, median[range]	6.9 [1.7, 12.7]	4.2 [1.7, 12.3]	7.7 [4.1, 12.7]	<0.001
Female, n(%)	170 (66)	40 (69)	130 (65)	0.65
Age at CIS, years, mean(SD)	34.3 (7.9)	36.2 (8.2)	33.8 (7.8)	0.06
Time of last FU, years, mean (SD)	7.9 (2.5)	7.8 (2.7)	7.9 (2.5)	0.95
Annualised relapse rate, mean (SD)	0.25 (0.18)	0.23 (0.16)	0.26 (0.19)	0.22
EDSS at CIS, median[range]	1.5 [0.0, 4.5]	1.5 [0.0, 4.5]	1.5 [0.0, 3.5]	0.99
EDSS at last FU, median[range]	1.0 [0.0, 6.0]	2.0 [0.0, 6.0]	1.0 [0.0, 5.0]	<0.001
CIS topography, n(%)				0.66
Brainstem	49 (19)	13 (22.4)	36 (17.9)	
Optic nerve	100 (39)	22 (37.9)	78 (38.8)	
Spinal Cord	81 (31)	17 (29.3)	64 (31.8)	
Polyregional	12 (5)	1 (1.7)	11 (5.5)	
Other	17 (6)	5 (8.6)	12 (6.0)	
Presence of OB, n(%)				0.27
Positive	142 (55)	36 (62)	106 (52.7)	
Negative	117 (45)	22 (38)	95 (47.3)	
Spinal cord lesions, n(%)				0.72
0	126 (49)	27 (46.6)	99 (49.3)	
1	54 (21)	14 (24.1)	40 (19.9)	
2-3	27 (10)	6 (10.3)	21 (10.4)	
>3	34 (13)	9 (15.5)	25 (12.4)	
unknown	18 (7)	2 (3.4)	16 (8.0)	
Brain lesions, n(%)				0.32
0	60 (23)	10 (17)	50 (25)	

1-3	28 (11)	4 (7)	24 (12)	
4-8	43 (16)	10 (17)	33 (16)	
≥9	128 (49)	34 (59)	94 (47)	
Treated during FU, n (%)	161 (62)	41 (71)	120 (60)	0.17
Proportion time treated, mean(SD)	0.53 (0.43)	0.58 (0.39)	0.51 (0.44)	0.24
Scanner model, n(%)				0.89
Tim Trio	236 (91.1)	53 (91.4)	183 (91)	
Symphony	12 (4.6)	3 (5.2)	9 (4.5)	
Avanto	9 (3.5)	2 (3.4)	7 (3.5)	
Symphony Tim	2 (0.8)	-	2 (1)	
Lesion load, mL, mean(SD)	3.4 (4.3)	4.8 (6.3)	3.1 (3.4)	0.05
BPF, mean(SD)	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)	0.15
WMF, mean(SD)	0.46 (0.02)	0.46 (0.02)	0.46 (0.02)	0.14
GMF, mean(SD)	0.54 (0.02)	0.54 (0.02)	0.54 (0.02)	0.14

PIRA: progression independent of relapse activity; *CIS*: clinically isolated syndrome, *OBs*: oligoclonal bands, *FU*: follow-up, *BPF*: brain parenchymal fraction, *GMF*: grey matter fraction, *WMF*: white matter fraction.

All the scans were pre-processed with the methodology described in Chapter 3 (see Section 3.3.1). Moreover, we obtained the different descriptive volumetric information from brain MRIs as explained in Section 3.3.2.

6.3 Proposed model

We propose the use of an end-to-end deep learning model to predict the survival curve of patients with MS experiencing a first PIRA event only using brain MRI scans acquired after the first demyelinating attack.

Our proposed image-based discrete-time survival model is illustrated in Figure 6.1. It is based on a discrete-time survival model implementation [203], but using as feature extractor an EfficientNet-b0 [58] followed by a fully connected layer that performs as predictor to output a vector of surviving probabilities, i.e., not reaching a PIRA. T1-w and T2-FLAIR scans, previously pre-processed, containing centred axial brain information were used as unique input to train

and test the network. After inference, different analyses were conducted in terms of cohort survival prediction using the probability or hazard rate at the first time-interval by: (i) assessing model's goodness-of-fit through a risk stratification strategy, (ii) testing model's performance through assessing the predicted risk on a classical statistical survival model, i.e., a CPH model [198], and (iii) interpreting the most relevant regions for the network to make its predictions, using SHAP [74].

6.3.1 Network architecture

The proposed 2D-CNN architecture was an EfficientNet-b0 [58], illustrated in Figure 6.2. EfficientNet [58] is a family of CNN classifiers or bottleneck encoders that differentiates from other popular architectures such as ResNet [56] or VGG [55], due to its unique compound scaling method that optimises the network's depth, width, and resolution simultaneously, allowing it to achieve better accuracy with fewer parameters. From all the different network sizes, we chose the smaller one, *b0*, since it is the one with fewer parameters, that balances with our dataset size.

The base architecture is the same for any size of network, they differ in the number of times a layer-block is repeated (products for *b0* are shown in Figure 6.2). This architecture is composed of a first convolutional layer and successive mobile inverted bottlenecks (MBConv) [210] with squeeze-and-excitation (SE) optimisation [211], followed by a last 1×1 convolutional layer, a GAP and a fully connected layer activated with Sigmoid to obtain the different output probabilities.

Conforming the MBConv block there are depth-wise convolutional layers, which expand features instead of reducing them. After that, SE blocks are used to improve the quality of representations by explicitly modeling the inter-dependencies between the channels. Additionally, in between this two blocks a Swish activation is used, defined as $f(x) = x \cdot Sigmoid(x)$, that tends to work better than ReLU on deeper models.

6.3.2 Training and inference procedures

A 5-fold cross validation strategy was used to train and test the proposed model. We sampled the folds to have a similar event-time distribution in each one. In each iteration, 3 folds were used for training (156 scans), one-fold (52 scans) for

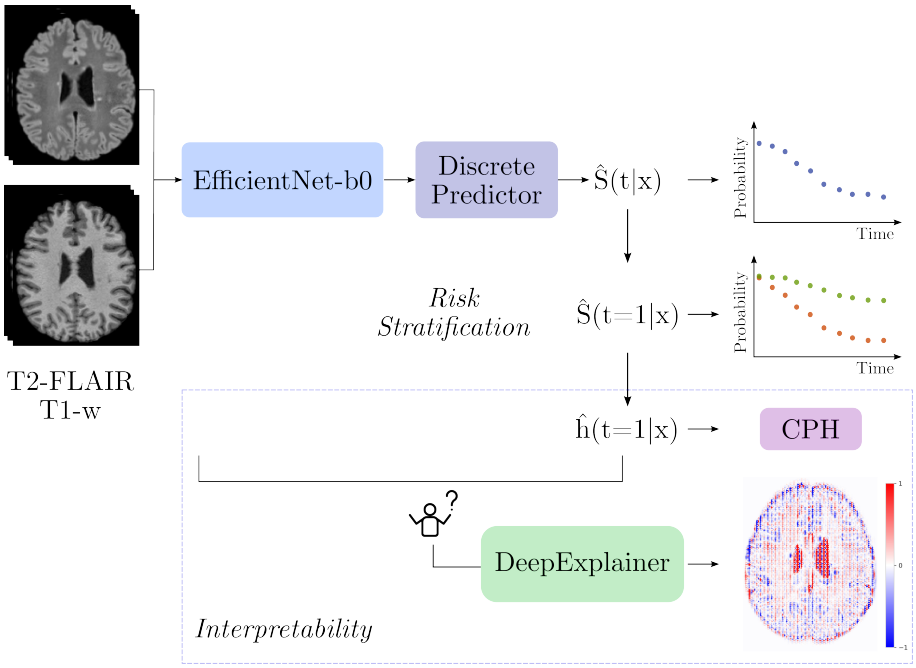

Figure 6.1: Overview of the proposed pipeline for the prediction of survival probabilities and their evaluation. The T2-FLAIR and T1-w 2D slices are used as input to the network to predict time-discrete probabilities of not reaching PIRA. After inference, the predicted probability at the first time is used to stratify the patients in high- vs low risk of reaching a PIRA. Also, the hazard at first time-interval is used for model's verification on a CPH model and interpretability with SHAP values DeepExplainer algorithm. $\hat{S}$: survival function, $\hat{h}$: hazard rate, *CPH*: Cox proportional hazard.

validation and the remaining one for testing. From each baseline scan (T1-w and T2-FLAIR scans), we used the central 40 axial slices as input to the 2D end-to-end deep learning model. This decision was taken after comparing the models' performance using different samplings, which included the total number of slices, 100, 60 and 40 slices.

We trained the model using the pre-trained weights on the ImageNet natural image dataset [212]. At the sixth convolutional block, we "unfroze" the weights and fine-tuned the last convolutional block parameters on our specific dataset, followed by the training of the predictor. The model was trained using the negative log likelihood loss function. For an easier implementation, it was

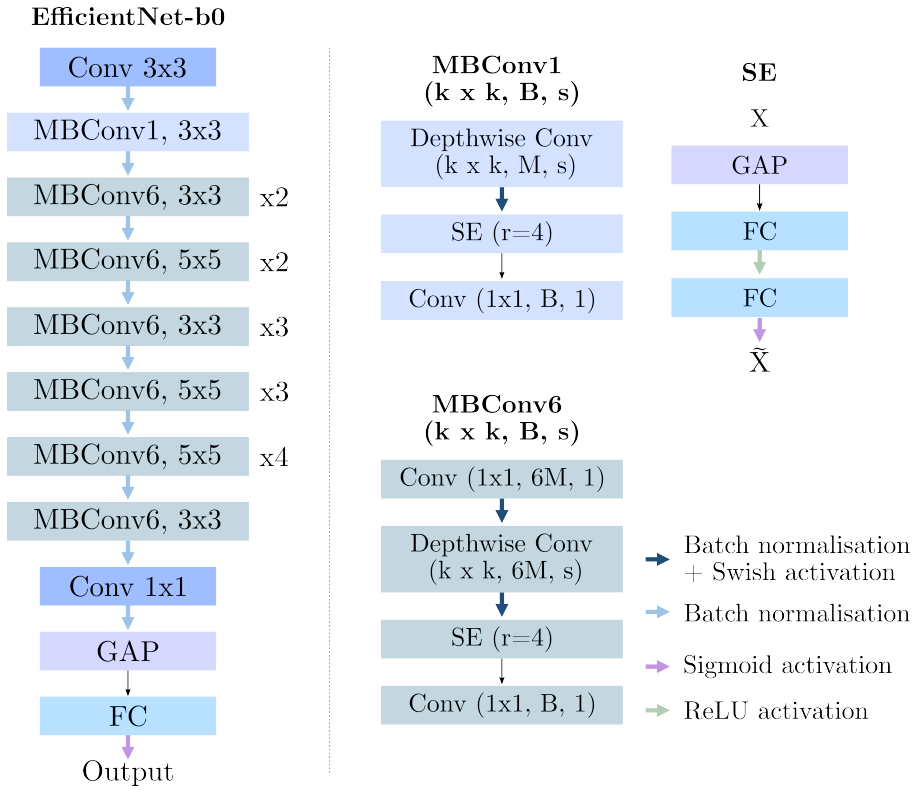

Figure 6.2: EfficientNet-b0 architecture and the main layers forming the convolutional and residual blocks. *SE*: squeeze-and-excitation, *MBConv*: mobile inverted bottleneck, *Conv*: convolutional layer, *GAP*: global adaptive pooling, *FC*: fully connected, k: kernel size, M: input feature maps, B: output feature maps, s: stride, r: reduction ration of SE.

evaluated by subject instead of time-interval [205]:

$$\sum_{i=1}^{d_j} log(h_j^{(i)}) + \sum_{i=P_j+1}^{r_j} log(1 - h_j^{(i)}) \tag{6.3}$$

where $h_j^{(i)}$ is the hazard probability for the i subject during time interval j, and P_j is the number of patients that had a first PIRA event during this interval j. This loss is evaluated at each interval, taking into account all input-patient that had the event (or were censored) later than the beginning of that interval [203].

