Deep Learning Methods for the Segmentation of White Matter Hyperintesities

Robust automatic segmentation for medical imaging analysis

Alice Schiavone

Supervised at University of Copenhagen by
Prof. Mads Nielsen
Asst. Prof. Mostafa Mehdipour Ghazi

Supervised at Università degli Studi di Milano by
Prof. Vincenzo Piuri

Faculty of Computer Science
Università degli Studi di Milano

October, 2023

Student ID: 986325

To Christian

I hope that I will inspire you to do great things.

Acknowledgements

I would like to express my deepest gratitude to Prof. Vincenzo Piuri, my supervisor at the University of Milan, who put time and effort into making this project possible in the beginning, and in finalizing its completion.

I'm extremely grateful to Prof. Mads Nielsen, Asst. Prof. Mostafa Mehdipour Ghazi and Sebastian Nørgaard Llambias, my supervisors at the University of Copenhagen, who allowed me to start my journey in the field of medical imaging in the beautiful country of Denmark as an expat. You welcomed me into your research group with no hesitation and provided me with the tools and expertise required to make this thesis happen, and so much more.

I am deeply indebted to Dr. Akshay Pai, Jacob Johansen, and Dr. Silvia Ingala from Cerebriu. You didn't only provide supervision for my project, but all of you gave me, in different ways, a new way to look at the world and powerful insights to help me reach my life goals. My gratitude can also be extended to all my colleagues from Cerebriu, with whom I built memories that I will always treasure.

Big thanks should also go to my family, who supported me since the start of my education to see me succeed today: thank you to Nikolas, mom, dad, Simone, and Bea. I am also thankful to my extended family, including my grandparents, my aunties and uncles, and my cousins.

Lastly, my academic journey would not have been the same without Davide D.A., Davide R. and Simon. Thank you for the fun and support along the way.

Abstract

White matter hyperintesities (WMH) are associated with an increased risk of stroke, cognitive decline, and dementia. A robust, yet accurate detection of WMH can help with the prevention of more lesions from forming. The task is still challenging as the lesions are often small and irregular. In this thesis, a robust deep learning-based method for the automatic segmentation of WMH is proposed, only using fluid attenuated inversion recovery (FLAIR) scans and MRI-specific data augmentation and comparing it with state-of-the-art methods. Different methods have been tested on public and private data, and we can show that one of these models is more robust to domain shift and achieves higher segmentation accuracy than the alternatives.

Ethical Considerations

This thesis includes medical images and data that have been collected and used with utmost consideration for patient privacy and confidentiality. Throughout the whole reaserch process, we have only worked with anonymized images, meaning the images did not have any additional data about patients names, dates of birth or other personal identifiers. We are aware that it is possible to derive a person portrait from their medical images, however no such personal information is made public by being published in this manuscript. All the images displayed in this work have been acquired either through other research studies, or through public datasets. We also picture some cases from a in-house dataset, however, no personal identification is possible through a single image slice, and no further information about patient is included. We are grateful to the patients and healthcare institutions who participated in this study and kindly provided their consent for the use of their medical data, either through studies we referenced or data that we worked with first hand.

Contents

List of Figures

List of Tables

List of Abbreviations

WMH white matter hyperintesities vii
FLAIR fluid attenuated inversion recovery vii
MS Multiple Sclerosis x
DSC Sørensen–Dice Similarity Coefficient xiii
VS Volumetric Similarity xiii
HDD Hausdorff Distance xiii
LF Low (Magnetic) Field MRI xv
MRI medical resonance imaging 1
CSVD cerebral small vessel disease 5
DICOM Digital Imaging and Communications in Medicine 15
NII NifTI-1 data format 15
STRIVE STandards for ReportIng Vascular changes on nEuroimaging 25
FCNE Fully Convolutional Network Ensembles 56
OA other abnormality 84

1

Introduction

1.1 | Motivation

White matter hyperintesities (WMH) have been long studied as a biomarker for a variety of cognitive impairment conditions, like dementia and Alzheimer's, and their presence is associated with an increased risk of stroke, disability, and mortality.

Clinicians rely on medical imaging analysis to assess the presence and importance of pathologies. In neuroimaging, different brain tissues are shown as having different voxel intensities. Visually, a magnetic resonance image is a simple grey-scale image. The lack of color or other nuance leaves it up to the interpretation of a medical professional to distinguish the different pathologies that might occur in a patient. To get more precise information about a particular tissue, we can use different imaging techniques. White matter hyperintesities appear white in fluid attenuated inversion recovery (FLAIR), a medical resonance imaging (MRI) techniques, as do other conditions like gliosis, edema, and stroke. In a FLAIR image, white matter hyperintesities are easier to notice, due to the suppression of the signal of cerebrospinal fluid, which also appears bright in other sequences. While these conditions might look similar in color and intensity, they are extremely different in position and shape, but most importantly in the reason of how they came to be, and in how they affect a patient. We focus our attention on age-related white matter hyperintesities, lesions that occur in otherwise healthy subjects as they age.

Many clinical studies rely on the quantification of the number and the volume of these lesions, through different techniques: some rely on a categorical scale that indicates how extensive are the lesions, and other methods identify the volume of the diseased tissue by producing a segmentation map. For the purpose of this work, we define segmentation as the division between healthy and not healthy tissue, performed on an

image acquired through medical imaging techniques. Being able to accurately segment white matter hyperintesities helps researchers in studying their nature and causes. Observing biomarkers such as white matter hyperintensities can help in aiding studies of diseases where they are found to be co-occurring. Moreover, the presence of white matter hyperintensities can interfere where they are not the main focus of a given study, as they can often be mistaken as other pathologies. Consequently, in some situations, we could benefit from their exclusion.

1.2 | Aims and Objectives

Manual annotation of medical images is expensive and long, that can only be carried out by a expert radiologist. We developed a method for the automatic segmentation of white matter hyperintesities with deep learning methods. The segmented lesions would have been used for the neurological assessment of a large cohort of patients, in the order of thousands, coming from many Danish hospitals. In particular, the initial research hypothesis was about the changes in the volume of white matter hyperintesities in Covid-19 patients, before, after, and during the course of the infection. Unfortunately, the data couldn't be delivered in time for the purpose of this thesis, forcing us to change the aim of this work. As the statistical correlations between the patients and the volume of WMH would have been only the last validation step of the process, we are still concerned with the implementation of an automatic segmentation method that is able to generalize on such a large cohort. Therefore, we focus our efforts on the development of this method, excluding the initial motivation but keeping it as the foundation for the method's performance final evaluation. We can formally define our research question as:

> ***Research hypothesis:*** *we can develop a robust automatic method for the segmentation of white matter hyperintesities in magnetic resonance images, that is able to generalize on images coming from a variety of hospitals and patients.*

In medical imaging, and more in general, the task of segmentation aims at assigning to a specific class each point of a sample (a pixel in two dimensions, a voxel in three dimensions). Our aim is to design a segmentation method that is robust to domain changes, which is the main challenge in a medical imaging setting. In recent years, the state-of-the-art techniques in segmentation tasks are machine learning methods, and in particular deep learning. Encouraged by the promising results in similar problems, we also propose a deep learning method. However, these methods only perform nicely

when many images are available. Due to the limited access to images and the expensive process of annotation, any machine learning method on this type of task suffers from a lack of data. We show that specific techniques can improve performance without acquiring more data.

1.3 | Proposed Solution

In order to solve this task, we work with machine learning methods, and in particular deep learning, to segment automatically white matter hyperintensities in brain magnetic resonance imaging. We trained models on images from a public dataset published by a medical imaging challenge. We evaluate these models on the testing dataset released by the same challenge, and on an in-house dataset, that is richer in other pathologies segmentations, but not in white matter hyperintesities: this setup is a nice example of a out-of-distribution setting because the data from challenges are often well curated and don't display problems that arise in the real world. The method that won the challenge is still the state-of-the-art method, and we will compare our experiments' performance to the former.

Specifically, our solution is based on U-Net architecture, trained with the NnUNet framework on the MICCAI 2017 WMH Segmentation Challenge dataset, with the help of MRI specific data augmentations. We test our method on the testing set from said challenge and on additional private data. Previous methods rely on the presence of more information, such as T1-weighted, FLAIR, or more sequences, but we reach a par performance training on FLAIR only.

The goal of this work is to address the challenges of WMH segmentation and to develop a method that performs well even on out-of-distribution data. The evaluation of results can also be problematic because the nature of the ground truth annotation is in itself affected by variance and disagreement between medical professionals. For this reason, qualitative metrics are as important as quantitative ones, and the discussion of results is as important as the results themselves.

1.4 | Document Structure

This thesis is divided into five main chapters. The first chapter, the *Introduction*, has already been presented.

In Chapter 2, titled *Background & Literature Overview*, we explore five different sections. Each section covers the essential knowledge needed to grasp the problem at hand

and the reasons behind our proposed solution. We will discuss white matter hyperintensities, their current identification methods, and the rationale for automating part of the medical imaging analysis. Given that our solution involves deep learning methods, we will explain how this method evolved from traditional machine learning techniques and how it has become a prominent approach in various fields, including medical imaging. Throughout this chapter, we go through previous work by referencing commonly adopted standards, including research on other automated segmentation techniques.

In Chapter 3, named *Materials & Methods*, we will outline the methodology used to conduct our experiments. This chapter includes information about the developed learning methods and the differences between them.

In Chapter 4, labeled *Results & Discussion*, we will present the findings obtained from our experiments and engage in a thorough discussion and analysis of these results. The presentation of results will include standard evaluation metrics utilized in the field, as well as a qualitative visual examination of the method's performance on normal and edge test cases.

In Chapter 5, titled *Conclusions*, we will provide a comprehensive summary of our work and the results we obtained. Finally, we will explore potential future directions to extend this research.

The reader will find *references* to the cited authors and their work at the end of this manuscript, along with additional appendixes.

A small section of the results reported in this thesis have been submitted and accepted as a short paper [19] at the *Medical Imaging with Deep Learning* (MIDL) conference, which took place in Nashville (USA) from July 10th to July 12th, 2023. The paper is provided at the end of this thesis.

2

Background & Literature Overview

In this chapter, we will go over all the information needed to understand the task at hand and the methods that have been developed to solve it. We will introduce the motivation for the segmentation of white matter hyperintesities, with all the clinical implications. We will explain how brain information is extracted through magnetic resonance imaging. Lastly, we will go over different techniques to achieve an accurate automatic segmentation of the lesions in question, and in particular, we will focus on machine learning and related methods.

2.1 | Brain Anatomy and Pathology

While the gray matter is commonly recognized as responsible for computational power and information storage in the human body, white matter has historically been considered less significant. Nonetheless, in the human brain, white matter occupies a much larger proportion of space compared to the brains of other animals. [20]

To communicate, brain cells extend in a cable-like structure called the axon, which transports electrical signals throughout the brain. To do it more efficiently, the axon is coated in myelin, a fatty material that insulates the axon and increases its signaling rate.

White matter is key in transporting information from and to different regions of the brain, but this was its only suspected function for decades. We know today that white matter is critical in mastering a variety of skills, and it changes throughout the life of an individual.

Cerebral small vessel disease (CSVD) is a condition that affects small arteries and capillaries of the brain and refers to several pathologies. Neuroimaging features of CSVD include, among others, white matter hyperintesities. The segmentation and quantification of white matter hyperintesities is the main focus of this work.