The sum of all losses at each time-interval is the total loss.

We trained each fold-model for a maximum of 150 epochs, with a fixed batch size of 64 and an early stopping strategy based on the validation loss behaviour to prevent overfitting. The model was optimised with Adam [165] with an initial learning rate of 10^{-5} and a cosine annealing schedule [213].

For inference, we used the same sampling as for training, passing through the model each single 2D slice. After that, the median of the 40 slices was taken as final patient-output. Thus, for every patient we obtained a predicted-survival function expressing the probability at each time-interval to reach a first PIRA event, as well as the hazard rate at the first time-interval.

6.3.3 Model evaluation

Performance

To assess the model performance, we used the time-discrete output, the one capturing a time-varying prognostic performance. For that, we used two different time-dependent metrics. To evaluate the survival model accuracy, we used the extended version of the widely used Harrell's concordance index or c-index [214], i.e., the time-dependent concordance index (c^{td}) [215]. The c^{td} considers time-to-event information by comparing predicted and observed event-times at multiple time-points over the course of follow-up. At each time-point, it determines the proportion of pairs of individuals where the predicted event-time is greater than the observed event-time and then computes the average of these proportions over all time-points [215]. A valid score ranges from 0.5 to 1.0, with 1.0 indicating a perfectly discriminative model and 0.5 indicating random probabilities.

The second accuracy metric used was the integrated Brier score (IBS) [216], which quantifies the mean square difference between the predicted survival probabilities and the observed event-time. The smaller the IBS, the better the predictions are. For a model with an incidence of outcome of 50%, the maximum IBS is 0.25 [217], which suggests a non-informative model, whereas an IBS=0 indicates a perfect model. Since our outcome incidence is 22%, we will consider an acceptable IBS value if it is between 0 and 0.17, calculated from [217]. Additionally, the numerical value of the IBS would be time-interval comparable by graphical representation of the predicted survival probabilities and its relative true observations by means of the well-known Kaplan-Meier curve [197].

Risk stratification

After evaluating the model performance, we assessed the model's goodness-of-fit through a risk stratification strategy. We divided our patients based on the median probability (across all patients) of presenting a first PIRA event during the first-time interval (i.e., from year 1 to year 2 of follow-up) into high-risk (i.e., probability above the median) and low-risk (i.e., probability below the median) groups of reaching a first PIRA event [205]. The purpose of this step was the assessment of the model's goodness-of-fit. That is, we assessed the patients of such risk stratification to correctly identify those patients who eventually had a PIRA event during the follow-up. Association was also obtained for those with early PIRA, i.e., for those patients whose first PIRA event took place in the first 5 years after the first attack.

Additionally, this risk stratification allowed us to describe the clinical and MRI profiles of those patients from the high- and low-risk PIRA groups. For that, we focused on two main aspects: (i) demographic and clinical history information at baseline, and (ii) relevant MRI volumetric measurements obtained at baseline.

Descriptive statistical analysis

We performed a descriptive statistical analysis of the main demographic and clinical information at baseline between patients who reached the event (PIRA) and patients who did not (censored). The same analysis was also performed for patients that were stratified as high vs low risk of PIRA based on deep learning-based outputs. We used a chi-square contingency test and t-test, as appropriate, considering as significant a p-value<0.05. Additionally, to test the separability of the risk groups, we used a log-rank test.

6.4 Proposed interpretability

To support the quantitative performance results we proposed two different interpretability implementations. On one hand, we proposed to use a traditional statistical model, the CPH [198], to be fit with our deep learning model's output (i.e., the predicted hazard rate of PIRA during the first-time interval) and a known predictive covariate, i.e., the age at CIS, which has been proposed as the most important predictor of PIRA by several authors [14, 16, 17]. On the other hand, our deep learning model is submitted to its own interpretability,

using a back-propagation method to explain its decisions, specifically using the DeepExplainer of SHAP [74]. Both of them are calculated from the first time-interval hazard rate obtained as model output.

6.4.1 Classical statistical model

CPH models are widely used and accepted for survival analysis in the medical field, including the prediction of PIRA [17]. At this stage we aimed to assess whether our deep learning-based outputs (i.e., hazard rate of PIRA during the first time interval and PIRA-risk group) could improve an accepted survival model of PIRA including age as the only predictor [16, 17].

We first built a CPH model with age at the first demyelinating attack as the only predictor, based on previous publications [16, 17]. Afterwards, the hazard rate of PIRA during the first-time interval (min-max normalised) was added to the model as covariate. Hazard ratios (HRs) and their 95% CI were estimated for all the predictors. Whenever the HR of the newly added variable was significant at $p<0.05$ we assumed that the deep learning-based estimates improved the age-based CPH model of time to PIRA, providing independent information. CPH models' accuracies were assessed through the c-index [214].

Two additional CPH models were built to test the ability of deep learning-based estimates (hazard rate) to improve classical survival models of time to PIRA: (i) a model with age, lesion load, GM volume, WM volume, and total intracranial volume as covariates; (ii) a model with age, lesion load, GM volume, WM volume, total intracranial volume, and deep learning-based hazard rate as covariates. We obtained the c-index for each model. We assumed that any improvement in model performance was significant whenever the c-index obtained after adding the deep learning-based hazard rate was higher than that of the model without it and the HR for the deep learning-based hazard rate was significant ($p<0.05$).

6.4.2 Relevance maps: Deep SHAP

In our study, the interpretability of our model predictions was crucial. To achieve this, we used the SHAP framework, introduced in Chapter 2.4, which provides both local and global explanations for the model's behavior [74]. As a back-propagation interpretability method, SHAP operates within a post-hoc framework, ensuring model-specific insights. The essence of SHAP lies in

assigning relevance or importance values to individual features, such as pixels in our context, with respect to the model's output prediction.

In a classification scenario, the SHAP framework attributes positive or negative contributions to different features, indicating their impact on the predictive class. However, since we are dealing with a more likely regression task —predicting the first time-interval hazard rate— we align the positive values with an increase in the network's output and the negative values with a reduction. In the context of our study, positive values correspond to an increased survival, implying a lower probability of experiencing a first PIRA, whereas negative values suggest a higher probability of encountering a PIRA event.

We used the DeepExplainer method within the SHAP framework to quantify the importance of each pixel's contribution to our model's prediction of hazard risk [74, 218]. By summing up the SHAP values across all pixels, we effectively express the predicted hazard risk. The visualisation of these SHAP values presents a balanced depiction of areas that exert positive or negative influences on the network's output, thereby influencing survival probabilities. More precisely, the resulting SHAP value maps are presented on a blue-to-red scale. Negative relevance, indicated by the colour blue, suggests a higher likelihood of encountering a PIRA, while positive relevance, shown in red, suggests a decreased probability of encountering a PIRA, thereby indicating a greater chance of survival.

Once obtaining the individual SHAP maps for each patient, we computed an average map population to reveal to which regions the model was paying more attention while performing the prediction. This was achieved by multiplying the SHAP maps by a brain parcellation map. For this purpose, the SHAP maps were normalised and a binary threshold was set at 95% percentile of positive and negative relevance. Both were assessed separately to identify the most relevant brain anatomical areas associated with the risk of experiencing a first PIRA event.

6.5 Results

6.5.1 Survival performance

The discrete-time survival model achieved a mean c^{td} across folds of 0.72 (range=[0.68, 0.78]) and a mean IBS of 0.1 (SD=0.04), reflecting the model's accuracy. Such accuracy was particularly high until the 8-year follow-up period,

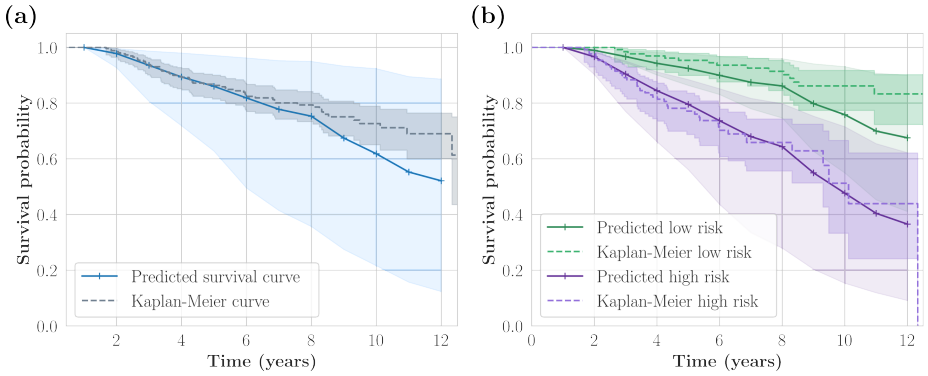

Figure 6.3: Model predictions. (a) Average of the predicted survival curves (continuous line) for all the patients included in the analysis compared with the Kaplan-Meier estimator (dashed line) output. (b) Post-analysis risk stratification groups. The predicted survival curves are overlayed by the Kaplan-Meier estimated curves of each low- and high-risk estimations. In both graphics, the shaded areas represent the 95% CI of each represented output.

when the deep learning-based survival estimates started to differ from the empirical survival function (Figure 6.3(a)). Figure 6.3(a) shows the mean (across all patients) discrete survival function predicted by our model compared with the actual Kaplan-Meier survival curve only described by the target time-events distribution. The visual assessment of the curves, combined with the IBS score, demonstrates that the IBS falls within the informative range of values [0-0.17].

6.5.2 Risk stratification

Once the accuracy of the model was evaluated, patients were ranked into low (N=130, 65% female, mean age (SD)=34(8) years) and high (N=129, 66% female, mean age (SD)=35(8) years) risk of PIRA, depending on whether they were below or above the median first-year estimated probabilities, respectively. This stratification procedure showed that 74% of patients who developed PIRA were correctly identified in the high-risk group. This proportion raised up to 83% when it came to identifying patients with early PIRA (within the first five years of the disease).

Figure 6.3(b) shows the Kaplan-Meier curve and the predicted survival curve for each risk group. We visually assessed that the survival curves were separable

as well as statistically significant (log-rank test p=3.4×10^{-7}).

Descriptive analysis

As seen in Table 6.1, no statistically significant differences were found in any of the demographic, clinical and MRI characteristics presented at baseline between patients who reached a PIRA event and patients who did not. However, there was some borderline evidence of patients with PIRA being older at the first demyelinating attack (p=0.06) and having higher lesion load (p=0.05).

When we compared the baseline characteristics between the two risk groups of patients, i.e., low- and high-risk, summarised in Table 6.2, we did not find any significant differences. Of note, no differences in the proportion of patients treated or the proportion of time under treatment were observed either.

6.5.3 Interpretability

Classical statistical model

The CPH model with age as the only predictor achieved a c-index of 0.59. An older age at the first demyelinating attack was significantly associated with a greater risk of PIRA (HR=1.04 [95% CI 1.01-1.08] p=0.02). When we added the min-max normalised deep learning-estimated hazard of PIRA to the age-based CPH model, the resulting CPH model with both covariates showed a c-index of 0.74. The HR for the deep learning-based hazard was 43.2 [95% CI 14.62-127.63] p=9.6×10^{-12}.

Similar model improvements were observed when the deep learning-estimated hazard was added to models adjusted for age, lesion load and GM, WM, and total intracranial volumes (see Table 6.3).

Cohort regional-based analysis

The cohort average SHAP values maps revealed that the most relevant areas for predicting PIRA were the frontal and parietal cortex, lateral ventricles, periventricular, frontal and parietal WM, as shown in Figure 6.4. Cortical areas had a more negative relevance, lateral ventricles a more positive one, while a similar relevance value was observed in periventricular WM.

Table 6.2: Demographic, clinical history and brain MRI characteristics of patients at baseline included in the analysis grouped by the post-inference risk stratification.

	Full cohort	**High-risk**	**Low-risk**	***p***
N, (%)	259	129 (50)	130 (50)	
Outcome, years, mean(SD)	7.9 (2.5)	6.8 (2.3)	8.8 (2.5)	< 0.001
Female, n(%)	170 (66)	86 (67)	84 (65)	0.83
Age at CIS, years, mean(SD)	34.3 (7.9)	34.9 (7.9)	33.8 (8.0)	0.24
Time of last FU, years, mean (SD)	7.9 (2.5)	6.8 (2.3)	8.8 (2.5)	< 0.001
Annualised relapse rate, mean (SD)	0.25 (0.18)	0.25 (0.14)	0.26 (0.22)	0.74
EDSS at CIS, median[range]	1.5 [0.0, 4.5]	1.0 [0.0, 3.5]	1.5 [0.0, 4.5]	0.37
EDSS at last FU, median[range]	1.0 [0.0, 6.0]	1.5 [0.0, 4.0]	1.0 [0.0, 6.0]	0.85
CIS topography, n(%)				0.76
Brainstem	49 (19)	26 (20.2)	23 (17.7)	
Optic nerve	100 (39)	45 (35)	55 (42.3)	
Spinal Cord	81 (31)	44 (34)	37 (28.5)	
Polyregional	12 (5)	6 (4.6)	6 (4.5)	
Other	17 (6)	8 (6.2)	9 (7)	
Presence of OB, n(%)				0.85
Positive	142 (55)	72 (56)	70 (54)	
Negative	117 (45)	57 (44)	60 (46)	
Spinal cord lesions, n(%)				0.77
0	126 (49)	65 (50.4)	61 (47)	
1	54 (21)	29 (22.5)	25 (19.2)	
2-3	27 (10)	13 (10.1)	14 (10.8)	
>3	34 (13)	15 (11.6)	19 (14.5)	
unknown	18 (7)	7 (5.4)	11 (8.5)	
Brain lesions, n(%)				0.2
0	60 (23)	26 (20.2)	34 (26)	

1-3	28 (11)	19 (14.7)	9 (7)	
4-8	43 (16)	21 (16.3)	22 (16)	
≥9	128 (49)	63 (48.8)	65 (50)	
Treated during FU, n (%)	161 (62)	81 (62.8)	80 (61.5)	0.94
Proportion time treated, mean (SD)	0.53 (0.43)	0.52 (0.42)	0.54 (0.44)	0.71
Scanner model[a], n(%)				0.79
Tim Trio	236 (91.1)	120 (93)	116 (89)	
Symphony	12 (4.6)	1 (0.8)	11 (84.5)	
Avanto	9 (3.5)	7 (5.4)	2 (1.5)	
Symphony Tim	2 (0.8)	1 (0.8)	1 (0.8)	
Lesion load, mL, mean(SD)	3.4 (4.3)	3.8 (4)	3.1 (4.4)	0.2
BPF, mean(SD)	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)	0.63
WMF, mean(SD)	0.46 (0.02)	0.46 (0.02)	0.46 (0.02)	0.23
GMF, mean(SD)	0.54 (0.02)	0.54 (0.02)	0.54 (0.02)	0.23

[a] By the central limit theorem, we can only calculate statistics on the Tim Trio scanner since the other scanners sample size was not large enough to assume a normal distribution.
CIS: clinically isolated syndrome, *OBs*: oligoclonal bands, *FU*: follow-up, *EDSS*: expanded disability status scale, *BPF*: brain parenchymal fraction, *GMF*: grey matter fraction, *WMF*: white matter fraction.

6.6 Discussion

In this Chapter, we have developed the first deep learning image-based survival model for future prediction of PIRA. Our findings show that a deep learning-based model only using brain MRI scans performed at the time of the first demyelinating event is able to accurately predict the survival function of the studied dataset. Thus, the survival probabilities estimated from our deep learning model could correctly identify in the high-risk group 74% and 83% of patients experiencing PIRA and early PIRA, respectively. Furthermore, our deep learning-based hazard remained as a significant predictor of PIRA when added to a survival model built with the strongest predictor of PIRA so far, according to the literature, which is

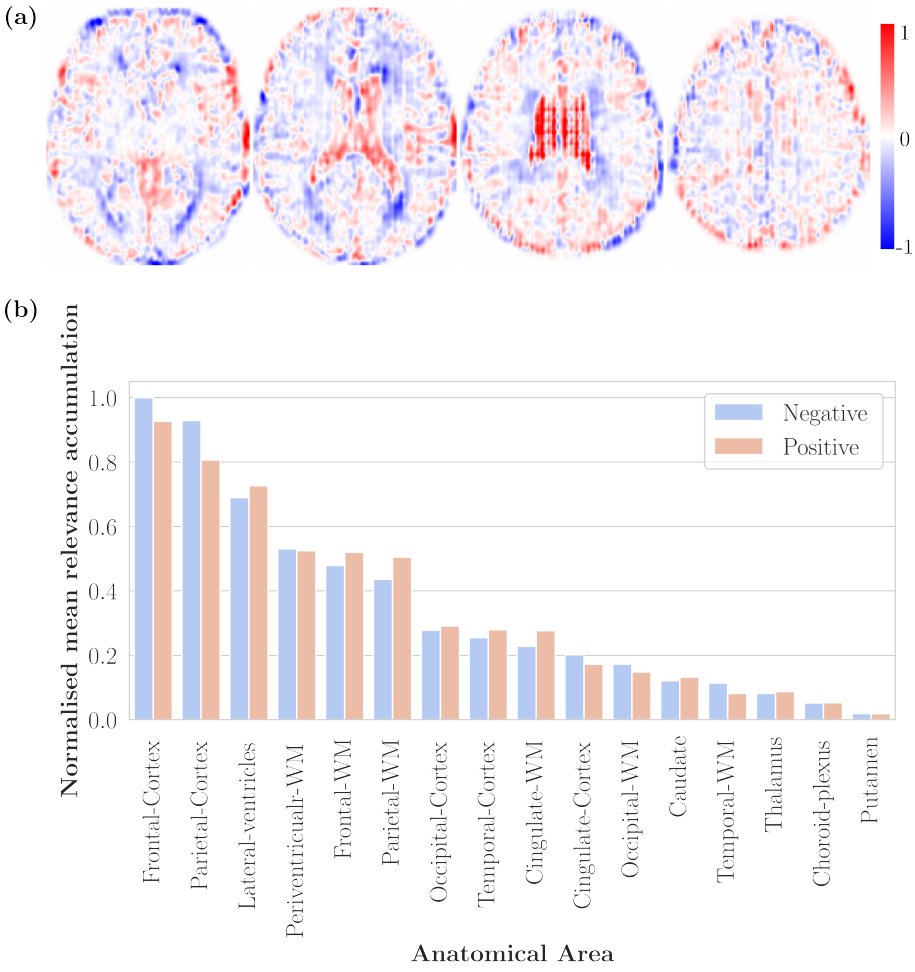

Figure 6.4: Average SHAP maps analysis. (a) Different slices of the mean SHAP maps calculated using the entire cohort. Negative (blue) values are associated with a higher risk of reaching a PIRA and positive (red) values to lower risk. The average map has been smoothed for a better visualisation. Additionally, the individual maps are binarised at 95% percentile of each relevance to quantify the pixel accumulation. (b) Mean pixel accumulation by each anatomical area for positive and negative relevance values.

Table 6.3: Resultant CPH models built with different input variables.

Model	Predictors	**HR** (95% CI), p	c-index
A	Age, years	1.04 (1.01, 1.08), p=0.018	0.59
B	Age, years	1.04 (1.00, 1.07), p=0.045	0.74
	Deep learning-estimated hazard rate	43.2 (14.62, 127.63), $p=9.6 \times 10^{-12}$	
C	Age, years	1.05 (1.01, 1.09), p=0.013	0.64
	Lesion load, mL	1.07 (1.02, 1.12), p=0.003	
	WM volume, mL	1.01 (1.0, 1.02), p=0.040	
	GM volume, mL	1.01 (1.0, 1.02), p=0.117	
	Total intracranial volume, mL	0.99 (0.98, 1.0), p=0.042	
D	Age, years	1.05 (1.01, 1.09), p=0.014	0.75
	Lesion load, mL	1.08 (1.04, 1.13), $p=3.7 \times 10^{-4}$	
	WM volume, mL	1.01 (1.0, 1.02), p=0.076	
	GM volume, mL	1.01 (1.0, 1.02), p=0.091	
	Total intracranial volume, mL	0.99 (0.99, 1.0), p=0.056	
	Deep learning-estimated hazard rate	67.45 (20.17, 225.51), $p=7.9 \times 10^{-12}$	

HR: Hazard Ratio, *CI*: Confidence Interval, *c-index*: concordance index, *GM*: grey matter, *WM*: white matter.

age at the time of the first attack. Finally, thanks to the relevance maps obtained from the deep learning model, we revealed that the anatomical areas with the greatest contribution to the output were the frontal and parietal cortex followed by lateral ventricles and periventricular WM.

Although several publications based on clinical data at baseline have shown the descriptive survival curves of large cohorts of patients presenting -or not- PIRA [16, 17], none of these studies has provided future prognostic predictions or has exploited the whole capacity of features from MRI that might be relevant for MS progression. Nonetheless, they all agree on the association between older age at CIS and a greater risk of PIRA [16, 17].

In this context, during the examination of our baseline descriptive clinical data

from patients who achieved a PIRA and from those who did not, we did not find any features that were significantly different between the two groups. The two variables that came closest to reaching statistical significance were age at the first attack and lesion load noted in the initial MRI. Notably, a higher age at CIS tended to be associated with patients who experienced a PIRA event (p=0.06), a pattern consistent with findings from earlier research [17]. Patients who reached a PIRA associated with a higher lesion load, an MRI-extracted biomarker, showed no significant difference between groups (p=0.05). In studies relying solely on clinical data [43], the presence of lesions is typically expressed using categorical numbers or lesion range, and in our case, there was no significant difference between groups either (p=0.32).

Our findings showed not only an accurate prediction of PIRA based on the deep learning estimates, but also that such estimates were able to improve conventional survival models of PIRA based on age at first attack, the strongest predictor of PIRA at the time of the first attack. The survival curve prediction closely followed the observable survival curve, represented with the Kaplan-Meier curve estimator (Figure 6.3(a)). Our estimations at earlier time intervals proved to be better, according to the high correct identification of PIRA —and especially early PIRA— of our survival probability threshold, based on estimations exactly at the first time-interval (Figure 6.3(b)). This might be explained by the fact that the density of population still "surviving" on those intervals was higher than in the later ones. This reinforces the importance of large cohorts with extended clinical follow-up, as they offer a more comprehensive representation of potential findings, even when patient variability remains a factor.