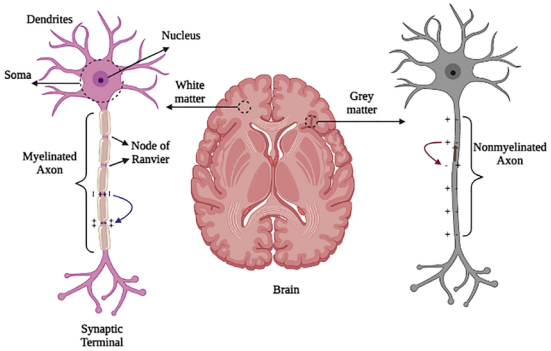


Figure 2.1: Schematic structure of two different neurons, one (a) that is myelinated and belongs to white matter tissue, and the other (b) that is nonmyelinated, which can be found in gray matter. The insulation of the axon of white matter cells makes it much more efficient in sending and receiving signals for long distances, even in the nervous system. [1]

Several pathologies can affect the human brain. [21] We can divide them into three main groups:

- Diseases that damage brain structures. These can be due to traumatic injuries, or due to internal tissue disruption such as cerebrovascular disease, infections, and tumors. This group includes also communication and sensory disorders.
- Neurodegenerative diseases and mental disorders that raise from functional brain disorders with detectable destruction of brain communication networks. Among neurodegenerative diseases we have Parkinson's disease, Alzheimer's disease, and other types of dementia; mental disorders included in this category are schizophrenia, depression, bipolar disorder, alcoholism, and drug abuse.
- Other brain disorders, like migraines and sleep disorders, that cannot be detectable by structural or functional impairment

While the latter group does not affect our work, as there is no visual effect on the images we analyze, we are interested in the first two: age-related white matter hyperintesities are a biomarker for the presence and development of neurodegenerative diseases, but a patient can be affected by one or more brain abnormalities at the same time. An old patient may develop asymptomatic age-related white matter hyperintesities but may be screened for other disruptive disorders, like tumors or strokes. The pathologies may appear in the same subjects in the same image, so focusing on the automatic segmentation of a single pathology may be hindered by the presence of another. We

will show how our work is affected by these other abnormalities when we look at the segmentation results.

2.1.1 | White Matter Hyperintesities

White matter hyperintesities of presumed vascular origin are the most studied feature of cerebral small vessel disease (CSVD). [22]

We are concerned with age-related white matter hyperintesities, which appear frequently in CT and MR scans of older individuals. They are a heterogeneous process, meaning that they may appear commonly as a semi-normal finding or can be pathological, as they have been associated with cognitive impairment and other body dysfunctions. [23] While it is commonly reported that, in healthy populations, WMH generally increase with age and time, it is not clear why they do so. It is also reported that, sometimes, they may decrease in size. [22]

WMH are a common finding in lacunar stroke and all strokes, while predicting a worse prognosis after stroke. Their presence can aid in clinical decision-making. [24] WMH are associated with an increased risk of dementia, cognitive and functional impairment, an increased insurgence of recurrent ischemic and non-ischemic stroke, and an increase in cardiovascular and all-cause mortality. [25] They have also been found to affect all main cognitive domains, including memory. The worst clinical outcomes are when the lesions are not only extensive but also confluent.

White matter hyperintesities appear bright in T2-weighted MRI sequences, and in particular on FLAIR. They appear in the white matter tissue and are typically symmetrical between hemispheres. [22] The vast majority of studies on WMH uses neuroimaging to quantify them, typically on computed tomography (CT) and magnetic resonance imaging (MRI). Various rating systems and scores have been used throughout the years, resulting in diversity in results when evaluating changes in WMH. The main disadvantage of using a rating scale is their subjective interpretation, which increases intra-rater variability. [26]

As the white matter is the messenger of the brain, WMHs cause cognitive decline, especially in information processing speed. This affects executive function and may lead to dementia. Althought a progression in WMHs has shown a decline in cognitive skills, there is little evidence that a reduction in their progression will prevent further decline. [27]

2.1.2 | Assessment of White Matter Hyperintesities

White matter hyperintese signal was scored by Fakezas [28] in order to assess the extent of the lesions. (Table 2.1) A focus (foci, plural) is the area were the disease develops. We call bridging when thin lesions connect other lesions. This scoring method is actively used in clinical practice and research studies. Segmentations of white matter hyperintesities are not, on the contrary, very common: this is due to the intense and expensive annotation process, done by radiologists. It has often been deemed enough to score a patient based on a simple scale. However, if we want to study white matter hyperintesities as a biomarker for other diseases, the Fakezas scale is not enough.

Fazekas (PWMH)	
0	None or a single WMH lesion
1	Caps or pencil-thin lining
2	Smooth halo
3	Irregular periventricular signal extending into the deep white matter
Fazekas (DWMH)	
0	None or a single WMH lesion
1	Multiple punctate foci
2	Beginning confluence of foci (bridging)
3	Large confluent areas

Table 2.1: Fazekas scale for the assessment of WMH.

First, the scale has been adapted over the years by different researchers, based on the different tasks it was needed for. We can divide the white matter into more regions, depending on spatiality or brain regions. [28] proposed a first division based on deep and periventricular white matter, that can be differentiated based on the vicinity to brain ventricles. We define as periventricular white matter hyperintesities (PWMH) the lesions that appear within a 10mm margin from the brain ventricles. The remaining lesions will be considered as deep white matter hyperintesities (DWMH). The Fazekas score can be assigned for each of these regions of the brain.

Other Fazekas scale variations are based on six or more brain regions, based on a standard brain atlas, or give more complex ratings. There is not an agreed standard on which scoring method is best to use. Additionally, any visual scoring is highly influenced by graders' opinions: two radiologists might be in disagreement when one case cannot clearly be categorized in one class or the other. As the method is already imprecise, intra-grader disagreement greatly affects the scoring process. [29] argues that there exists a very good correlation between many of these rating scales, but still advocates

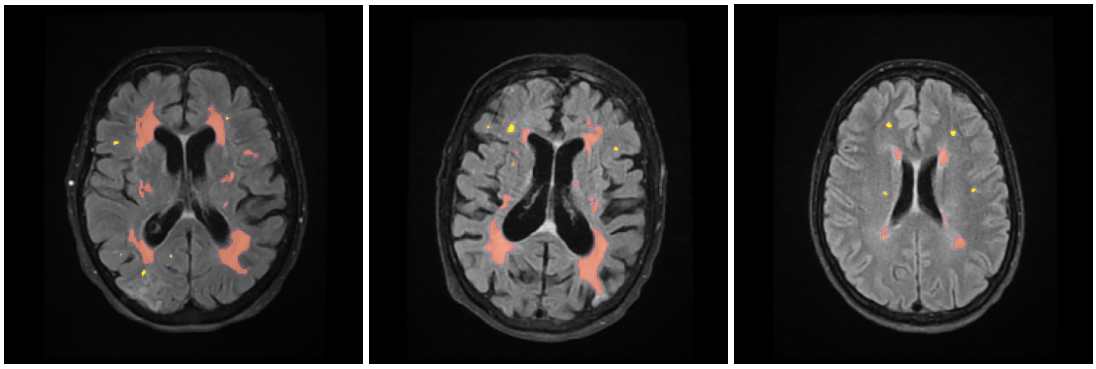


Figure 2.2: Deep (yellow) and periventricular (red) white matter hyperintesities, segmented by our proposed method. We define as periventricular the lesions that fall into a 10mm margin from within the ventricles and deep all the others. There are other types of white matter lesions based on location, but they will not be mentioned as they do not concern age-related WMH.

for a computerized measurement of white matter hyperintesities.

A volume-based metric of assessment for the presence of white matter hyperintensities is the best possible scenario. We aim at developing one that can be robust to multiple data distributions so that it can be a first stepping stone in other research projects that use white matter lesions as a biomarker for other diseases.

2.2 | Medical Imaging

Medical imaging refers to the use of various imaging techniques to visualize the internal structures and functions of the human body for diagnostic and therapeutic purposes. These images provide valuable information to healthcare professionals to aid in the diagnosis, treatment planning, and monitoring of various medical conditions. Images can also be used for image analysis and interpretation by medical researchers. An image can be produced through different mediums, like computed tomography (CT), ultrasound, or X-rays, but we will focus on magnetic resonance imaging (MRI). Each imaging technique has its own strengths and weaknesses, but the choice of preferring one or the other can also come from a cost optimization point of view. Healthcare providers may choose to rely on a cheaper and faster technique if a doctor deems unnecessary the use of more precise, but more expensive, techniques. Among the challenges that we could face by working with data acquired through medical imaging, we can find:

- Data quality: medical images can be affected by various factors such as low resolution, high noise levels, low contrast, and imaging artifacts. Diagnosis, but also accuracy and reliability of analysis techniques, can be affected by these imperfections.
- Data size and management: the size of medical image datasets poses difficulties in terms of storage capacity, data transfer speeds, and computational requirements.
- Data availability: while medical imaging generates large volumes of data every day, most of it is not publicly available. This hinders research efforts, but it prevents the leak and misuse of private medical information. Most publicly available datasets are made of only a few dozen cases, but thousands of images are acquired every day.
- Image noise: images are affected by the signal-to-noise ratio (SNR), which measures how much of the signal information is present with respect to noise that appears due to external factors.
- Image interpretation: different radiologists may have different opinions about the same image, or there can be great variability between different levels of expertise.
- Annotation and analysis: while other types of images can be easily classified by anyone, medical images require the opinion of a trained medical professional. Segmenting a single image manually, depending on the task and on the type of image,

can take from a few minutes to multiple hours. It is easy to see how a single image annotation can get pretty expensive.

2.2.1 | Magnetic resonance imaging

Magnetic resonance imaging, abbreviated as MRI, is a non-invasive radiology technique used to look at detailed anatomical representations of a subject. The final output of this process is a 3D image, or a sequence of 2D slices, that represent parts or organs of a patient. If the image is three-dimensional, the data is stored in voxels, the 3D correspective of a pixel in a two-dimensional picture. The process exploits the water or lipids that make up most of our human bodies, in particular hydrogen. [30] How hydrogen responds to a *pulse sequence* determines how that tissue will look in the final image. A strong magnetic field forces the only proton in hydrogen atoms into aligning with the magnet field, oriented along the MRI scanner axis. The different reaction levels in time and energy required for the protons in different tissues is what is measured by the scanner. The magnetic field is perturbed by a radio frequency energy (RF), which calms the protons back to their relaxation state, producing energy measured by a receiving coil. The process in which the protons relax, the repetition time (TR) and time-to-echo (TE) is what differentiate different types of MRIs. We measure TR as the time between successive pulse sequences applied to the same slice; TE is the time between the delivery of the RF pulse and the receipt of the echo signal. [2]

The resulting images change also based on the proton density, which is different in different tissues: proton density is higher in fluids like blood and cerebrospinal fluid, and lower in bones and tendons. This is what is displayed as shades of grey after a Fourier transformation of the signal. [30]

Doctors request a type of MRI scan based on which tissue or area is most important for them to inspect to help the patient at that time. We describe the few we needed and used for this study. Current limitations of magnetic resonance imaging are motion artifacts and prolonged acquisition time, as it can also be stressful for claustrophobic patients.

By varying the parameters of an MRI scan, we obtain different images. We describe the following three sequences in more detail: T1-weighted, T2-weighted, and Fluid-attenuated Inversion Recovery.

For more details about magnetic resonance imaging and the acquisition process, read Section 2.3.

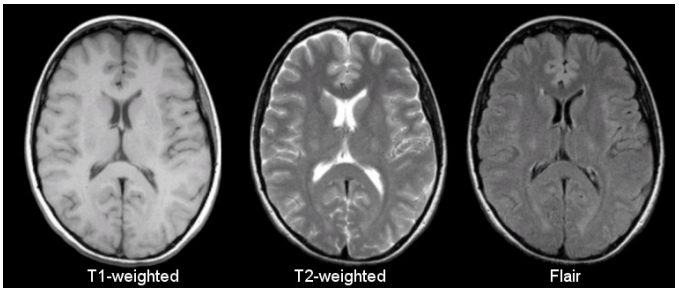


Figure 2.3: A comparison between three MRI sequences: T1-weighted, T2-weighted and FLAIR. [2]

2.2.1.1 | T1-weighted (T1)

With short TE and TR times we are able to amplify the T_1 differences between tissues, obtaining a image with a clear contrast. [30] Tissues with more water have a lower signal because they take longer to realign after a RF pulse, thus they appear dark. T1-weighted images provide valuable information about the anatomy of the body, and are especially useful for visualizing soft tissues. In neuroimaging, T1-weighted images are used to investigate brain morphology and asses structural changes due to various diseases and conditions.

2.2.1.2 | T2-weighted (T2)

T_2 properties of the tissues result in a entirely different image. With respect to T1-weighted scan, T2-weighted sequences require a longer acquisition time. Abnormal fluid are brighter against normal tissue, which we see darker. This is why T2 are considered to be "pathology" scans. [30] It is easy to tell a T1 from a T2 by looking at the cerebrospinal fluid (CSF): it is darker on the first, and brighter on the latter. T2-weighted images are sensitive to changes in water content in tissues, which is mostly affected by inflammatory processes such as infections or autoimmune disorders. We can also identify ischemic brain tissue because it looks hyperintense, being parts of the brain where there is a lack of blood flow.

2.2.1.3 | Fluid-attenuated inversion recovery (FLAIR)

A third type of MRI is also common, the Fluid-attenuated inversion recovery, or FLAIR. FLAIR is similar to a T2-weighted image, but TE and TR are even longer. Because the cerebrospinal fluid is hyperintense in T2-weighted images (Table 2.2), to detect bright lesions such as white matter lesions, we prefer FLAIR, which suppresses the cerebrospinal

Sequence	TR (sec)	TE (msec)
T1	0.5	14
T2	4	90
FLAIR	9	114

Table 2.2: A comparison between MRI sequences and their approximate TR and TE times. [2]

Sequence	Grey Matter	White Matter	CSF	Inflammation
T1	Grey	Light Grey	Dark	Dark
T2	Light Grey	Dark Grey	Bright	Bright
FLAIR	Light Grey	Dark Grey	Dark	Bright

Table 2.3: A comparison between MRI sequences and their shades of grey. [2]

fluid signal. Doing so enables us to distinguish normal tissue from lesions deriving from ischemic strokes and intracerebral hemorrhages. For this thesis, we look at the inflamed or abnormal areas of the brain, which show up as white matter hyperintensities in MRI images, and in particular, on the age-related lesions.

2.2.2 | Annotation of medical images

The task of looking at the MRI sequences and identifying abnormalities is performed by radiologists and similar medical experts. The standard protocol involves different clinicians and tasks. Take, for example, a patient that arrives at the ER with stroke symptoms. This person's condition would be evaluated by a neurologist, that could request an MRI or CT scan to have a complete picture of the situation. A radiologist will look at the scan and annotate on it any abnormalities of the brain physiology so that the neurologist could confirm their hypotheses and provide better treatment.

It is clear that in the case of an emergency such as that of a stroke, any minute matters for the well-being of the patient. The bottleneck in this process would probably be that of taking and annotating the scan of the patient's brain because the radiologist has to manually highlight any voxel that could be relevant to the diagnosis. Other than time-consuming, this process is also highly subjective and delicate, because of the many variables in play. These could be things such as the scanner manufacturer, the chosen sequence parameters, the ability of the patient of staying still for the whole duration of the process, and finally the inter-observed variability that introduces further biases, as no radiologist is the same as one other.

If the process of annotation could be automated, the medical implications in research

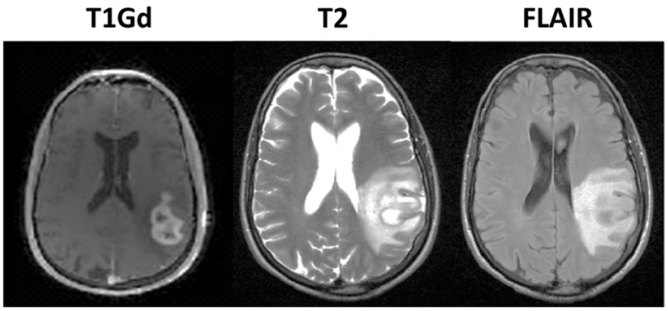


Figure 2.4: Three primary imaging modalities are used to visualize different aspects of brain tumors. T1Gd, which is T1-weighted with a gadolinium-based contrasting agent, helps identify areas with compromised blood-brain barrier due to tumor-induced angiogenesis, a particularly aggressive form of tumor. T2-weighted and FLAIR images are useful for detecting edema, the hyperintense region around the tumor. FLAIR provides better visibility of the edema by suppressing cerebrospinal fluid signal. Figure by [3].

and treatment of patients would be significant for many applications. In this thesis, we aim at developing a robust automated annotation method that would help in the analysis of medical images.

Image segmentation is the process that divides said image into different regions. These regions are characterized by properties like color (in practice, grey level), brightness, and contrast. [31] Segmentations are useful to study the anatomical structure, identify regions of interest like tissue abnormalities, measure said regions of interest, and, lastly, help in treatment planning. The output of a segmentation process, especially if automated, is affected by intensity inhomogeneity, artifacts, and closeness in the gray level of different soft tissue.

In practice, segmentation is done by separating the tissue into regions, which could be a simple binary relationship between them, like "healthy" or "not healthy". The annotation could be more complex to include different types of tissue abnormality, as it could be the case that a patient is affected by more than one pathology at a time: it is common, for example, that a patient with a tumor displays also other related conditions such as edema, like in Figure 2.4. The figure shows also why multiple MRI sequences are acquired at once, to have more context about the patient's conditions.

2.3 | Medical images: acquisition and archival

Medical images are data over three-dimensional spatial domains, or four-dimensional if acquired over time. [32] Fluid characteristics of different tissues are sampled and stored as discrete data points in formats like Digital Imaging and Communications in Medicine (DICOM) and NifTI-1 data format (NII). In this section, we will give an overview of how MR images are acquired and stored, and what kind of challenges arise when compromising quality over lower costs. For this section, we consulted [33] and [34].

Magnetic Resonance Imaging (MRI) is a powerful medical imaging technique used to visualize internal structures of the body with high clarity and detail. MRI employs strong magnetic fields and radiofrequency pulses to excite the protons in the body's tissues. As these return to their equilibrium state, they emit radiofrequency signals that are detected by the MRI scanner and used to create detailed images.

By utilizing a powerful, uniform external magnetic field to align protons within the water nuclei of the tissues, MRI exploits the magnetization properties of atomic nuclei to reconstruct images from the reconstructed signal. It does so by perturbing the alignment by introducing external Radio Frequency (RF) energy, causing the nuclei to emit RF energy as they return to their resting alignment through relaxation processes. After a specific time, the emitted signals are measured and transformed using Fourier transformation to create images with varying RF pulse sequences.

Unlike other imaging techniques, magnetic resonance imaging is greatly affected by the choice of parameters during acquisition.

A medical image dataset typically comprises one or more images that can represent different aspects, such as:

- The projection of an anatomical volume onto an image plane, through a projection or planar imaging.
- A series of thin slices through a volume, through tomographic or multislice two-dimensional imaging.
- Data from a isotropic volume, through volume or three-dimensional imaging.
- Multiple acquisitions of the same tomographic or volume image are taken over time to create a dynamic series of acquisitions, resulting in four-dimensional imaging.

Examples for each category are provided in Table 2.5.

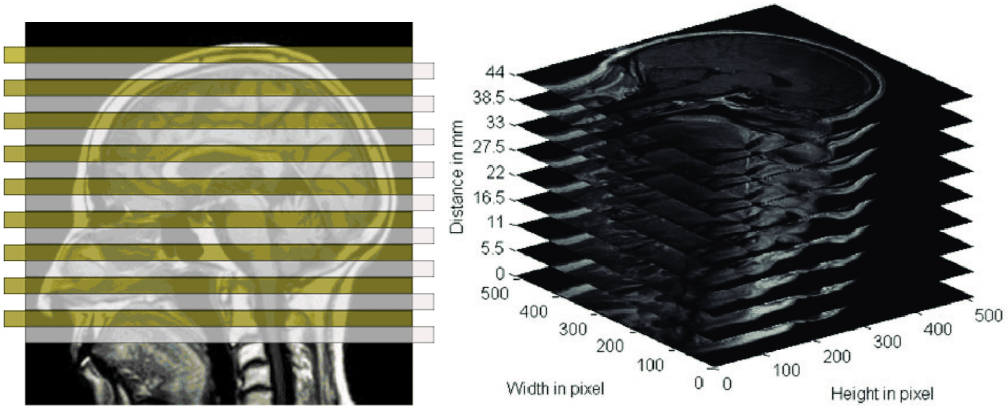


Figure 2.5: MRI sequences can be acquired as 3D volumes. but it is more common to acquire individual 2D slices, that are then stuck together. The slice gap in MRI refers to the space between consecutive image slices during data acquisition, which can lead to incomplete coverage, artifacts, and reduced image quality. Illustrations by [4] and [5].

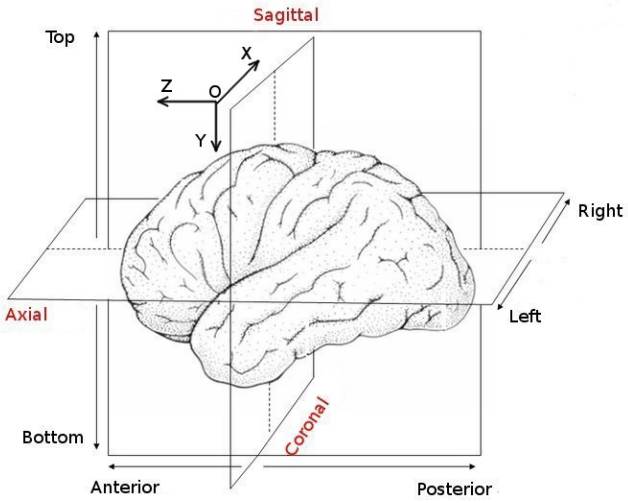


Figure 2.6: Illustration of the coronal, sagittal, and axial plane.

Parameter	Description
Radiofrequency pulse (RF)	Brief burst of electromagnetic energy that causes protons to flip their alignment.
Repetition Time (TR) and Time to Echo (TE)	They determine the timing between RF pulses and the acquisition of the echo signals. They influence image contrast and signal intensities.
Field Strength	The strength of the magnetic fields varies greatly. A stronger magnet is much more expensive but provides a better signal-to-noise ratio and higher resolution.
Slice Thickness and Gap	When we are not acquiring isotropic 3D volumes, we have to balance the trade-off between scan time and image quality. More slices provide more detail on smaller structures, but a gap between slices will always be present. (Figure 2.5)
Matrix Size and Field of View (FOV)	The spatial resolution and the coverage area of the image are determined by the matrix size and FOV, respectively. Increased resolution and area require increased scan times and storage space.
Imaging plane	We choose the optimal imaging plane based on which structures or lesions we are interested in visualizing. (Figure 2.6)
Contrast Agents	A patient may be injected with contrast agents if it is crucial for the visualization of enhanced, specific tissues or pathologies.

Table 2.4: Parameters of a magnetic resonance imaging protocol. They can be chosen based on the tissue that wants to be analyzed, but also by constraints given by time and cost resources.

2.3.1 | Image Formats

Image file formats offer a standardized means of storing information that describes an image. Standardization is critical in medical applications, where different healthcare sites and research institutes need due communicate and be able to display the images. The file format specifies how the image data are structured within the file and how the pixel data should be interpreted by each software to ensure proper loading and visualization.