When we compared low-risk vs high-risk groups, we found that these were not significantly different either in terms of clinical or paraclinical data. Age and lesion load at first attack, which were the ones closest to being significant when comparing patients with vs without a future PIRA event, either were not significantly different when comparing high- vs low-risk PIRA groups. Two main ideas may be extracted from this behaviour. First, having in mind that the risk groups are balanced, we divided the dataset into two equally sized groups and our cohort was not. However, the consistent, non-separable behavior observed in both the real and predicted groups, based on clinical data, supports our hypothesis that if there is predictive power at baseline, it could be found in the MRI scans. Secondly, losing the significance of lesion load in the predicted risk groups (p=0.2) revealed that, at least at baseline MRI, the presence of T2-lesions might be

relevant but the deep learning model did not only focused on that.

Furthermore, our deep learning survival model provided relevance maps based on the estimated SHAP values, which inform of the most relevant regions in the input data, i.e., brain regions in our case, for the estimation of the survival probabilities. These relevance maps highlighted the importance of the integrity of the cortical GM for the future development of PIRA. More specifically, frontal and parietal cortical areas were particularly relevant for negative (unfavourable) prognostic predictions, that is, for the prediction of PIRA. This is in line with the findings from Cagol *et al.* [43], who found a strong correlation between an increased cortical GM volume loss in people with MS and a concurrent experience of PIRA. However, to the best of our knowledge, we are the first to provide evidence of the relevance of the cortical GM integrity at the time of the first attack for the future development of PIRA. Of note, after the frontal and parietal cortical areas, the most relevant regions to predict PIRA were the lateral ventricles, and the periventricular WM, followed by the subcortical WM in the frontoparietal regions. This would support the relevance of WM integrity too for the development of PIRA, as other authors have already pinpointed [14, 43, 199, 200]. On the other hand, the lateral ventricles were not directly associated with a worse prognosis. In longitudinal or cross-sectional studies, the increase in size of lateral ventricles has found to be correlated with lower whole brain volume [219]. Therefore, this positive contribution can be viewed as a significant focus on a potential feature that remains undiscovered. Periventricular WM (coloured in blue and red) demonstrated relevance in both prognostic directions. This brain region is where T2-lesions most frequently appear, and a higher lesion load has been correlated with disease progression, including reaching PIRA [14, 220]. The bidirectional nature of these contributions may reflect the influence of patients with and without baseline lesions. In this context, the model assigns negative relevance (higher risk of reaching a PIRA) when lesions are present, and positive relevance when they are absent, capturing the distinct impact of lesion presence on the prognosis. Additionally, it is important to consider that our study specifically focuses on the outcome of PIRA, whereas other mechanisms contribute to disease progression in MS. Nevertheless, the most striking aspect of our research is that we have been able to ascertain the importance of those tissue types and regions using a single MRI time-point and without the need for a pre-selection of MRI metrics, such as longitudinal atrophy rates, whose estimation is always subjected to measurement

error. All this again, brings to light the potential for deep learning-based models for predictive purposes in clinical practice and research.

This study is not without limitations. The sample size and more importantly the low density of patients at longer follow-up time intervals made it difficult to define how robustly our model could generalise to future patients or patients from another cohort. Moreover, the uniqueness of our deeply phenotyped first-attack cohort, both in terms of the quality of the data and length of follow-up, made it very difficult to find a similarly rich first-attack cohort where we could test our deep learning model. On the other hand, this comparison, once it is possible, will be of great value to prove the robustness of our model and its capacity to generalise to unseen cohorts. Methodologically, our choice to use a 2D CNN was made based on the available data to build such a model. Using 2D slices as input increases the available training data, but it comes at the expense of the spatial dimensionality offered by processing the entire 3D volume. Most of the deep learning image-based survival models published so far, apart from using the image itself as input, include clinical or demographic data also as input to such models, which could be a future work to implement on our proposal. Another important future addition to the models may be the inclusion of spinal cord data apart from brain MRI data, which has already proved to be relevant for the development of PIRA [200]. Additionally, our predictive model could not take into account the potential effects of treatment, since its input data was only the MRI scans at the time of the first attack, before any treatment could be even proposed. Finally, concerning the interpretability maps, by definition they are limited by the lack of a ground truth. Particularly, the ambiguous outcomes regarding potential biomarker extraction from these maps make this particular question even more open and ripe for exploration in future research endeavors. The development or refinement of explainability algorithms for regression tasks or tasks that are not only based on the classification of known classes is therefore required.

In conclusion, to the best of our knowledge, we present the first deep learning image-based survival model to predict future prognosis of patients with MS from routinely acquired structural MRI scans performed at the time of first demyelinating event. Our model accurately predicted who would develop PIRA, especially early PIRA, and improved a classical statistical survival model based only on age at symptom onset. These results may have important implications for clinical practice, since for the first time, we may be able to identify, with an only

brain MRI scan performed at symptom onset, and with a more than acceptable risk association, who will develop PIRA, especially early PIRA. Furthermore, our deep learning survival model provided relevance maps, highlighting the importance of the integrity of the cortical GM for the future development of PIRA. In sum, our results suggest that the proposed deep learning model could be a useful tool for the clinical management of patients with MS at symptom onset and therefore it may be worth exploring its impact in clinical practice in the future.

Chapter 7

Conclusions

The aim of this PhD Thesis has been to propose and develop deep learning tools able to detect and predict future disability accumulation in MS, and thus possibly assist physicians in their clinical practice in the near future. After an extensive examination of the current state of the art concerning predictive models for early and long-term prognosis in MS, using routinely-acquired structural MRI data, it has become clear that the literature falls short in providing conclusive evidence regarding the direct relationship between outputs obtained from MRI scans and the gradual development of disability. Although several studies have explored correlations between MS imaging biomarkers and the progression of disability, they have not yet sufficiently addressed the fundamental question: "What is the precise association between EDSS and the characteristics of its corresponding brain MRI scan?" With the aim to address this question, we have centred on the use and analysis of different deep learning models to do so. We first conducted a cross-sectional study attempting to identify if a single brain MRI scan, at any disease time-point, could correlate with its corresponding clinical score evaluation, EDSS or PDDS. Although we used the entire brain MRI volume as input to a classification model, we also explored and validated several classification models based on regional image inputs, assessing the potential impact of each one of those when compared to the whole brain volume. Finally, we presented the first deep learning image-based survival model for predicting

the development of a PIRA event. In both cross-sectional and future prognosis models, we introduced a pipeline that could evaluate the interpretability of the results obtained with a deep learning model, thereby revealing the black-box nature of those implementations.

Following the same objectives outlined in the Introduction, the following Section summarises the main conclusions and contributions of this PhD Thesis: (i) two different deep learning models predictors associated with MS disability progression, each with their respective (ii) interpretability algorithms that highlight the brain regions implicated in higher disability status or an increased risk of disease progression.

7.1 Contributions

First, we proposed and evaluated the use of a ResNet-based architecture to create a binary classification model to stratify brain MRI scans based on their cross-sectional EDSS stage: those with EDSS$<$3.0 and those with EDSS$\geq$3.0. For that, we used a single brain MRI scan, T1-w and FLAIR sequences, acquired at any time-point of the disease course. The results demonstrated a high accuracy (79%), which was consistently validated using an external database. Additionally, the use of an interpretability algorithm allowed us to reveal that frontotemporal cortex and cerebellum areas could be key regions regarding the mechanisms contributing to disability accumulation in MS. The main findings of this study have been published in the following journal paper:

- Deciphering multiple sclerosis disability with deep learning attention maps on clinical MRI. *NeuroImage: Clinical*, 38, art 103376, 2023. [JCR N IF 4.891, Q2(5/14)]

We further evaluated and compared the proposed cross-sectional model by exploring various regional input strategies to use within the same network architecture. The results obtained from these regional deep learning models –WM, GM, subcortical GM, lateral ventricles and brainstem structures– revealed that a CNN-based model could successfully extract features leading to accurate classification, regardless of the chosen input strategy. Throughout this analysis, we directed the network to focus on specific regions that, a priori, are considered important in MS. Notably, larger regions, GM and WM tissues outperformed smaller regions such as ventricles and brainstem regions. When assessing the

model's performance on an external database, the whole brain global approach was the one that kept offering the best trade-off in both datasets. This study and the obtained results have been published in the following journal paper:

- Global and regional deep learning models for multiple sclerosis stratification from MRI. *Journal of Magnetic Resonance Imaging* [JCR N IF 5.119, Q1(34/136)]

Finally, we proposed and evaluated a discrete-time survival image-based deep learning model designed to predict the survival function of patients with MS who may experience PIRA. As far as we know, this is the first proposal of this kind, i.e., future prediction independent of the follow-up time, in MS. Once again, we solely used brain MRI scans (T1-w and FLAIR sequences) to predict the future prognosis of patients with MS at the time of first demyelinating attack. The obtained results showed a promising accuracy to predict future development of PIRA, particularly in its early stages. The achieved performance insights can also be harnessed to enhance the effectiveness of conventional statistical models, which are widely accepted and used in the clinical domain. Moreover, we successfully derived relevance maps to ensure interpretability, treating our results as a regression task. These maps revealed the importance of the cortical GM, especially within the frontal and parietal regions, to predict a future PIRA event.

The proposal and results presented in Chapter 6, have been recently submitted for publication under the following title:

- Deep learning to predict progression independent of relapse activity at first demyelinating event.

7.2 Future work

The analysis of brain MRI scans for patients with MS is a multifaceted topic that encompasses several aspects and involves multiple lines of research. In this PhD Thesis, grounded in a consistent clinical perspective, we have endeavored to develop robust and comprehensible solutions tailored for the end-users, the clinicians, while holding a potentially inestimable value for people with MS, who may benefit, in the future, from enhanced decision algorithms based on deep learning-based research on disease prediction. To extend beyond the scope of the aforementioned studies, we introduce different potential directions that can

be pursued, encompassing (i) enhancements to our proposed solutions, and (ii) prospects for long-term future research.

7.2.1 Short-term proposal improvements

In this PhD Thesis, we have introduced two different image-based deep learning pipelines for predicting the disability stage of patients with MS, from both cross-sectional and longitudinal perspectives. Additionally, for each of the pipelines, we have proposed the integration of an interpretability algorithm to uncover the decision-making processes within the model itself. As demonstrated across various Chapters, this is an emerging and significant subject. The primary challenge, alongside the limited availability of data for constructing and evaluating such models, lies in the absence of definitive explanations for the widely variable disease progression observed among individuals. A more extensive dataset and external validation encompassing diverse cases are, therefore, needed. In our specific context, significant advancements in domain adaptation could be achieved through the inclusion of scans from various MR vendors and mulitcentre studies, thus improving models' generalisability.

There are several short-term improvements that can be made to our proposals. Concerning the proposed deep learning models, which only used the brain MRI scans as main input, it could be beneficial to explore the inclusion of additional clinical data such as demographics, available for all patients. Similarly, in terms of cohort selection, we did not consider factors like the presence of treatment or modifying behaviours, which are of great clinical significance. Furthermore, potential modifications to our pipelines could involve integrating classical statistical models in conjunction with deep learning models for the final output operation.

More specifically, regarding Chapters 4 and 5, a potential test for domain adaptation could involve reversing the study: training on MS PATHS dataset and using the VHUH cohort as independent dataset. This approach could help unveil any potential differences that may arise due to this exchange, even when using data from the same MR vendor. In Chapter 6, the availability of data to assess the performance on an external and independent dataset would greatly enhance its value. Additionally, the proposed pipeline could also be evaluated on different clinical outcomes, such as predicting future disability stages or exploring other mechanisms of progression.