After the acquisition, we have to choose in which image format we would like to store the images. Along the image, we store metadata, that [34] describes as *information that describes the image*. In particular, metadata includes a header with information about the image matrix dimensions or spatial resolution, but also on how the image was pro-

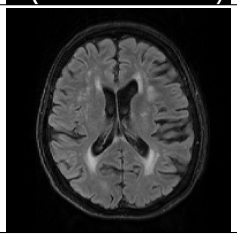
Imaging modality	Example	
Projection imaging		In X-ray imaging, a two-dimensional image is created by passing X-rays through the body and capturing the transmitted X-rays on a detector.
Tomographic imaging		Computed Tomography (CT) scans use X-rays and advanced computer processing to create cross-sectional images of the body.
Volume imaging		MRI can acquire isotropic volume data, meaning the resolution is the same in all three dimensions.
Four-dimensional imaging		In cardiac MRI, multiple images are acquired over time to create a dynamic series of images, commonly known as cine imaging. This allows visualization of the beating heart and assessment of its function.

Table 2.5: Different acquisition modalities for medical imaging.

duced, such as pulse sequence, timing information, et cetera. Metadata is used by the software application that opens the image, and by clinicians for diagnostic purposes.

To sum it up, the file format specifies how the image data are structured within the file and how the pixel data should be interpreted by software to ensure proper loading and visualization. This information influences all the steps of our automatic segmentation pipeline, from pre-processing to processing, and finally post-processing.

There are two main categories of medical image file formats: the first is the one intended for the standardization of images for diagnostic purposes, and the second category is mainly used for analysis. The Dicom format belongs to the first category, and it is the most commonly used by every medical imaging department. Dicom images are

rich in metadata, and citing [34], "[the Dicom header] contains the most complete description of the entire procedure used to generate the image ever conceived in terms of acquisition protocol and scanning parameters". DICOM enables interoperability among different medical imaging devices and healthcare systems.

The most common file format used for the analysis of medical images in the research community is the Nifti format. A Nifti's header is relatively simple if compared to the Dicom's one. It is limited to the information required for the visualization of the images, but more data can be added if needed. The conversion from one file format to another is often possible, but maybe not trivial. For this project, we relied on the Nifti format.

2.3.2 | Imaging artifacts

Medical imaging artifacts are unintended and often undesirable anomalies that can appear in medical images, compromising their accuracy and diagnostic value. These artifacts arise from various sources, such as patient motion, the presence of metal on the patient, or physical limitations of the imaging techniques. Radiologists rely heavily on the quality and fidelity of medical images to make assess the condition of a given patient. Artifacts can lead to misinterpretation, confusion, and even erroneous diagnoses, because they may alter the image by hiding or disrupting relevant information. We display in more detail which artifacts may arise in Appendix A.

2.4 | Automatic Segmentation Methods

The task of automatic annotation of medical images, among others, is particularly expensive. While it is easier for other categories of tasks to just have humans, with no specific qualification required, do the manual annotation, this does not hold in medical imaging. Expert radiologists and physicians are required in order to have valuable data, data which in its nature is also hard to gather, mainly for privacy and regulatory reasons. It might also happen that the already available images have not been segmented with a particular goal in mind, so they will end up lacking information or precision.

Automatic segmentation methods come to aid two groups of professionals: researchers, that are studying particular conditions and diseases in order to get insights on them, which could develop into new treatments in the future; or presently, to help doctors in clinical diagnosis and treatment planning for their patients, faster and more efficiently.

Compute-aided diagnosis goals are, mainly, to automate the diagnosis process so that more patients can be handled with no loss in accuracy. This can be achieved by the

faster processing time of machines, which lower costs and support faster communication, even in remote areas where specific patient care is not always guaranteed. [31]

Technically speaking, automatic segmentation is related to another image-related task, that of image classification. While image classification aims at answering the question 'is object X in the picture?', *semantic segmentation* wants to answer the question 'what objects are in the picture, and where are they?'. In other words, semantic segmentation doesn't assign a single class to the image as a whole, but assign a class to each and every pixel in the picture, usually including a class for the *background*. There are other types of image segmentation task, like instance segmentation and panoptic segmentation, but they are out of the scope of this work. Semantic segmentation outputs a segmentation map for each input image fed. Segmentation maps are usually mono-channel binary images, while the input can be multi-channel (e.g. RGB) and can take values in a predefined scale. In the stages that precede the final output, a segmentation map can take more values, for example, continuous values in [0,1] to indicate the *probability* of that pixel belonging to a certain class, values which are then thresholded to get a binary mask. If the classes are more than one, instead of a binary segmentation map, we can have one with discrete numbers to which we can assign one class each. Of course, one pixel can only belong to one class at the final stage of segmentation.

2.4.1 | Evaluation Criteria

We call x an image and A its associated manual annotation, made by a radiologist. We evaluate the performance of any given automatic segmentation method h by computing a metric over A and the segmentation map B given by $h(x)$.

We can measure the quality of segmentation by computing different similarity measures commonly found in the literature. Some of these are based on confusion tables, while others are based on similarities between overlapping voxels in the image. We can also compute distances between misclassified voxels in the segmentation map. [35]

Until May 2023, the STandards for ReportIng Vascular changes on nEuroimaging was the standard for the categorization of the features of cerebral small vessel disease. The guidelines have been updated by [22] to create STRIVE-2. According to this new standard, the suggested quality metrics for the evaluation of automatic segmentation methods of white matter hyperintesities (WMH) are the following measures: the Sørensen–Dice Similarity Coefficient (DSC), the Hausdorff Distance (HDD), and the Volumetric Similarity (VS). Consequently, we base our evaluation strategy on these metrics.

2.4.1.1 | Sørensen–Dice Similarity Coefficient (DSC)

The Sørensen–Dice Similarity Coefficient (DSC) is a spatial overlap index. It is the most used metric when validating medical volume segmentations. Commonly referred to simply as Dice, it allows a direct comparison between the automatic segmentation B and its manual annotation A. It ranges from 0 to 1, where 0 indicates no spatial overlap between two sets, and 1 is their perfect overlap. [6]

[36] defined DSC as

$$\mathrm{DSC}(A,B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2.1}$$

This measure has also been used as a statistical validation for manual annotations, where the same MRI image was repeatedly annotated to measure reproducibility between human raters.

2.4.1.2 | Volumetric Similarity (VS)

We define the Volumetric Similarity (VS) as $1 - VD$, where VD is the volumetric distance between A and B. We follow the definition by [6], that indicates the volumetric distance as the absolute volume difference divided by the sum of the compared volumes. It ranges from 0 to 1.

$$\mathrm{VS}(A,B) = 1 - \frac{||A| - |B||}{|A| + |B|} \tag{2.2}$$

This is not an overlap-based metric, as the overlap between the segments is absolutely not considered: only the absolute volume of the segmented region in one segmentation is compared with the corresponding volume in the other segmentation. We can have a maximum VS value even when the region overlap is zero. This fact highlights

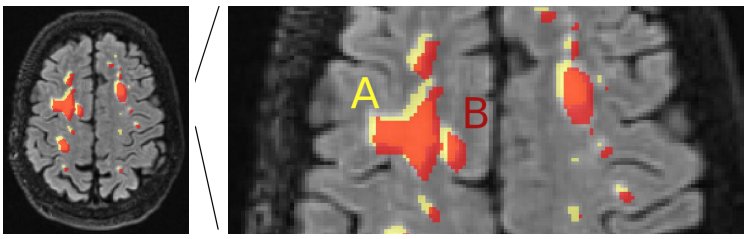


Figure 2.7: Given an image x and its associated manual annotation A, we evaluate the performance of a given automatic segmentation method h by computing a metric over A and the segmentation map B given by $h(\mathbf{x})$. To understand how each metric works, we can visualize A as the yellow spots on the image, and B as the red ones.

the divergent correlation between Volumetric Similarity and Sørensen–Dice Similarity Coefficient, as VS implicitly assumes that the segments are optimally aligned, which only makes sense when the overlap is high. As the overlap decreases, the probability of two segments not being aligned increases. This explains why the same segmentation map B can yield very different results for DSC and VS. We focus on the boundary and alignment of the segmented regions when computing metrics such as DSC, but if we are interested in the overall volume of white matter hyperintesities we might prefer to focus on Volumetric Similarity.

2.4.1.3 | Hausdorff Distance (HDD)

The Hausdorff Distance between two finite sets A and B can be defined as

$$d_H(A,B) = \max\left(\max_{a\in A}\min_{b\in B} d(a,b), \max_{b\in B}\min_{a\in A} d(a,b)\right) \tag{2.3}$$

where $d(a,b)$ is some norm, like the Euclidean distance.

Hausdorff Distance (HDD) is a spatial distance metric based on calculating the distances between all pairs of voxels from each mask (A or B). It measures the distance between crisp volumes, in terms of maximum distance from one point of a segmentation mask to the nearest point in the other mask. In general, spatial distance metrics are used as a dissimilarity measure, meaning metrics used to identify boundaries rather than magnitude. This metric is not upper-bounded, while a perfect match yields a score of zero. The difference between overlap and spatial distance metrics is explained in more detail by Figure 2.9.

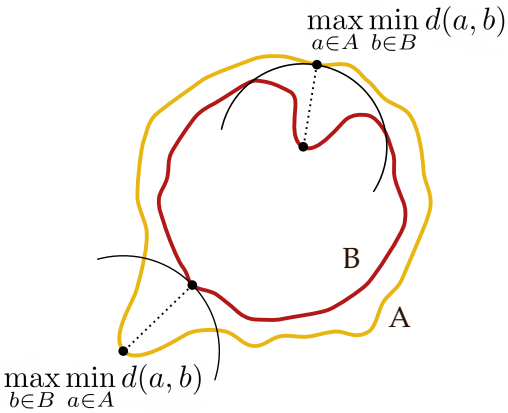


Figure 2.8: Visualization of the components for the calculation of the Hausdorff distance, between a yellow set A and a red set B.

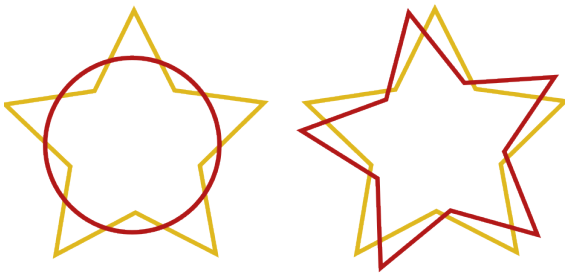


Figure 2.9: Visualization based on the one by [6], to show how metrics fail in the discovery of boundary errors. Given a black manual annotation in the shape of a star, we draw two red annotations with an equal magnitude. An overlap-based metric like Sørensen–Dice Similarity Coefficient (DSC) will assign to the red circle the highest similarity, while a spatial distance based metric like Hausdorff Distance (HDD) would correctly identify the red star as the closest prediction. However, the Volumetric Similarity (VS) will yield an exact match in both cases, as it is the metric that is most invariant to boundary errors.

2.4.1.4 | Confusion Matrix

We can reformulate 2.1 and 2.2 in terms of a confusion matrix $\mathcal{C}$. A segmentation task can be evaluated by viewing it as a binary classification of each individual voxel, and asking the following: how many voxels from my automatic annotation B have been correctly identified as a voxel from the manual annotation A? We will call *positive* a voxel that represents the target prediction (in our task, white matter hyperintesities), and *negative* a voxel that represents the background. Doing so, we can count the following:

True Positive	TP	number of voxels correctly predicted as positive in A and B
True Negative	TN	number of voxels correctly predicted as negative in A and B
False Positive	FP	number of voxels incorrectly predicted as positive in B but that are negative in A
False Negative	FN	number of voxels incorrectly predicted as negative in B but that are positive in A

We will have a perfect prediction when both the number of False Positives and False Negatives are zero.