Regarding interpretability, several additional analyses could be performed to enhance the clarity of our findings. One essential aspect is to discern whether the relevant imaging features highlighted in cortical regions are a result of registration or segmentation imperfections, or genuine target-relevance. Moreover, specifically in relation to Chapter 6, there is a strong demand for exploring and implementing alternative interpretability algorithms able to explain regression tasks or scenarios where the ground truth is not explicitly defined.

As a final aim for future implementation, the applicability of the presented tools in clinical practice needs to be tested. For this purpose, the creation of standalone tools to facilitate the sharing and easy deployment of applications becomes essential.

7.2.2 Future research lines

In the long-term, there are several new research lines departing from this PhD Thesis that could be studied. For example, the use of longitudinal MRI data to predict future prognosis. Our studies only included baseline data to predict how the patient will act in the future. However, having a baseline and follow-up scan, could actually quantify the longitudinal differences that may characterise a progression profile or another. To achieve this goal, available longitudinal data is therefore required. If it is not available, a possible additional methodological implementation could be the use of synthetic images to expand the available cohort or to generate longitudinal data.

Another potential study that could stem from this PhD Thesis is the use of a different type of data as the primary input. For instance, functional MRI, although less commonly used in clinical practice, but whose research in MS is rapidly evolving. Ongoing studies continue to shed light on the complex relationship between brain function and MS, which can contribute to a deeper understanding of the MS pathological underpinnings of the disease, potentially leading to improved diagnostic techniques and prediction of progression profiles.

Finally, the methods and concepts presented in this PhD Thesis could also be applied to the study of other brain diseases with similar characteristics, with the appropriate retraining and evaluation.

Bibliography

[1] Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology. 2018 2;17(2):162-73.

[2] Steinman, M D L. Multiple Sclerosis: A Coordinated Immunological Attack against Myelin in the Central Nervous System. Cell. 1996 5;85(3):299-302.

[3] Chiaravalloti ND, DeLuca J. Cognitive impairment in multiple sclerosis. The Lancet Neurology. 2008 12;7(12):1139-51.

[4] Vandebergh M, Degryse N, Dubois B, Goris A. Environmental risk factors in multiple sclerosis: bridging Mendelian randomization and observational studies. Journal of Neurology. 2022;269(8):4565-74.

[5] Kim W, Patsopoulos NA. Genetics and functional genomics of multiple sclerosis. Seminars in Immunopathology. 2022;44(1):63-79.

[6] Hone L, Giovannoni G, Dobson R, Jacobs BM. Predicting Multiple Sclerosis: Challenges and Opportunities. Frontiers in Neurology. 2022;12.

[7] MS International Federation. Atlas of MS; 2023. Available from: `www.atlasofms.org`.

[8] McGinley MP, Goldschmidt CH, Rae-Grant AD. Diagnosis and Treatment of Multiple Sclerosis: A Review. JAMA. 2021 2;325(8):765-79.

[9] Gaitán MI, Correale J. Multiple Sclerosis Misdiagnosis: A Persistent Problem to Solve. Frontiers in Neurology. 2019;10.

[10] McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, et al. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. Annals of Neurology. 2001 7;50(1):121-7.

[11] Brownlee WJ, Hardy TA, Fazekas F, Miller DH. Diagnosis of multiple sclerosis: progress and challenges. The Lancet. 2017;389(10076):1336-46.

[12] Müller J, Cagol A, Lorscheider J, Tsagkas C, Benkert P, Yaldizli O, et al. Harmonizing Definitions for Progression Independent of Relapse Activity in Multiple Sclerosis: A Systematic Review. JAMA Neurology. 2023 10.

[13] Kappos L, Butzkueven H, Wiendl H, Spelman T, Pellegrini F, Chen Y, et al. Greater sensitivity to multiple sclerosis disability worsening and progression events using a roving versus a fixed reference value in a prospective cohort study. Multiple Sclerosis Journal. 2017 5;24(7):963-73.

[14] Kappos L, Wolinsky JS, Giovannoni G, Arnold DL, Wang Q, Bernasconi C, et al. Contribution of Relapse-Independent Progression vs Relapse-Associated Worsening to Overall Confirmed Disability Accumulation in Typical Relapsing Multiple Sclerosis in a Pooled Analysis of 2 Randomized Clinical Trials. JAMA Neurology. 2020;77(9):1132-40.

[15] Lublin FD, Häring DA, Ganjgahi H, Ocampo A, Hatami F, Čuklina J, et al. How patients with multiple sclerosis acquire disability. Brain : a journal of neurology. 2022;145(9):3147-61.

[16] Portaccio E, Bellinvia A, Fonderico M, Pastò L, Razzolini L, Totaro R, et al. Progression is independent of relapse activity in early multiple sclerosis: a real-life cohort study. Brain. 2022;145(8):2796-805.

[17] Tur C, Carbonell-Mirabent P, Cobo-Calvo A, Otero-Romero S, Arrambide G, Midaglia L, et al. Association of Early Progression Independent of Relapse Activity with Long-term Disability after a First Demyelinating Event in Multiple Sclerosis. JAMA Neurology. 2023 2;80(2):151-60.

[18] Tintore M, Rovira A, Río J, Otero-Romero S, Arrambide G, Tur C, et al. Defining high, medium and low impact prognostic factors for developing multiple sclerosis. Brain. 2015;138(Pt 7):1863-74.

[19] De Stefano N, Matthews PM, Filippi M, Agosta F, De Luca M, Bartolozzi ML, et al. Evidence of early cortical atrophy in MS. Neurology. 2003 4;60(7):1157 LP 1162.

[20] Calabrese M, Castellaro M. Cortical Gray Matter MR Imaging in Multiple Sclerosis. Neuroimaging Clinics of North America. 2017 5;27(2):301-12.

[21] Sastre-Garriga J, Pareto D, Battaglini M, Rocca MA, Ciccarelli O, Enzinger C, et al. MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. Nature Reviews Neurology. 2020;16(3):171-82.

[22] Bonacchi R, Filippi M, Rocca MA. Role of artificial intelligence in MS clinical practice. NeuroImage: Clinical. 2022;35(May):103065.

[23] Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Martí R, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artificial Intelligence in Medicine. 2019;95:64-81.

[24] Lladó X, Ganiler O, Oliver A, Marti R, Freixenet J, Valls L, et al. Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology. 2012;54(8):787-807.

[25] Kontopodis E, Papadaki E, Trivzakis E, Maris T, Simos P, Papadakis G, et al. Emerging deep learning techniques using magnetic resonance imaging data applied in multiple sclerosis and clinical isolated syndrome patients (Review). Experimental and Therapeutic Medicine. 2021;22(4):1-17.

[26] Ma Y, Zhang C, Cabezas M, Song Y, Tang Z, Liu D, et al. Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images: Techniques and Clinical Applications. IEEE Journal of Biomedical and Health Informatics. 2022;26(6):2680-92.

[27] Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in Biology and Medicine. 2022;140(October 2021):105111.

[28] van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis. 2022;79:102470.

[29] Seccia R, Romano S, Salvetti M, Crisanti A, Palagi L, Grassi F. Machine Learning Use for Prognostic Purposes in Multiple Sclerosis. Life. 2021;11(2):122.

[30] Wottschel V, Alexander DC, Kwok PP, Chard DT, Stromillo ML, De Stefano N, et al. Predicting outcome in clinically isolated syndrome using machine learning. NeuroImage: Clinical. 2015;7:281-7.

[31] Roca P, Attye A, Colas L, Tucholka A, Rubini P, Cackowski S, et al. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. Diagnostic and Interventional Imaging. 2020 12;101(12):795-802.

[32] Storelli L, Azzimonti M, Gueye M, Vizzino C, Preziosa P, Tedeschi G, et al. A Deep Learning Approach to Predicting Disease Progression in Multiple Sclerosis Using Magnetic Resonance Imaging. Investigative radiology. 2022;57(7):423-32.

[33] Coll L, Pareto D, Carbonell-Mirabent P, Cobo-Calvo A, Arrambide G, Vidal-Jordana A, et al. Deciphering multiple sclerosis disability with deep learning attention maps on clinical MRI. NeuroImage: Clinical. 2023;38(March):103376.

[34] Coll L, Pareto D, Carbonell-Mirabent P, Cobo-Calvo A, Arrambide G, Vidal-Jordana A, et al. Global and Regional Deep Learning Models for Multiple Sclerosis Stratification From MRI. Journal of Magnetic Resonance Imaging. 2023:1-10.

[35] Hectors SJCG, Jacobs I, Moonen CTW, Strijkers GJ, Nicolay K. MRI methods for the evaluation of high intensity focused ultrasound tumor treatment: Current status and future needs. Magnetic Resonance in Medicine. 2016 1;75(1):302-17.

[36] Gui L, Loukas S, Lazeyras F, Hüppi PS, Meskaldji DE, Borradori Tolsa C. Longitudinal study of neonatal brain tissue volumes in preterm infants

and their ability to predict neurodevelopmental outcome. NeuroImage. 2019;185:728-41.

[37] Derks SHAE, van der Veldt AAM, Smits M. Brain metastases: the role of clinical imaging. The British Journal of Radiology. 2021 12;95(1130):20210944.

[38] Wattjes MP, Ciccarelli O, Reich DS, Banwell B, de Stefano N, Enzinger C, et al. 2021 MAGNIMS CMSC NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. The Lancet Neurology. 2021 8;20(8):653-70.

[39] Courchesne E, Chisum HJ, Townsend J, Cowles A, Covington J, Egaas B, et al. Normal Brain Development and Aging: Quantitative Analysis at in Vivo MR Imaging in Healthy Volunteers. Radiology. 2000 9;216(3):672-82.

[40] Jacobsen C, Hagemeier J, Myhr KM, Nyland H, Lode K, Bergsland N, et al. Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. Journal of Neurology, Neurosurgery & Psychiatry. 2014 10;85(10):1109 LP 1115.

[41] Rocca MA, Battaglini M, Benedict RHB, Stefano ND, Geurts JJG, Henry RG, et al. Brain MRI atrophy quantification in MS. Neurology. 2017 1;88(4):403 LP 413.

[42] Sastre-Garriga J, Pareto D, Rovira A. Brain Atrophy in Multiple Sclerosis: Clinical Relevance and Technical Aspects. Neuroimaging Clinics of North America. 2017;27(2):289-300.

[43] Cagol A, Schaedelin S, Barakovic M, Benkert P, Todea RA, Rahmanzadeh R, et al. Association of Brain Atrophy with Disease Progression Independent of Relapse Activity in Patients with Relapsing Multiple Sclerosis. JAMA Neurology. 2022;79(7):682-92.

[44] Mulder ER, de Jong RA, Knol DL, van Schijndel RA, Cover KS, Visser PJ, et al. Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. NeuroImage. 2014;92:169-81.

[45] Battaglini M, Jenkinson M, De Stefano N. SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI. Human Brain Mapping. 2018;39(3):1063-77.

[46] Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, et al. Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis. NeuroImage. 2002;17(1):479-89.

[47] Nakamura K, Guizard N, Fonov VS, Narayanan S, Collins DL, Arnold DL. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. NeuroImage: Clinical. 2014;4:10-7.

[48] Leung KK, Ridgway GR, Ourselin S, Fox NC. Consistent multi-time-point brain atrophy estimation from the boundary shift integral. NeuroImage. 2012;59(4):3995-4005.

[49] Cao P, Gao J, Zhang Z. Multi-view based multi-model learning for MCI diagnosis. Brain Sciences. 2020;10(3).

[50] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021;2(3):160.

[51] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

[52] Goodfellow I, Bengio Y, Courville A. Deep learning . Adaptive computation and machine learning series. Cambridge, MA: MIT Press; 2017.

[53] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net. ICLR 2015. 2014 12.

[54] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10); 2010. p. 807-14.

[55] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint. 2014 9.

[56] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016:770-8.

[57] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017:2261-9.

[58] Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97; 2019. p. 6105-14.

[59] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. 9th International Conference on Learning Representations. 2021.

[60] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6.

[61] Santos CFGD, Papa JP. Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. ACM Computing Surveys. 2022;54(10 s).

[62] Ras G, Xie N, van Gerven M, Doran D. Explainable Deep Learning: A Field Guide for the Uninitiated. Journal of Artificial Intelligence Research. 2022;73:329-96.

[63] Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. npj Digital Medicine. 2022;5(1).

[64] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015;07-12-June:3156-64.

[65] Zhang Z, Chen P, Sapkota M, Yang L. TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References. In: Medical Image Computing and Computer Assisted Intervention MICCAI 2017; 2017. p. 320-8.

[66] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997 11;9(8):1735-80.

[67] Uehara K, Murakawa M, Nosato H, Sakanashi H. Prototype-based interpretation of pathological image analysis by convolutional neural networks. In: Asian Conference on Pattern Recognition. Springer; 2019. p. 640-52.

[68] Wang CJ, Hamm CA, Savic LJ, Ferrante M, Schobert I, Schlachter T, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. European Radiology. 2019;29(7):3348-57.

[69] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ICLR 2014. 2013 12.

[70] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Computer Vision – ECCV 2014; 2014. p. 818-33.

[71] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2921-9.

[72] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. IEEE International Conference on Computer Vision. 2017.

[73] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE. 2015;10(7).

[74] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;2017-Decem(Section 2):4766-75.

[75] Jetley S, Lord NA, Lee N, Torr PHS. Learn to pay attention. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. 2018:1-14.

[76] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;13-17-Augu:1135-44.

[77] Fong RC, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. Proceedings of the IEEE International Conference on Computer Vision. 2017;2017-Octob:3449-57.

[78] Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: Prediction difference analysis. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. 2017:1-12.

[79] Uzunova H, Ehrhardt J, Kepp T, Handels H. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In: Proc.SPIE. vol. 10949; 2019. p. 1094911.

[80] Zhu P, Ogino M. Guideline-based additive explanation for computer-aided diagnosis of lung nodules. vol. 11797 LNCS. Springer International Publishing; 2019.

[81] Shoeibi A, Khodatars M, Jafari M, Moridian P, Rezaei M, Alizadehsani R, et al. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. Computers in Biology and Medicine. 2021 9;136:104697.

[82] McKinley R, Wepfer R, Aschwanden F, Grunder L, Muri R, Rummel C, et al. Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks. Scientific Reports. 2021;11(1):1-11.

[83] Billot B, Cerri S, Leemput KV, Dalca AV, Iglesias JE. Joint segmentation of multiple sclerosis lesions and brain anatomy in mri scans of any contrast and resolution with CNNs. Proceedings - International Symposium on Biomedical Imaging. 2021;2021-April:1971-4.

[84] Gabr RE, Coronado I, Robinson M, Sujit SJ, Datta S, Sun X, et al. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. Multiple Sclerosis Journal. 2020;26(10):1217-26.

[85] Clèrigues A, Valverde S, Salvi J, Oliver A, Lladó X. Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry. Computerized Medical Imaging and Graphics. 2023;103(May 2022).

[86] Tang Z, Cabezas M, Liu D, Barnett M, Cai W, Wang C. LG-Net: Lesion Gate Network for Multiple Sclerosis Lesion Inpainting. In: Medical Image Computing and Computer Assisted Intervention 2021; 2021. p. 660-9.

[87] Xiong H, Wang C, Barnett M, Wang C. Multiple Sclerosis Lesion Filling Using a Non-lesion Attention Based Convolutional Network. In: Neural Information Processing; 2020. p. 448-60.

[88] Manjón JV, Romero JE, Vivo-Hernando R, Rubio G, Aparici F, de la Iglesia-Vaya M, et al. Blind MRI Brain Lesion Inpainting Using Deep Learning. In: Simulation and Synthesis in Medical Imaging; 2020. p. 41-9.

[89] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. Advances in Neural Information Processing Systems. 2014 6.

[90] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. Artificial Intelligence in Medicine. 2020;109:101938.

[91] Salem M, Valverde S, Cabezas M, Pareto D, Oliver A, Salvi J, et al. Multiple Sclerosis Lesion Synthesis in MRI Using an Encoder-Decoder U-NET. IEEE Access. 2019;7:25171-84.

[92] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention 2015; 2015. p. 234-41.

[93] Basaran BD, Qiao M, Matthews PM, Bai W. Subject-Specific Lesion Generation and Pseudo-Healthy Synthesis for Multiple Sclerosis Brain Images. In: Simulation and Synthesis in Medical Imaging; 2022. p. 1-11.

[94] Kamraoui RA, Mansencal B, Manjon JV, Coupé P. Longitudinal detection of new MS lesions using deep learning. Frontiers in Neuroimaging. 2022;1.

[95] Wei W, Poirion E, Bodini B, Durrleman S, Colliot O, Stankoff B, et al. Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis. Journal of Medical Imaging. 2019 2;6(1):14005.

[96] Valencia L, Clèrigues A, Valverde S, Salem M, Oliver A, Rovira A, et al. Evaluating the use of synthetic T1-w images in new T2 lesion detection in multiple sclerosis. Frontiers in Neuroscience. 2022.

[97] Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Scientific Reports. 2018;8(1):1-17.

[98] Zhang H, Oguz I. Multiple Sclerosis Lesion Segmentation - A Survey of Supervised CNN-Based Methods. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries; 2021. p. 11-29.

[99] Roy S, Butman JA, Reich DS, Calabresi PA, Pham DL. Multiple Sclerosis Lesion Segmentation from Brain MRI via Fully Convolutional Neural Networks. arXiv preprint. 2018.

[100] Aslani S, Dayan M, Storelli L, Filippi M, Murino V, Rocca MA, et al. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage. 2019;196:1-15.

[101] Birenbaum A, Greenspan H. Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. Engineering Applications of Artificial Intelligence. 2017;65:111-8.

[102] Zhang H, Valcarcel AM, Bakshi R, Chu R, Bagnato F, Shinohara RT, et al. Multiple Sclerosis Lesion Segmentation with Tiramisu and 2.5D Stacked Slices. In: Medical Image Computing and Computer Assisted Intervention; 2019. p. 338-46.

[103] Valverde S, Cabezas M, Roura E, González-Villà S, Pareto D, Vilanova JC, et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage. 2017;155:159-68.

[104] Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. Medical Image Analysis. 2020;59:101557.

[105] La Rosa F, Abdulkadir A, Fartaria MJ, Rahmanzadeh R, Lu PJ, Galbusera R, et al. Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. NeuroImage: Clinical. 2020;27:102335.

[106] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021;18(2):203-11.

[107] Carass A, Roy S, Jog A, Cuzzocreo JL, Magrath E, Gherman A, et al. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage. 2017;148:77-102.

[108] Commowick O, Cervenansky F, Cotton F, Dojat M. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. MICCAI 2021 - 24th International Conference on Medical Image Computing and Computer Assisted Intervention. 2021:126-2021.

[109] Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. IEEE Transactions on Biomedical Engineering. 2022;69(3):1173-85.

[110] Valverde S, Salem M, Cabezas M, Pareto D, Vilanova JC, Ramió-Torrentà L, et al. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clinical. 2019;21:101638.

[111] Dewey BE, Zhao C, Reinhold JC, Carass A, Fitzgerald KC, Sotirchos ES, et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magnetic Resonance Imaging. 2019;64:160-70.

[112] Ackaouy A, Courty N, Vallée E, Commowick O, Barillot C, Galassi F. Unsupervised Domain Adaptation With Optimal Transport in Multi-Site Segmentation of Multiple Sclerosis Lesions From MRI Data. Frontiers in Computational Neuroscience. 2020;14.

[113] Karaca Y, Cattani C, Moonis M. Comparison of Deep Learning and Support Vector Machine Learning for Subgroups of Multiple Sclerosis. In: Computational Science and Its Applications; 2017. p. 142-53.

[114] Marzullo A, Kocevar G, Stamile C, Durand-Dubief F, Terracina G, Calimeri F, et al. Classification of Multiple Sclerosis Clinical Profiles via Graph Convolutional Neural Networks. Frontiers in Neuroscience. 2019;13.

[115] Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. NeuroImage: Clinical. 2019;24:102003.

[116] Eshaghi A, Young AL, Wijeratne PA, Prados F, Arnold DL, Narayanan S, et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. Nature Communications. 2021 12;12(1):2078.

[117] Cruciani F, Brusini L, Zucchelli M, Retuci Pinheiro G, Setti F, Boscolo Galazzo I, et al. Interpretable deep learning as a means for decrypting disease signature in multiple sclerosis. Journal of Neural Engineering. 2021;18(4).

[118] Yoo Y, Tang LYW, Brosch T, Li DKB, Kolind S, Vavasour I, et al. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. NeuroImage: Clinical. 2018;17:169-78.

[119] Siar H, Teshnehlab M. Diagnosing and Classification Tumors and MS Simultaneous of Magnetic Resonance Images Using Convolution Neural Network. In: 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS); 2019. p. 1-4.

[120] Geraldes R, Ciccarelli O, Barkhof F, De Stefano N, Enzinger C, Filippi M, et al. The current role of MRI in differentiating multiple sclerosis from its imaging mimics. Nature Reviews Neurology. 2018;14(4):199-213.

[121] Kim H, Lee Y, Kim YH, Lim YM, Lee JS, Woo J, et al. Deep Learning-Based Method to Differentiate Neuromyelitis Optica Spectrum Disorder From Multiple Sclerosis. Frontiers in Neurology. 2020;11(November):1-8.

[122] Wang Z, Yu Z, Wang Y, Zhang H, Luo Y, Shi L, et al. 3D Compressed Convolutional Neural Network Differentiates Neuromyelitis Optical Spectrum Disorders From Multiple Sclerosis Using Automated

White Matter Hyperintensities Segmentations. Frontiers in Physiology. 2020;11.

[123] Rocca MA, Anzalone N, Storelli L, Del Poggio A, Cacciaguerra L, Manfredi AA, et al. Deep Learning on Conventional Magnetic Resonance Imaging Improves the Diagnosis of Multiple Sclerosis Mimics. Investigative Radiology. 2021;56(4).

[124] Maggi P, Fartaria MJ, Jorge J, La Rosa F, Absinta M, Sati P, et al. CVSnet: A machine learning approach for automated central vein sign assessment in multiple sclerosis. NMR in Biomedicine. 2020;33(5):1-11.

[125] Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. Journal of Neuroscience Methods. 2021;353(August 2020):109098.

[126] Bendfeldt K, Taschler B, Gaetano L, Madoerin P, Kuster P, Mueller-Lenke N, et al. MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. Brain Imaging and Behavior. 2019;13(5):1361-74.

[127] Wottschel V, Chard DT, Enzinger C, Filippi M, Frederiksen JL, Gasperini C, et al. SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis. NeuroImage: Clinical. 2019;24:102011.

[128] Tousignant A, Lemaître P, Doina P, Arnold DL, Arbel T. Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data. Proceedings of Machine Learning Research. 2019;102:483-92.

[129] Yoo Y, Tang LYW, Li DKB, Metz L, Kolind S, Traboulsee AL, et al. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2019 5;7(3):250-9.