Precision and *recall* are commonly used to evaluate a model. Precision is seen as a measure of quality, meaning that the model returns more relevant results than irrelevant ones when precision is high. Recall can be seen as a measure of quantity, higher when

the model returns most the the relevant results. Precision and recall are also know as *Positive Predictive Value* and *True Positive Rate*, and can we formulated as

$$\text{Precision}(A, B) = \frac{TP}{TP + FP} = PPV \tag{2.4}$$

$$\text{Recall}(A, B) = \frac{TP}{TP + FN} = TPR \tag{2.5}$$

However, precision and recall penalize errors in small segments more than they do in large segments, as they are very sensible to segment size. This fact hinders their use in medical image segmentation. [6]

More complex metrics can be derived from the components of the confusion matrix. The F_β *-Measure* (FMS_β) is a trade-off between precision and recall, defined as

$$FMS_\beta = \frac{(\beta^2 + 1) \cdot PPV \cdot TPR}{\beta^2 \cdot PPV + TPR} \tag{2.6}$$

With $\beta = 1$, the FMS is mathematically equivalent to the Sørensen–Dice Similarity Coefficient. We can write 2.1 as

$$DSC(A, B) = \frac{2TP}{2TP + FP + FN}$$

Additionally, we can write 2.2 as

$$VS(A, B) = 1 - \frac{|FP + FN|}{2TP + FP + FN}$$

2.4.2 | Previous work

Many different approaches have been developed and implemented to tackle the problem of white matter hyperintesities [37]. For example, a threshold-based method in 2001 derived a simple threshold from a regression analysis on the histogram of the FLAIR images. [38] It is also possible to use a k-Nearest Neighbors algorithm (kNN) [39] or a Support Vector Machine (SVM) [40] by extracting feature vectors from the images and then building a WMH probability map from it.

As shown by an analysis of [10] of the MICCAI 2017 WMH challenge (later discussed in section 3.1.1), the twelve top-performing methods for the segmentation of WMH use deep learning, with a U-Net-like architecture being the overall most common. These methods relied on FLAIR and T1-weighted images.

The results of the challenge, and further results on other imaging segmentation tasks, encourage us to follow the same path. We report a summary of previous methods in Table 2.6, including the test set size and on which MRI sequence they were trained on.

Another interesting study compared WMH segmentation tools most frequently used in research studies by clinicians to methods developed for the MICCAI 2017 WMH challenge, to compare the evolution and performance of new methods compared to the already available and vastly used in practice. [41] They selected the tools based on other performance studies: Brain Intensity Abnormality Classification Algorithm (BIANCA) [42], LST LGA [43] and LST LPA [44]. They compare these tools to the top two methods for the MICCAI 2017 WMH challenge at the time of the study, *sysu_media* [13] and *pgs* [45]. To assess the performance of these methods on 50 FLAIR test images, they used metrics from the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE), which have been updated recently. [22] Without re-training *sysu_media* on a subset of the available FLAIR test images showed the highest performance (DSC$= 0.63 \pm 0.19$), excluding the performance of said method but re-trained on a subset of the data. Given that we are looking for a method able to generalize on a wide cohort of subjects, we consider more valuable the performance on out-of-distribution data, thus excluding techniques that rely on re-training. [22] concludes that while *sysu_media* is the best method, its performance was close to traditional methods.

Lastly, we are motivated into the research of a more robust method because of the drop in performance of *sysu_media* when trained on FLAIR only. When adding T1-w information, *sysu_media* scored DSC=0.80 on the MICCAI 2017 WMH challenge test set, but relying on FLAIR only yielded a much lower DSC=0.63 on the out-of-distribution test set. [22]. We also leverage the power of deep learning to accelerate the translation

of volumetric brain imaging data to quantitative information.

A small portion of the work reported in this thesis was published and presented at the *Medical Imaging with Deep Learning* (MIDL) conference, held in Nashville in July 2023. [19] One reviewer suggested reporting more comprehensive comparisons between more recent multiple state-of-the-art methods. Given the discussion from results from [41] and [10], we conclude that it is only worth comparing our new method concerning *sysu_media*, the method by [13]. The latter has been extensively compared to other machine learning methods, both new (*pgs*, from 2021) and commonly used tools in medical imaging research (like BIANCA). We aren't aware of other recent studies that compare multiple methods on an out-of-distribution cohort. In the available studies, no other traditional or deep learning method reported a significant performance improvement over *sysu_media*, even if this method is, at the time of writing, already five years old. As we are purely interested in a robust automatic segmentation algorithm, which relies on as little information as possible, we are not interested in solely out-performing *sysu_media* or other state-of-the-art methods that rely on more MRI sequences or re-training to yield better performance. Furthermore, the purpose of this algorithm was to be tested on a very large cohort of patients from multiple Danish hospitals. As it is difficult to obtain standardized data across different sites, we chose to rely on a single MRI sequence. As being able to generalize to a wide distribution of data is often more complex than yielding a good performance on a single dataset, we prioritized this property over performance on the data from the challenge. As we will discuss in more detail at the end of the thesis, our method trained on FLAIR performs as well as *sysu_media* trained on FLAIR and T1-w, and performs better than *sysu_media* trained on FLAIR only. In conclusion, we only compare our experiments concerning *sysu_media*, as to our knowledge, no new state-of-the-art other method has been as extensively researched on its generalization capabilities.

Authors	Method	Test set size	Reported results
Jack et al. [38]	Segmentation of the intensity histogram of FLAIR	9	Mean absolute error of 6.6% on WMH volume
Anbeek et al. [39]	k-NN classification from voxel intensities and spatial information, using information from T1-w, inversion recovery (IR), proton density-weighted (PD), T2-w, and FLAIR.	20	A similarity index between 0.7 and 0.8, depending on the patient WMH volume
Kruggel et al. [40]	Probability map of WMH based on feature vectors extracted from T1-w and T2-w	116	Sensitivity of 0.901 and specificity of 0.913
sysu_media [13]	Ensemble of three fully convolutional neural networks similar to U-Net	110	Winner method of the MICCAI 2017 WMH Segmentation challenge. Highest DSC (0.80), lowest Hausdorff Distance (95th percentile), highest recall.
cian [10]	Multi-dimensional gated recurrent units (MD-GRU) trained on 3D patches	110	Second place at the MICCAI 2017 WMH Segmentation challenge. Lowest absolute percentage volume difference (0.193).
nlp_logix [10]	Multiscale Convolutional Neural Network, trained on ten folds and selecting the average of the best three performing checkpoints.	110	Third place in the MICCAI 2017 WMH Segmentation challenge. Highest F1-score (0.78).
nist [10]	Random Decision Forest classifier trained on location and intensity features	110	Highest non-deep learning method in the MICCAI 2017 WMH Segmentation challenge, ranked 12th. DSC of 0.63.

Table 2.6: A summary of previous work and the proposed methods for the automatic segmentation of white matter hyperintesities. The MICCAI 2017 WMH Segmentation challenge was open for new submissions after the challenge ended, so the ranking changed throughout the years. The methods reported in this table are based on the evaluation performed by Kuijf et al. [10].

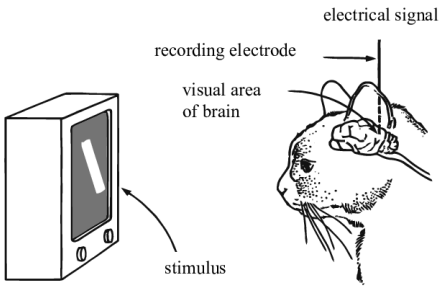


Figure 2.10: The experiment done by Hubel and Wiesel on a domestic cat visualizing bars of light. [7]

2.5 | Deep Learning

> "Why did AI hit so many roadblocks in previous decades? The reason is that most of the knowledge we have of the world around us is not formalized in written language as a set of explicit tasks — a necessity for writing any computer program. [...] That is where deep learning comes in.""
>
> *Yoshua Bengio for Scientific American, "Springtime for AI: The Rise of Deep Learning" (2016)*

In the context of artificial intelligence and machine learning, deep learning has emerged as a powerful paradigm, revolutionizing our ability to solve complex tasks across various domains. At the heart of this revolution lies the Multilayer Perceptron (MLP) [46], a fundamental neural network architecture that has paved the way for remarkable advancements in a wide variety of tasks. MLPs with sigmoid activation functions can approximate any continuous function, but the computations required would be prohibitively expensive. [47]

How can we leverage the nice approximation properties of multilayer perceptrons, if we do not have the resources to handle the exponential number of hidden units required to model nonlinearity? Instead of increasing the number of units *horizontally* in a single hidden layer, we can go *deeper*. In this context, the word *deep* means adding more hidden layers between our input and output layers: with a smaller number of neurons in each of them.

In 1981, David H. Hubel and Torsten Wiesel receive the Nobel Prize in Physiology or Medicine for their work in sensory processing and the development of the visual system. In particular, in 1959, they projected lights in patterns in front of a cat, which was previously anesthetized to insert a microelectrode in its visual cortex. [48] They noticed

that not all neurons react to lines of different angles equally, in fact, some activate when seeing horizontal lines and some when seeing vertical ones. Moreover, not every neuron sees the same, they instead react to only a small portion of the visual field. By overlapping the receptive fields of different neurons, the brain puts together the overall image seen by our eyes. To detect all the complex patterns, some neurons react to combinations of the straight lines detected by the 'lower layers' connected to them. This means that the brain is an intricate network of layers, that rely on different levels to notice patterns, starting from simple ones. This research prompted scientists working on artificial neurons to develop an architecture that resembles what an organic brain normally does, to detect patterns in digital images, following the *distributed representation* idea. The first to introduce and successfully implement this concept was [49], who used a multilayered neural network with hierarchies for character recognition of Japanese handwritten characters. Soon after, deep learning became popular due to its efficacy at solving very complex tasks. Moving deeper into a network permits the practical application of the universal approximation theorem under mild conditions. Together with an unbounded number of hidden layers, modern variations of the multilayer perceptron include techniques that have been empirically found to improve efficiency and trainability, but that are still difficult to prove theoretically. In this context, machine learning has shifted from a purely mathematical discipline into a more practical one, at the expense of understandability. A big concern with this family of models is that it is difficult for us to understand *why* a decision, such as classification, is made because the information is transformed many times from input to output. The strength of neural networks seems, nonetheless, to be enriched when the number of layers and neurons is increased, and when other techniques and optimizations are implemented.

2.5.1 | Cost and optimization

Training deep networks is challenging because we must be able to propagate the changes backward, from output to input. When we compute $\ell(\hat{y}, y)$ we know how much we are wrong on the prediction, but which weights should we update, and by how much? Deep learning algorithms use a process called *optimization*, where we try to minimize or maximize some function $f(\mathbf{x})$ by changing $\mathbf{x}$. In our case, we minimize our loss function ℓ by changing $\mathbf{w}$. *Gradient descent* is a technique that uses the derivative of a function to find in which direction we can move to find a minimum. (Figure 2.11) We can think of a nonlinear function as a landscape with valleys and hills. If we are on top of a hill, we want to move downwards toward the lowest valley, which we call the *global minimum* of the function. Any other valley that isn't the lowest one is called *local minimum*. Given

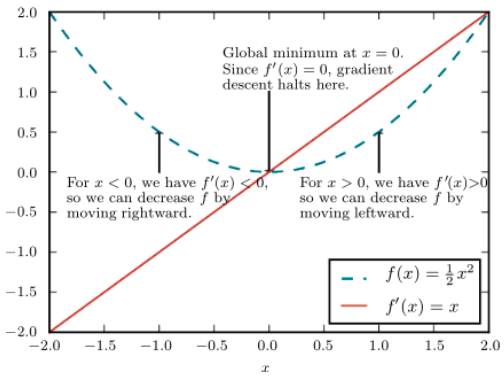


Figure 2.11: The gradient descent algorithm uses derivatives to follow a function downhill toward its minimum. [8]

a function $y = f(x)$, where $x, y \in \mathbb{R}$, we can move in small steps toward a minimum by computing the derivative of f, denoted as $f'(x)$ or $\frac{\partial y}{\partial x}$. The derivative $f'(x)$ is the *slope* of $f(x)$ in the point x. We obtain a small change ϵ in the output by scaling it by ϵ with the derivative of the function

$$f(x + \epsilon) \approx f(x) + \epsilon f'(x)$$

Knowing how to change x to minimize the function allows us to make small improvements in y, and we do it by moving x in small steps in the opposite sign of the derivative. This optimization process can become difficult, because we may get stuck in a local minima or find ourselves in a very flat region that is difficult to descend. Lastly, our inputs are solely multidimensional, which further increases the complexity of this process. Functions with multiple inputs must make use of *partial derivatives*.

We define a partial derivative $\frac{\partial}{\partial x_i} f(\mathbf{x})$ the function that measures how f changes when the single variable x_i increase at point $\mathbf{x}$. We can generalize the gradient to the case where we have multiple input vectors: we denote as $\nabla_{\mathbf{x}} f(\mathbf{x})$ the gradient of f, being the vector containing all the partial derivatives for each x_i. In other words, each element i of the gradient is the partial derivative of f with respect to its input x_i. Following our landscape analogy, the gradient points directly uphill, so if we want to decrease f we must move in the direction of the negative gradient. The *method of the steepest descent* proposes a computation with the gradient and a new term called the *learning rate* ϵ, by computing the next step $\mathbf{x}'$ as

$$\mathbf{x}' = \mathbf{x} - \epsilon \nabla_{\mathbf{x}} f(\mathbf{x})$$

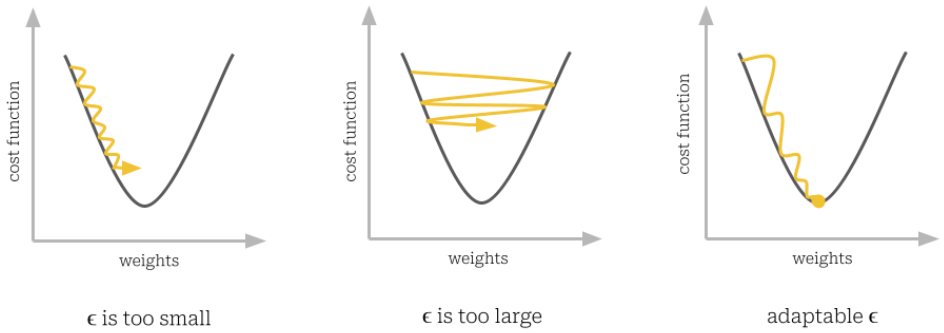


Figure 2.12: The learning rate ϵ, when set too small, increases the number of training steps needed to reach the minimum. On the contrary, if ϵ is too large we might completely miss the valley. It is advisable to have an adaptable ϵ, based on some conditions based during learning. The most common strategy is to start with a large learning rate and to reduce it when our predictions don't improve after a certain number of training steps.

The learning rate ϵ is often set to a small constant or can be chosen with more complex strategies. Controlling the step size can be challenging, as it affects the number of iterations needed to reach the minimum, or if we will reach it at all (Figure 2.12).

A common choice of loss function ℓ in medical imaging segmentation is the Dice loss [50], or a variation of it. Given a *ground truth* annotation G, which in our task is the manual annotation of an image made by a radiologist, we call P the predicted probabilistic maps. Both are computed over N voxels, so that $G = \{g_1, ..., g_N\}$ and $P = \{p_1, ..., p_N\}$. We define the Dice loss between two binary volumes as

$$\mathcal{DL} = \frac{2\sum_{n=1}^{N} p_n g_n}{\sum_{n=1}^{N} p_i^2 + \sum_{n=1}^{N} g_i^2} \tag{2.7}$$

This function can be differentiated yielding the gradient

$$\frac{\partial D}{\partial p_j} = 2\left[\frac{g_j\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) - 2p_j\left(\sum_i^N p_i g_i\right)}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right)^2}\right] \tag{2.8}$$

that we compute over the j-th voxel of the prediction. If we treated segmentation as a simple per-voxel classification task, we would encounter a problem called *class imbalance*, because most of the voxels in the segmentation map belong to the *background* class. The target class, in our case the WMH voxels, represent the minority over the whole volume. That is why a metric like the *accuracy* is not useful (Table 2.7). Using the

Metric	Score
Accuracy	0.99995
Precision	0.53784
Recall	0.61053
DSC	0.57188
VS	0.93669

Table 2.7: Metrics computed over a manual segmentation A and the predicted segmentation B by some model h. The accuracy, defined as $\frac{TP+TN}{TP+TN+FP+FN}$ is a misleading metric in segmentation tasks. In medical imaging, for instance, if the background is much larger than the foreground, high accuracy can be achieved by simply predicting everything as background, even though the model might be failing to accurately segment the important regions. In fact, the Dice score is pretty low. In this example, we can also notice the difference between DSC and VS: even though the VS is pretty high, indicating that the automatic segmentation volume B is close to the ground truth, this volume doesn't overlap well with the manual annotation A. This contradicting comparison is made possible because this is a very particular case, as this image has only 873 voxels annotated as WMH over 19'158'000 total voxels in the image.

formulation in (2.7) and (2.8) allows us to assign weights to samples of different classes to balance the voxels of background and foreground.

A deep learning algorithm is often trained over hundreds, thousands, or even millions of samples. Loss functions are computed over a single training example, meaning computing gradient descent over a set of examples Θ can be formulated as

$$\nabla_\Theta \ell(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_\Theta \ell(\hat{y}^{(i)}, y^{(i)}) \tag{2.9}$$

Computing gradient descent has a total cost of $O(m)$, meaning that a single training step over thousands or millions of samples can take very long. Instead, we can compute the gradient as an expectation using the *stochastic gradient descent*, or SGD. We build a *minibatch* $\mathbb{B}$ of uniformly sampled examples from Θ. The size of $\mathbb{B}$ is fixed and usually small with respect to Θ. We can formulate the estimate $\mathbf{g}$ of the gradient as

$$\mathbf{g} = \frac{1}{m'} \nabla_\Theta \sum_{i=1}^{m'} \ell(\hat{y}^{(i)}, y^{(i)}) \tag{2.10}$$

Consequently, given the learning rate ϵ, the stochastic gradient descent algorithm follows the estimated gradient downhill

$$\Theta \leftarrow \Theta - \epsilon \mathbf{g} \tag{2.11}$$

The optimization is not guaranteed to arrive at the global minimum, and not even at a local minimum, in a reasonable time. However, it still finds a very low value of the cost function quickly enough, as often the number of steps required to converge scale with Θ.

Optimization algorithms such as stochastic gradient descent provide a scalable way of training deep learning models over a large number of examples. However, the training is still affected by a variety of challenges.

2.5.2 | Validation and testing

Ensuring that a trained deep learning model generalizes well to unseen data is a critical step. Validation and testing procedures are employed to assess the model's performance. During training, a portion of the dataset is often set aside as a validation set. The model's performance on this validation set helps monitor its progress and prevents overfitting, where the model learns to memorize the training data rather than capturing underlying patterns. After training, the model is evaluated on a separate testing dataset that it has not seen before. This final evaluation provides an accurate estimation of the model's real-world performance. This test set is usually acquired from the same source as the training and validation set because of limitations in the data acquisition process. However, the generalization capabilities of a model should be evaluated on *out-of-ditribution* data, especially when we aim at deploying our model *in the wild*, meaning in real-life scenarios.

2.5.3 | Regularization

Underfitting is the problem that arises when a model is not capable of *fitting* the data properly. In other words, it is not capable of finding patterns in the data that would yield correct predictions. Underfitting can be solved by increasing model complexity, by adding new layers or nodes or acquiring more training data. If these strategies don't work, we could be facing the problem with inappropriate tools, or facing a too complex task. On the contrary, deep neural networks are usually concerned with the opposite problem, that of *overfitting*. When our model relies too much on learning patterns from the available data, it hinders its ability to generalize on new data. It might also happen when the training set is *class imbalanced*, meaning that the majority of examples belongs to few classes: in these case, the model will simply be rewarded for adjusting its weights to the most common case, rather than the few outliers that didn't affect the learning enough for it to adjust to them. Regularization techniques mitigate overfitting

and improve the model's generalization capabilities. Some techniques such as L1 and L2 regularization, affect the training process, by adding penalty terms to the cost function based on the magnitudes of the model's parameters, discouraging excessively large weights. Regularization techniques, like early stopping or ensembling, don't interfere with the model learning but leverage what it has learned so far. We will now name and shortly explain a few regularization techniques, but many more are available.

2.5.3.1 | Early Stopping

Early Stopping is a regularization technique that combats overfitting by monitoring the model's performance on a validation dataset during training. We can exploit the validation error to stop the learning process when we notice that the error starts to increase, indicating that the model's performance on unseen data is deteriorating. The validation error is usually computed at the end of each *epoch*, where an epoch can be defined differently through a learning pass. An epoch can be defined as a predefined number of training steps, or as a full pass over the training data.

2.5.3.2 | Dropout

Dropout is another popular regularization technique where randomly selected neurons are ignored during training. This prevents the network from relying too heavily on specific neurons and encourages the distribution of learning across the network. Given an input $\mathbf{x}$, we call $\mathbf{M}$ a binary mask matrix that indicates which neurons are active. We get the input $\mathbf{y}$ from

$$\mathbf{y} = \mathbf{M} \odot \mathbf{x} \tag{2.12}$$

where $\odot$ represents element-wise multiplication. Consequently, individual neurons are dropped out from the learning pass with a certain probability in $\mathbf{M}$.

2.5.3.3 | Data augmentation

The scarcity of annotated data, in particular in the medical imaging field, poses a big threat to deep learning methods, which are notoriously *data-hungry*. An efficient technique to increase the number of samples in a training set is *data augmentation*. Data augmentation is the process of applying various transformations to the original training set, such as rotations or other affine transforms, or more complex functions. This process artificially increases the training set, adding plausible variations to the data.

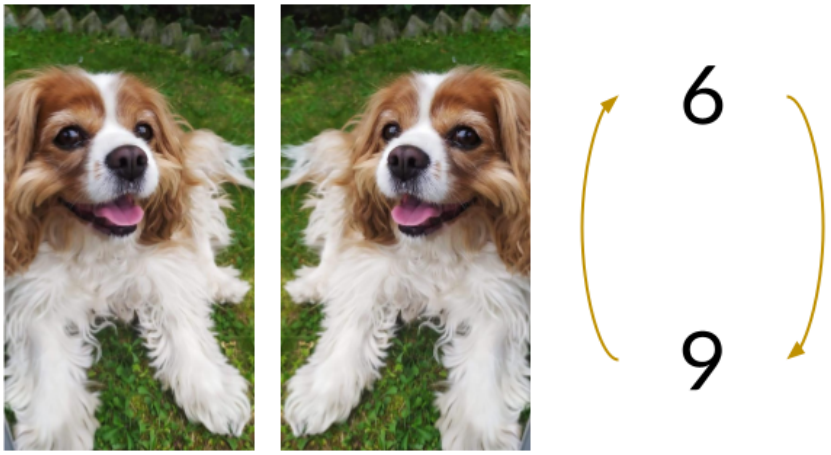


Figure 2.13: The transformations applied to data during the data augmentation process are usually driven by randomness but within some boundaries. For example, it would make sense to randomly decide whether or not to flip the image of a dog (a), as it would still clearly represent a dog (b). However, in a digit recognition task, the range within which we can accept a rotation should be bounded. Rotating a 6 by rotation close to 180° (c) would transform it into a 9. With a ground truth annotation of 6, a perfect model would predict a 9, which contradicts the annotation, thus confusing the model about what it is learning.