[130] Durso-Finley J, Falet J, Nichyporuk B, Arnold D, Arbel T. Personalized prediction of future lesion activity and treatment effect in multiple

sclerosis from baseline MRI. Canadian Journal of Neurological Sciences. 2022;49(s1):S36-6.

[131] Falet JPR, Durso-Finley J, Nichyporuk B, Schroeter J, Bovis F, Sormani MP, et al. Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. Nature Communications. 2022;13(1).

[132] Lassau N, Bousaid I, Chouzenoux E, Lamarque JP, Charmettant B, Azoulay M, et al. Three artificial intelligence data challenges based on CT and MRI. Diagnostic and Interventional Imaging. 2020;101(12):783-8.

[133] Mowry EM, Bermel RA, Williams JR, S Benzinger TL, de Moor C, Fisher E, et al. Harnessing Real-World Data to Inform Decision-Making: Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATHS). Frontiers in Neurology. 2020;11(632).

[134] Kister I, Bacon T, Cutter GR. How Multiple Sclerosis Symptoms Vary by Age, Sex, and Race/Ethnicity. Neurology: Clinical Practice. 2021;11(4):335-41.

[135] Dobson R, Ramagopalan S, Davis A, Giovannoni G. Cerebrospinal fluid oligoclonal bands in multiple sclerosis and clinically isolated syndromes: a meta-analysis of prevalence, prognosis and effect of latitude. Journal of Neurology, Neurosurgery & Psychiatry. 2013 8;84(8):909 LP 914.

[136] Kurtzke JF. Rating neurologic impairment in multiple sclerosis. Neurology. 1983 11;33(11):1444 1452.

[137] Tintoré M, Rovira A, Río J, Nos C, Grivé E, Téllez N, et al. Baseline MRI predicts future attacks and disability in clinically isolated syndromes. Neurology. 2006 9;67(6):968 972.

[138] Amato MP, Portaccio E, Stromillo ML, Goretti B, Zipoli V, Siracusa G, et al. Cognitive assessment and quantitative magnetic resonance metrics can help to identify benign multiple sclerosis. Neurology. 2008;71(9):632-8.

[139] Rizzo MA, Hadjimichael OC, Preiningerova J, Vollmer TL. Prevalence and treatment of spasticity reported by multiple sclerosis patients. Multiple Sclerosis Journal. 2004 10;10(5):589-95.

[140] Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D. Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. BMC Neurology. 2013;13:37.

[141] Solà-Valls N, Vicente-Pascual M, Blanco Y, Solana E, Llufriu S, Martínez-Heras E, et al. Spanish validation of the telephone assessed Expanded Disability Status Scale and Patient Determined Disease Steps in people with multiple sclerosis. Multiple Sclerosis and Related Disorders. 2019;27:333-9.

[142] Hou Z. A Review on MR Image Intensity Inhomogeneity Correction. International Journal of Biomedical Imaging. 2006;2006:49515.

[143] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Transactions on Medical Imaging. 2010;29(6):1310-20.

[144] Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping. 2019;40(17):4952-64.

[145] Valverde S, Coll L, Valencia L, Clèrigues A, Oliver A, Vilanova JC, et al. Assessing the Accuracy and Reproducibility of PARIETAL: A Deep Learning Brain Extraction Algorithm. Journal of Magnetic Resonance Imaging. 2021.

[146] Jenkinson M, Pechaud M, Smith S. BET2-MR-Based Estimation of Brain, Skull and Scalp Surfaces. Human Brain Mapping. 2002;17(2):143-55.

[147] Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. NeuroImage. 2012 8;62(2):782-90.

[148] Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, et al. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical. 2014;6:9-19.

[149] Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Scientific Reports. 2020;10(1):1-15.

[150] Danelakis A, Theoharis T, Verganelakis DA. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. Computerized Medical Imaging and Graphics. 2018;70:83-100.

[151] Schmidt P, Gaser C, Arsic M, Buck D, Förschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. NeuroImage. 2012 2;59(4):3774-83.

[152] Battaglini M, Jenkinson M, De Stefano N. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Human Brain Mapping. 2012 9;33(9):2062-71.

[153] Prados F, Cardoso MJ, Kanber B, Ciccarelli O, Kapoor R, Gandini Wheeler-Kingshott CAM, et al. A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. NeuroImage. 2016 10;139:376-84.

[154] Popescu V, Agosta F, Hulst HE, Sluimer IC, Knol DL, Sormani MP, et al. Brain atrophy and lesion load predict long term disability in multiple sclerosis. Journal of Neurology, Neurosurgery & Psychiatry. 2013 10;84(10):1082 1091.

[155] Dorent R, Booth T, Li W, Sudre CH, Kafiabadi S, Cardoso J, et al. Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. Medical Image Analysis. 2021;67.

[156] Pinaya WHL, Tudosiu PD, Gray R, Rees G, Nachev P, Ourselin S, et al. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Medical Image Analysis. 2022;79:102475.

[157] Dora L, Agrawal S, Panda R, Abraham A. State-of-the-Art Methods for Brain Tissue Segmentation: A Review. IEEE Reviews in Biomedical Engineering. 2017;10:235-49.

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

[159] González-Villà S, Oliver A, Valverde S, Wang L, Zwiggelaar R, Lladó X. A review on brain structures segmentation in magnetic resonance imaging. Artificial Intelligence in Medicine. 2016;73:45-69.

[160] Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. NeuroImage. 2020 10;219:117012.

[161] Fischl B. FreeSurfer. NeuroImage. 2012;62(2):774-81.

[162] Sarraf S, Desouza DD, Anderson JAE, Saverino C. MCADNNet: Recognizing stages of cognitive impairment through efficient convolutional fMRI and MRI neural network topology models. IEEE Access. 2019;7(Mci):155584-600.

[163] Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nature Communications. 2018;9(1):4273.

[164] De Meo E, Portaccio E, Giorgio A, Ruano L, Goretti B, Niccolai C, et al. Identifying the Distinct Cognitive Phenotypes in Multiple Sclerosis. JAMA Neurology. 2021 4;78(4):414-25.

[165] Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.

[166] Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Frontiers in Aging Neuroscience. 2019;10(JUL).

[167] Binder A, Montavon G, Lapuschkin S, Müller KR, Samek W. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In: Artificial Neural Networks and Machine Learning – ICANN 2016; 2016. p. 63-71.

[168] Lungu O, Pantano P, Kumfor F, Gallo A, Joel Shaw D, Cek MR, et al. Impaired Self-Other Distinction and Subcortical Gray-Matter Alterations Characterize Socio-Cognitive Disturbances in Multiple Sclerosis. Front Neurol. 2019;10:525.

[169] Bove R, Poole S, Cuneo R, Gupta S, Sabatino J, Harms M, et al. Remote Observational Research for Multiple Sclerosis. Neurology - Neuroimmunology Neuroinflammation. 2023 3;10(2):e200070.

[170] Signori A, Lorscheider J, Vukusic S, Trojano M, Iaffaldano P, Hillert J, et al. Heterogeneity on long-term disability trajectories in patients with secondary progressive MS: a latent class analysis from Big MS Data network. Journal of Neurology, Neurosurgery & Psychiatry. 2022 9:2022-329987.

[171] Tur C, Moccia M, Barkhof F, Chataway J, Sastre-Garriga J, Thompson AJ, et al. Assessing treatment outcomes in multiple sclerosis trials and in the clinical setting. Nature Reviews Neurology. 2018;14(2):75-93.

[172] Lopatina A, Ropele S, Sibgatulin R, Reichenbach JR, Güllmar D. Investigation of Deep-Learning-Driven Identification of Multiple Sclerosis Patients Based on Susceptibility-Weighted Images Using Relevance Analysis. Frontiers in Neuroscience. 2020;14.

[173] Gilmore CP, Donaldson I, Bö L, Owens T, Lowe J, Evangelou N. Regional variations in the extent and pattern of grey matter demyelination in multiple sclerosis: a comparison between the cerebral cortex, cerebellar cortex, deep grey matter nuclei and the spinal cord. Journal of Neurology, Neurosurgery & Psychiatry. 2009 2;80(2):182 LP 187.

[174] Tur C, Eshaghi A, Altmann DR, Jenkins TM, Prados F, Grussu F, et al. Structural cortical network reorganization associated with early conversion to multiple sclerosis. Scientific Reports. 2018;8(1):10715.

[175] Collorone S, Prados F, Kanber B, Cawley NM, Tur C, Grussu F, et al. Brain microstructural and metabolic alterations detected in vivo at onset of the first demyelinating event. Brain. 2021;144:1409-21.

[176] Cordano C, Nourbakhsh B, Yiu HH, Papinutto N, Caverzasi E, Abdelhak AC, et al. Differences in Age-Related Retinal and Cortical Atrophy Rates in Multiple Sclerosis. Neurology. 2022 8;15(99):e1685-93.

[177] Madsen MAJ, Wiggermann V, Marques MFM, Lundell H, Cerri S, Puonti O, et al. Linking lesions in sensorimotor cortex to contralateral hand function in multiple sclerosis: a 7 T MRI study. Brain. 2022 6;10(145):3522-35.

[178] Tintore M, Rovira A, Arrambide G, Mitjana R, Río J, Auger C, et al. Brainstem lesions in clinically isolated syndromes. Neurology. 2010 11;75(21):1933 1938.

[179] Savini G, Pardini M, Castellazzi G, Lascialfari A, Chard D, D'Angelo E, et al. Default Mode Network Structural Integrity and Cerebellar Connectivity Predict Information Processing Speed Deficit in Multiple Sclerosis. Frontiers in Cellular Neuroscience. 2019;13.

[180] Brown JWL, Pardini M, Brownlee WJ, Fernando K, Samson RS, Prados Carrasco F, et al. An abnormal periventricular magnetization transfer ratio gradient occurs early in multiple sclerosis. Brain. 2017 2;140(2):387-98.

[181] Eshaghi A, Marinescu RV, Young AL, Firth NC, Prados F, Cardoso MJ, et al. Progression of regional grey matter atrophy in multiple sclerosis on behalf of the MAGNIMS study group*. Brain. 2018;141:1665-77.

[182] Tur C, Grussu F, De Angelis F, Prados F, Kanber B, Calvi A, et al. Spatial patterns of brain lesions assessed through covariance estimations of lesional voxels in multiple Sclerosis: The SPACE-MS technique. NeuroImage: Clinical. 2022 1;33:102904.

[183] Cappelle S, Pareto D, Vidal-Jordana A, Alyafeai R, Alberich M, Sastre-Garriga J, et al. A validation study of manual atrophy measures in patients with Multiple Sclerosis. Neuroradiology. 2020;62(8):955-64.

[184] Frizzell TO, Glashutter M, Liu CC, Zeng A, Pan D, Hajra SG, et al. Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: A systematic review. Ageing Research Reviews. 2022;77:101614.

[185] Ahmed S, Kim BC, Lee KH, Yub H, Jung HY. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. PLoS ONE. 2020;15(12).

[186] Kwak K, Niethammer M, Giovanello KS, Styner M, Dayan E. Differential Role for Hippocampal Subfields in Alzheimer's Disease Progression Revealed with Deep Learning. Cerebral Cortex. 2021;32(3):467-78.

[187] Basheera S, Ram MSS. Deep learning based Alzheimer's disease early diagnosis using T2w segmented gray matter MRI. International Journal of Imaging Systems and Technology. 2021;31(3):1692-710.

[188] Mehmood A, Yang S, Feng Z, Wang M, Ahmad AS, Khan R, et al. A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images. Neuroscience. 2021;460:43-52.