2.5.3.4 | Model ensemble

Another technique that improves the generalization capabilities of a model is *ensembling*. There are various ensembling strategies, but they all involve training more than one model to decide on the prediction outcome. A common strategy is that of *voting*, or *averaging*, *ensemble*. In this method, multiple models are trained separately, either with the same learning method and different training sets splits, or training on the same dataset with different algorithms. Ensemble models generally improve the method generalization capabilities because different models can learn different attributes of the training data, as the models' weights are randomly initialized. Additionally, if a model error is due to bias and variance, when the variance decreases, the overall error would also likely decrease. We can lower the variance by averaging the model's outputs. We will further discuss model ensembling when discussing the different methods used for this task, in Chapter 3.

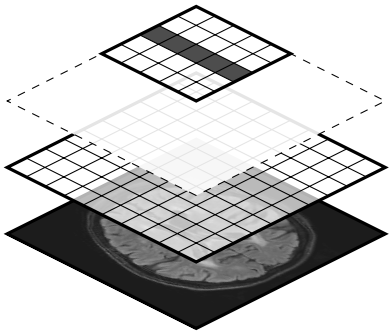


Figure 2.14: A convolutional layer scans the input image pixels with a filter to output a new feature map. In this toy example, a filter of size 5×5 detects a vertical line of 5 pixels.

2.5.4 | Convolutional Neural Networks

A neural network architecture determines the network's capacity to learn complex relationships in the data. We can design its architecture by deciding the number of layers, the number of neurons in each layer, activation functions, and connections between layers. Architectural innovations like convolutional layers for image data and recurrent layers for sequential data have significantly improved the performance of deep learning models in specific domains. Transfer learning, a technique where pre-trained models are fine-tuned for specific tasks, has also gained prominence, reducing the need for training large models from scratch. In practice we only talked about *fully-connected neural networks*, where we assumed that each neuron of a layer had a connection to the neuron in the next layer, and so on.

LeCun et al. [51] introduced the LeNet-5 architecture, a "complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing". This architecture classified hand-written digits and was successfully applied by banks to read cheques because it outperformed other models. The novelty of this approach is the *convolutional layer*.

Normally, a neural network layer has all its input nodes connected to the next layer, which implies that every pixel of the input image is computed at once. Instead, a convolutional layer loses the concept of pixel positioning in the image grid, and computes the output for a "window" of values, that moves over the image and whose output is directly connected to the next layer. In a two-dimensional task, we can visualize a filter as a smaller grid over an image, that is in itself a grid of pixels (see Figure 2.14). The window is called a *filter* or *kernel*. A filter is applied to the receptive field of the neuron in order to detect patterns: it performs an element-wise multiplication and sums up the

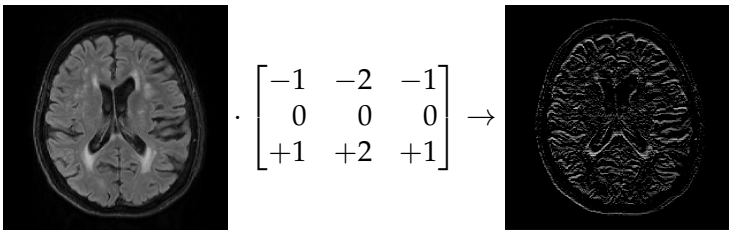


Figure 2.15: Feature detection example: Sobel filter (3×3) applied to detect horizontal lines in the image. A convolutional layer *learns* its own filters.

result to form the *feature map*. (See Figure 2.15) A convolutional layer is a stack of multiple filters and their respective feature maps. A *filter* or *kernel* is the neuron's learned weights represented as a small image, usually of the size of 3x3, 5x5 or more pixels. The size of a filter, or how many of them we want to learn per each given image window, is one of many of the network parameters. After a filter, we have a *pooling layer*, which reduces the feature space into something smaller, to which we can then apply another filter. Combining these layers we get a *convolutional neural network* (CNN), which in the case of image classification is still connected by fully connected layers at the end of the network, in order to predict the class from the extracted features. This last layer acts on a much smaller input vector, with respect to the full network input, because the pooling operation reduced the feature space into something much smaller than only keeps the information useful to the network to distinguish classes. (Figure 2.16)

Convolutions can be applied to any-dimensional input, from one to n, but in medical imaging, we usually work with two-dimensions if working with slices, in three-dimensions if working with volumes, and the less common four-dimension if time is another piece of information available. We can also work with multiple *channels*, which in pictures is usually the three *RGB* channels for color, or in medical imaging can be multiple MRI sequences. This is possible because the signal acquired by MRIs is in a single domain ($\mathbb{N}$ or $\mathbb{R}$), so we can exploit techniques developed for color channels by co-registering different sequences with respect to each other. In the end, the feature vector loses every information about the localization of a feature in the original image. This is not relevant in a classification task, but it is in a segmentation task, where we don't only want to assign to each pixel a class, but also know *where* the pixel is in the output. We can retain the spatial information by using transposed convolutions, also known as *deconvolutions*, which help upsample the features back to the original input dimensions. Additionally, skip connections, a technique originating from the *U-Net architecture*, further enhance spatial preservation by combining features from different

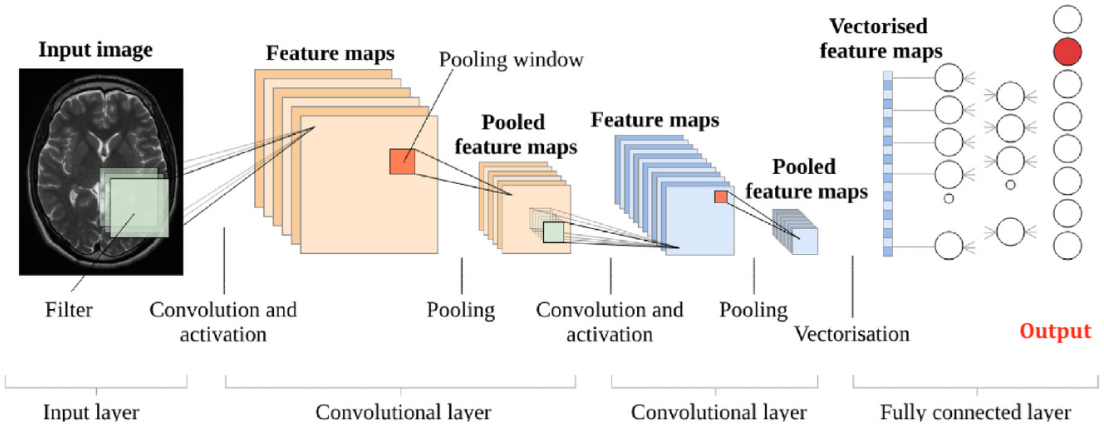


Figure 2.16: Building blocks of a CNN [9]

levels of the network hierarchy. We will discuss the U-Net architecture and its variant in the following chapters.

3

Materials & Methods

3.1 | Data

3.1.1 | MICCAI WMH dataset

In 2017, the MICCAI Conference (Medical Image Computing and Computer Assisted Interventions) hosted a challenge that invited any team to participate in developing a new algorithm for the automatic segmentation of White Matter Hyperintesities of presumed vascular origin. [10] The main conference was hosted in Quebec City, in Canada, from September 10th to 14th. Twenty teams participated in what is known as the *WMH segmentation challenge*. Each team was given 60 images to train their methods, and 110 images were kept a secret to be used in the evaluation of the performance, computing five different metrics. The dataset used for the models' training and testing has since been made public, including the images used to score the teams' performance. The dataset, in order to test the robustness of the algorithms for domain shift, was made by sets of images acquired from five different MRI scanners, made by three different vendors, coming from Singapore and the Netherlands. This implies that each image came from a different data distribution. The single site datasets details are summarized in 3.1. For each subject, the teams were provided with one 3D T1-weighted image and a 2D multi-slice FLAIR image. For the purpose of this work, we chose to not work with T1 in order to make our method as general as possible, so that it could be applied in contexts when only FLAIR is available. For each FLAIR image, the data provided includes the original FLAIR image, used by annotators to manually segment the WMH for the ground truth segmentation map, which is also provided. This image was also *bias field corrected*. This correction is a pre-processing process used to correct bias field inhomogeneities, and it was made using *SPM12 r6685*. SPM12 is an "academic software toolkit

for the analysis of functional imaging data". [52]

The manual reference for the image annotation contains three labels:

0. Background

1. White matter hyperintesities

2. Other pathologies

The teams participating in the challenges were evaluated only on their ability to segment WMH, so they created a rough mask of other brain abnormalities that might have occurred in the image in a single label "other pathologies". This is for them to ignore during evaluation when computing specific metrics on WMH only.

Table 3.1: MRI Scanner Parameters. [10]

Sites	Scanner	FLAIR Voxel (mm^3)	TR/TE (ms)	Train (n)	Test (n)
NUHS Singapore	3T Siemens TrioTim	$1.00 \times 1.00 \times 3.00$	9000/82	20	30
UMC Utrecht	3T Philips Achieva	$0.98 \times 0.98 \times 1.20$	11000/125	20	30
VU Amsterdam	3T GE Signa HDxt	$0.98 \times 0.98 \times 1.20$	8000/126	20	30
VU Amsterdam	3T Philips Ingenuity	$1.04 \times 1.04 \times 0.56$	4800/279		10
VU Amsterdam	1.5T GE Signa HDxt	$1.21 \times 1.21 \times 1.30$	6500/117		10
Total				**60**	**110**

The images were segmented following the STRIVE. Using a contour drawing technique, one expert observer delineated the outline of the WMH. A second observer performed a peer review of the manual delineations from Observer 1. In case of disagreement, Observer 1 would have corrected any mistakes, errors, or delineations not following STRIVE standards. The provided segmentation map is the corrected segmentation of Observer 1 after it was reviewed by Observer 2. Additionally, two other observers (3 and 4) manually segmented the images and were also subjected to peer review. The final images are the result of agreement between the observers.

To obtain binary masks, any voxel whose volume was equal or exceding 50% was included within the manually drawn contour, which ended up with a label of 0 for the background, and the foreground divided into *WMH* with label 1 and *other pathologies* with label 2. In case of overlap between label 1 and label 2, label 1 was assigned.

This dataset is the main source of testing because of the quality of the annotation and its large size. It also tests the model generalization capabilities because it includes 20 images from sites not included in the training set.