[189] Zhou P, Jiang S, Yu L, Feng Y, Chen C, Li F, et al. Use of a Sparse-Response Deep Belief Network and Extreme Learning Machine to Discriminate Alzheimer's Disease, Mild Cognitive Impairment, and Normal Controls Based on Amyloid PET/MRI Images. Frontiers in Medicine. 2021;7:621204.

[190] Zhu W, Sun L, Huang J, Han L, Zhang D. Dual Attention Multi-Instance Deep Learning for Alzheimer's Disease Diagnosis with Structural MRI. IEEE Transactions on Medical Imaging. 2021;40(9):2354-66.

[191] Haider L, Chung K, Birch G, Eshaghi A, Mangesius S, Prados F, et al. Linear brain atrophy measures in multiple sclerosis and clinically isolated syndromes: a 30-year follow-up. Journal of Neurology, Neurosurgery & Psychiatry. 2021 8;92(8):839 846.

[192] Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Medicine. 2020;3:136.

[193] Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Processing Letters. 2014;21(11):1389-93.

[194] Amiri H, de Sitter A, Bendfeldt K, Battaglini M, Gandini Wheeler-Kingshott CAM, Calabrese M, et al. Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI. NeuroImage: Clinical. 2018 1;19:466-75.

[195] AlTokhis AI, AlAmrani A, Alotaibi A, Podlasek A, Constantinescu CS. Magnetic Resonance Imaging as a Prognostic Disability Marker in Clinically Isolated Syndrome and Multiple Sclerosis: A Systematic Review and Meta-Analysis. Diagnostics. 2022 1;12(2):270.

[196] Eshaghi A, Prados F, Brownlee WJ, Altmann DR, Tur C, Cardoso MJ, et al. Deep Gray Matter Volume Loss Drives Disability Worsening in Multiple Sclerosis. Ann Neurol. 2018;83(2):210-22.

[197] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958 6;53(282):457-81.

[198] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187-220.

[199] Cree BAC, Hollenbach JA, Bove R, Kirkish G, Sacco S, Caverzasi E, et al. Silent progression in disease activity–free relapsing multiple sclerosis. Annals of Neurology. 2019 5;85(5):653-66.

[200] Bischof A, Papinutto N, Keshavan A, Rajesh A, Kirkish G, Zhang X, et al. Spinal Cord Atrophy Predicts Progressive Disease in Relapsing Multiple Sclerosis. Annals of Neurology. 2022 2;91(2):268-81.

[201] Faraggi D, Simon R. A neural network model for survival data. Statistics in Medicine. 1995 1;14(1):73-82.

[202] Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1):1-12.

[203] Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ. 2019;2019(1):1-19.

[204] Haarburger C, Weitz P, Rippel O, Merhof D. Image-based survival prediction for lung cancer patients using CNNS. Proceedings - International Symposium on Biomedical Imaging. 2019;2019-April:1197-201.

[205] Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Scientific Reports. 2021;11(1):1-12.

[206] Matsumoto T, Walston SL, Walston M, Kabata D, Miki Y, Shiba M, et al. Deep Learning–Based Time-to-Death Prediction Model for COVID-19 Patients Using Clinical Data and Chest Radiographs. Journal of Digital Imaging. 2022.

[207] Mirabnahrazam G, Ma D, Beaulac C, Lee S, Popuri K, Lee H, et al. Predicting time-to-conversion for dementia of Alzheimer's type using multi-modal deep survival analysis. Neurobiology of Aging. 2023 1;121:139-56.

[208] Uemura T, Näppi JJ, Watari C, Hironaka T, Kamiya T, Yoshida H. Weakly unsupervised conditional generative adversarial network for image-based prognostic prediction for COVID-19 patients based on chest CT. Medical Image Analysis. 2021;73:102159.

[209] Xu X, Prasanna P. Brain Cancer Survival Prediction on Treatment-Naïve MRI using Deep Anchor Attention Learning with Vision Transformer. Proceedings - International Symposium on Biomedical Imaging. 2022;2022-March.

[210] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018:4510-20.

[211] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Conference on Computer Vision and Pattern Recognition; 2018. p. 7132-41.

[212] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. 2015;115(3):211-52.

[213] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. 2017:1-16.

[214] Harrell Jr FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982 5;247(18):2543-6.

[215] Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. Statistics in Medicine. 2005 12;24(24):3927-44.

[216] Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal. 2006 12;48(6):1029-40.

[217] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010 1;21(1):128-38.

[218] Eder M, Moser E, Holzinger A, Jean-Quartier C, Jeanquartier F. Interpretable Machine Learning with Brain Image and Survival Data. BioMedInformatics. 2022;2(3):492-510.

[219] Ajitomi S, Fujimori J, Nakashima I. Usefulness of two-dimensional measurements for the evaluation of brain volume and disability in multiple sclerosis. Multiple Sclerosis Journal - Experimental, Translational and Clinical. 2022;8(1).

[220] Zhang Y, Cofield S, Cutter G, Krieger S, Wolinsky JS, Lublin F. Predictors of Disease Activity and Worsening in Relapsing-Remitting Multiple Sclerosis. Neurology: Clinical Practice. 2022;12(4):E58-65.

Appendix A

Publications

Journals

- **Llucia Coll**, D. Pareto, P. Carbonell-Mirabent, Á. Cobo-Calvo, G. Arrambide, A. Vidal-Jordana, M. Comabella, J. Castilló, B. Rodríguez-Acebedo, A. Zabalza, I. Galán, L. Midaglia, C. Nos, A. Salerno, C. Auger, M. Alberich, J. Río, J. Sastre-Garriga, A. Oliver, X. Montalban, À. Rovira, M. Tintoré, X. Lladó, C. Tur. Deciphering multiple sclerosis disability with deep learning attention maps on clinical MRI. *NeuroImage: Clinical*, 38, art 103376, 2023. [JCR N IF 4.891, Q2(5/14)] [33]

- **Llucia Coll**, D. Pareto, P. Carbonell-Mirabent, Á. Cobo-Calvo, G. Arrambide, A. Vidal-Jordana, M. Comabella, J. Castilló, B. Rodríguez-Acebedo, A. Zabalza, I. Galán, L. Midaglia, C. Nos, C. Auger, M. Alberich, J. Río, J. Sastre-Garriga, A. Oliver, X. Montalban, À. Rovira, M. Tintoré, X. Lladó, C. Tur. Global and regional deep learning models for multiple sclerosis stratification from MRI. *Journal of Magnetic Resonance Imaging* [JCR N IF 5.119, Q1(34/136)] [34]

- **Llucia Coll**, D. Pareto, P. Carbonell-Mirabent, Á. Cobo-Calvo, G. Arrambide, A. Vidal-Jordana, M. Comabella, J. Castilló, B. Rodríguez-Acebedo, A. Zabalza, I. Galán, L. Midaglia, C. Nos, C. Auger, M. Alberich, J. Río, J. Sastre-Garriga, A. Oliver, X. Montalban, À. Rovira, M. Tintoré, X. Lladó, C. Tur. Deep learning to predict progression

independent of relapse activity at first demyelinating event. (Under revision)

- Sara Collorone, **Ll. Coll**, M. Lorenzi, X. Lladó, J. Sastre-Garriga, M. Tintoré, X. Montalban, À. Rovira, D. Pareto, C. Tur. Artificial intelligence applied to MRI data to tackle key challenges in multiple sclerosis. (Under revision at *Multiple Sclerosis Journal*).

Conferences

- **Llucia Coll**, P. Carbonell, G. Arrambide, A. Vidal-Jordana, L. Midaglia, A. Cobo-Calvo, M. Comabella, J. Castilló, B. Rodrñiguezm A. Zabalza, I. Gallán, C. Nos, C. Auger, J. Rio, J. Sastre-Garriga, X. Montalban, M. Tintoré, Àlex Rovira, X. Lladó, D. Pareto, C. Tur. Spatial features of brain demyelinating lesions as prognostic factors in the clinically isolated syndrome. *ECTRIMS 2021 Multiple Sclerosis*, 13-15th October, 2021, Vienna, Austria. [JCR CN IF:6.312 Q1(28/208)]

- **Llucia Coll**, P. Carbonell, G. Arrambide, A. Vidal-Jordana, L. Midaglia, A. Cobo-Calvo, M. Comabella, J. Castilló, B. Rodrñiguezm A. Zabalza, I. Gallán, C. Nos, C. Auger, J. Rio, J. Sastre-Garriga, X. Montalban, M. Tintoré, Àlex Rovira, X. Lladó, D. Pareto, C. Tur. Características espaciales de las lesiones desmielinizantes cerebrales como factores pronóstico en el síndrome clínico aislado. *LXXIII Conference Sociedad Española de Neurología (SEN)*, November 2021, Digital meeting.

- **Llucia Coll**, P. Carbonell-Mirabent, Á. Cobo-Calvo, G. Arrambide, Á. Vidal-Jordana, M. Comabella, J. Castilló, B. Rodríguez-Acevedo, A. Zabalza, I. Galán, L. Midaglia, C. Nos, A. Salerno, C. Auger, M. Alberich, J. Río, J. Sastre-Garriga, A. Oliver, X. Montalban, À. Rovira, M. Tintoré, D. Pareto, X. Lladó, C. Tur. Classification of MS patients into disability stages using deep learning approaches based solely on routinely-acquired MRI. *ISMRM 31th International Society for Magnetic Resonance in Medicine*, 07-12th May 2022 London, UK.

- **Llucia Coll**, P. Carbonell-Mirabent, Á. Cobo-Calvo, G. Arrambide, Á. Vidal-Jordana, M. Comabella, J. Castilló, B. Rodríguez-Acevedo, A. Zabalza, I. Galán, L. Midaglia, C. Nos, A. Salerno, C. Auger, M.

Alberich, J. Río, J. Sastre-Garriga, A. Oliver, X. Montalban, À. Rovira, M. Tintoré, D. Pareto, X. Lladó, C. Tur. Convolutional neural networks on routinely-acquired MRI to strategy MS patients according to their disability level and understand anatomical regions involved in clinical progression. *ECTRIMS 2022 Multiple Sclerosis*. Amsterdam, 26-28th October 2022. [JCR CN IF:6.312 Q1(28/208)]

- Ana Harris, **Ll. Coll**, P. Carbonell-Mirabent, G. Arrambide, Á. Vidal-Jordana, L. Midaglia, Á. Cobo-Calvo, M. Comabella, J. Castilló, B. Rodríguez-Acevedo, A. Zabalza, I. Galán, C. Nos, A. Salerno, C. Auger, J. Río, M. Alberich, J. Sastre-Garriga, X. Montalban, M. Tintoré, À. Rovira, D. Pareto, C. Tur. Temporal dynamics of spatial distributional features of brain white matter lesions and concurrent clinical changes in relapsing MS. *ECTRIMS 2022 Multiple Sclerosis*, Amsterdam, 26-28th October 2022. [JCR CN IF:6.312 Q1(28/208)]

- **Llucia Coll**, P. Carbonell-Mirabent, Á. Cobo-Calvo, G. Arrambide, Á. Vidal-Jordana, M. Comabella, J. Castilló, B. Rodríguez-Acevedo, A. Zabalza, I. Galán, L. Midaglia, C. Nos, A. Salerno, C. Auger, M. Alberich, J. Río, J. Sastre-Garriga, A. Oliver, X. Montalban, À. Rovira, M. Tintoré, D. Pareto, X. Lladó, C. Tur. Prediction of progression independent of relapse activity at the first demyelinating attack with a deep learning image-based survival model. *ECTRIMS 2023 Multiple Sclerosis*, 11-13th October 2023, Milan, Italy. [JCR CN IF:5.800 Q1(31/212)]