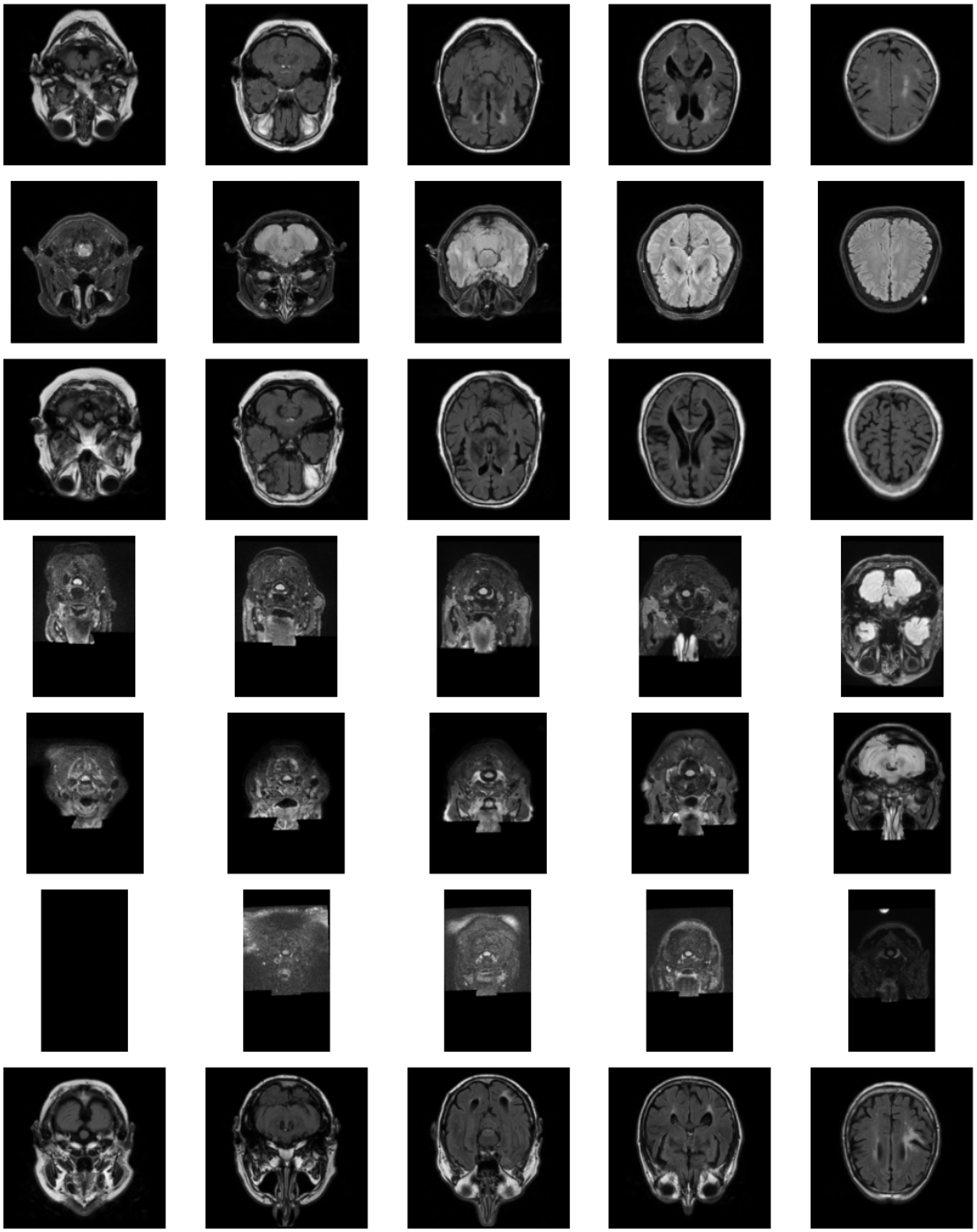


Figure 3.1: Samples from the MICCAI WMH Segmentation challenge training set. [10] To display the great variability within the same type of data, in each row it is plotted the same slice number for each image.

3.1.2 | INDUSA dataset

To test the robustness of the methods presented in this thesis, I compared their performance on an additional dataset that is not publicly available. This dataset, which we will refer to as *in-house dataset* or *INDUSA*, has been originally annotated for other purposes. In fact, this dataset was acquired and annotated for a different segmentation task, that is the segmentation of tumors, infarcts, and hemorrhages. Other pathologies are also present, and luckily, also white matter hyperintesities (WMH). We report the full annotation list in Appendix **??**, as it gives the reader an idea about what the original intent of the dataset is. This is important because this dataset was not intended for the segmentation of WMH. Rather, these lesions were annotated for the purpose of distinguishing them from other hyperintense pathologies, such as infarcts. This means that the annotations are not as precise as the ones from the MICCAI dataset. As the labels are not as precise, we can expect a lot of variation in the quantitative evaluation metrics. Visual inspection is important to assess the quality of the automatic segmentation.

Following the reasoning of the WMH segmentation challenge, this dataset labels were aggregated so that only two of them would actually make the cut in the final ground truth segmentation map: in order to have only two labels, except the background, all non-WMH labels were assigned the "other pathologies" (2) label.

The dataset has 22 test cases. Age, sex, and other personal data about the patients are not present. The images come from patients in India and the United States, so we can call this dataset INDUSA.

3.1.3 | Low-field MRI dataset

Until now, we have only talked about MRI scanners that use a magnetic strength of 1.5 or 3 Tesla. These scanners are the most commonly used in clinical practice, as they provide a lot of detail with reasonable resources. However, in recent years the radiology community has seen an increased interest in the deployment and use of *low field* scanners. Low-field-strength ($<$0.3T) and portable very-low-field strength ($<$0.1T) MRI systems are not yet widely adopted in clinical practice, mainly because of the images' poor resolution and low signal-to-noise ratio (SNR). [53] However, these systems have many advantages over higher field strength devices, such as:

- The lower sitting requirements reduce installation and maintenance costs, with commercial systems price scaling with the magnetic field strength ($1 million/Tesla). [54]

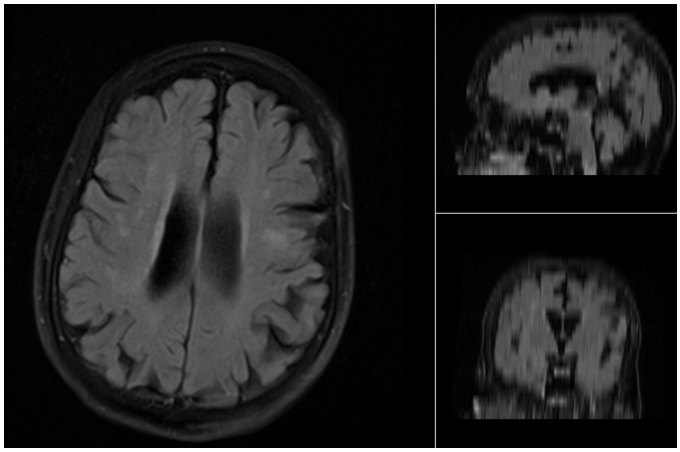


Figure 3.2: One case from the low-field MRI dataset.

- Clinicians in ICUs may prefer MRI over CT to rule out intracranial problems, however transporting patients on life-sustaining devices is challenging and can cause significant adverse events. The use of portable MRI has been found safe and feasible, and their use can reduce the need and cost of diagnostic test transportation, which constitutes 77% of total in-hospital transportation. [55; 56; 57]
- In stroke cases, where time is brain, very-low-field devices have been successfully implemented into fast clinical workflows, demonstrating high stroke sensitivity. [53; 58]
- MRI reduces the risk profile for implants and can be used during interventional procedures [53; 58]
- The acquisition time for a bedside 0.064T pMRI system is around 5-10 minutes per sequence. [59]
- Small and faster systems may improve patient acceptance for claustrophobic and pediatric patients. [53; 58]
- Patients with acute ischemic stroke who received care in an ambulance equipped with a portable CT scanner received intravenous thrombolysis 30 minutes quicker compared to patients with similar symptoms in conventional emergency care. [60]

Interest in portable low-field counterparts has been increasing in recent years due to technological innovations that have improved the quality of the images. Other than advances in hardware materials and design, artificial intelligence methods have been developed to improve the resolution and the signal-to-noise ratio of low-field MRI acquisitions, such as super-resolution and de-noising methods [61; 62]. These techniques

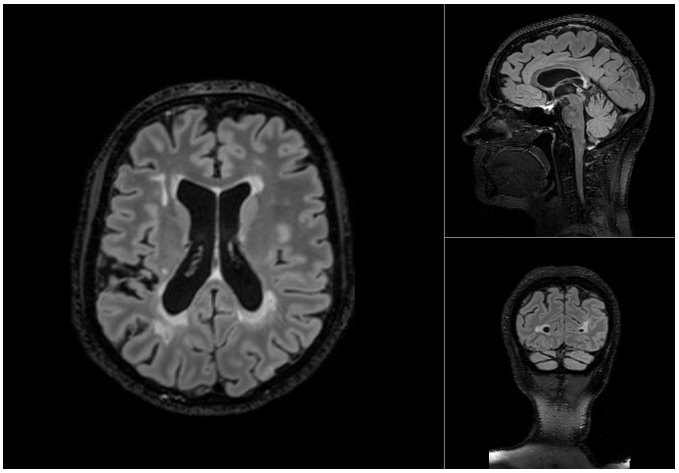


Figure 3.3: One case from the Multiple Sclerosis dataset. [11]

usually aim at making images obtained from low-field systems look like those from 1.5T or 3T scanners. This may help reduce the amount of training radiologists need in order to read sequences from these scanners. Automatic post-processing protocols for image reconstruction, when evaluated with respect to diagnostic quality by expert radiologists, have been judged good at accurately recovering fine details while still being robust at minimal perturbations. AI methods could improve patient experience and scanner efficiency, reducing costs for healthcare networks by having a 45% scan time reduction and non-inferior image quality [63]. However, most methods focus on optimizing common imaging metrics, such as the L1 distance between images, but neglect to optimize for clinical performance. While current solutions have shown that increasing resolution and de-noising can improve image detail, they may fail by introducing artifacts or by not reproducing some significant information. Most importantly, experts advise that the image contrast should be taken into consideration when assessing the post-processed images. This is due to the different physical properties of diverse magnetic fields which can get lost during reconstruction [63; 64].

As it is still not possible to rely on upscaled images, the algorithms have been tested directly on some available low-field images. No annotations are available, so we will only visually inspect the predicted segmentation. Moreover, as it is difficult to assess their quality and correctness even by expert radiologists, my comments are personal and may be incorrect. Nonetheless, it is still interesting to compare the methods on images with very low quality and completely out-of-distribution with respect to the original training set.

3.1.4 | Multiple Sclerosis Dataset

Multiple Sclerosis (MS) is a chronic inflammatory disease of the central nervous system. On magnetic resonance images, especially FLAIR, MS appears as hyperintesities just as the previously discussed age-related white matter hyperintesities. The main distinction between them is that MS lesions tend to be juxtacortical or intracortical, meaning that they appear adjacent to the cortex. There are also a few other ways to tell apart the two forms of lesion, but they are subtle and hard to notice for non-expert radiologists. The evolution of the disease can be monitored with MRI, thus giving insights into how clinicians can provide treatment to each patient.

As MS and WMH are visually similar, we can test the proposed method on a dataset from the Multiple Sclerosis Segmentation Challenge, hosted by MICCAI in 2016. [11] The dataset is made of 53 patient images from different centers, and they come with a ground truth annotation made by expert radiologists.

Dataset	Training	Testing
MICCAI WMH	60	110
INDUSA		22
LF		15$*$
MS		53
Total	60	200

Table 3.2: Number of cases available for each dataset, and how many have been used for training and testing. Note that the *LF* testing set has no ground truth annotation for quantitative evaluation (indicated with $*$).

3.1.5 | Data summary

Summing up, the algorithms were trained exclusively on the MICCAI WMH segmentation training set (60 images from 3 different scanners) and tested on four testing sets, the one used to evaluate the teams participating in the challenge (110 images, from 3+2 unseen scanners), an in-house dataset (INDUSA), a low-field MRI dataset with no available ground truth annotations (LF) and the Multiple Sclerosis dataset from MICCAI 2016 (MS). The number of data points is as seen in Table 3.2.

From Figure 3.4 we can see that the dimension that displays more variability is the third dimension z. Moreover, the average third dimension in the images slice shape is lower with respect to the other two. Both of these facts are due to the acquisition nature of MRI scans. The number of voxels per dimension (a) affects the size of the final image. The INDUSA and the LF datasets have similar x and y shapes, but z is very low, meaning that very few axial slices are available. From Figure 3.4 we can also see the voxel resolution distribution across datasets (b). If all three dimensions have the same value, the image is *isotropic*. That is the case for the MICCAI WMH training set, which has been resampled to an isotropic voxel resolution. A consequence of having very few slices is that INDUSA and LF need to cover more space between each slice, meaning having a very large z resolution. The MS dataset has a very high resolution, as it was acquired with 3D scanners.

The image volume size distribution is shown in Figure 3.5c. It is clear that the datasets that have more variability across used scanners, will show a wider distribution. The mean and standard deviation in Figures 3.5d and 3.5e are useful to show that a certain MRI scanner may have values within different ranges. The distribution is affected by the presence of hyperintesities, but generally speaking, each scanner has a different way of acquiring images that results in a more or less wide range of values.

It's also important to visualize how many WMH are present in each annotated dataset.

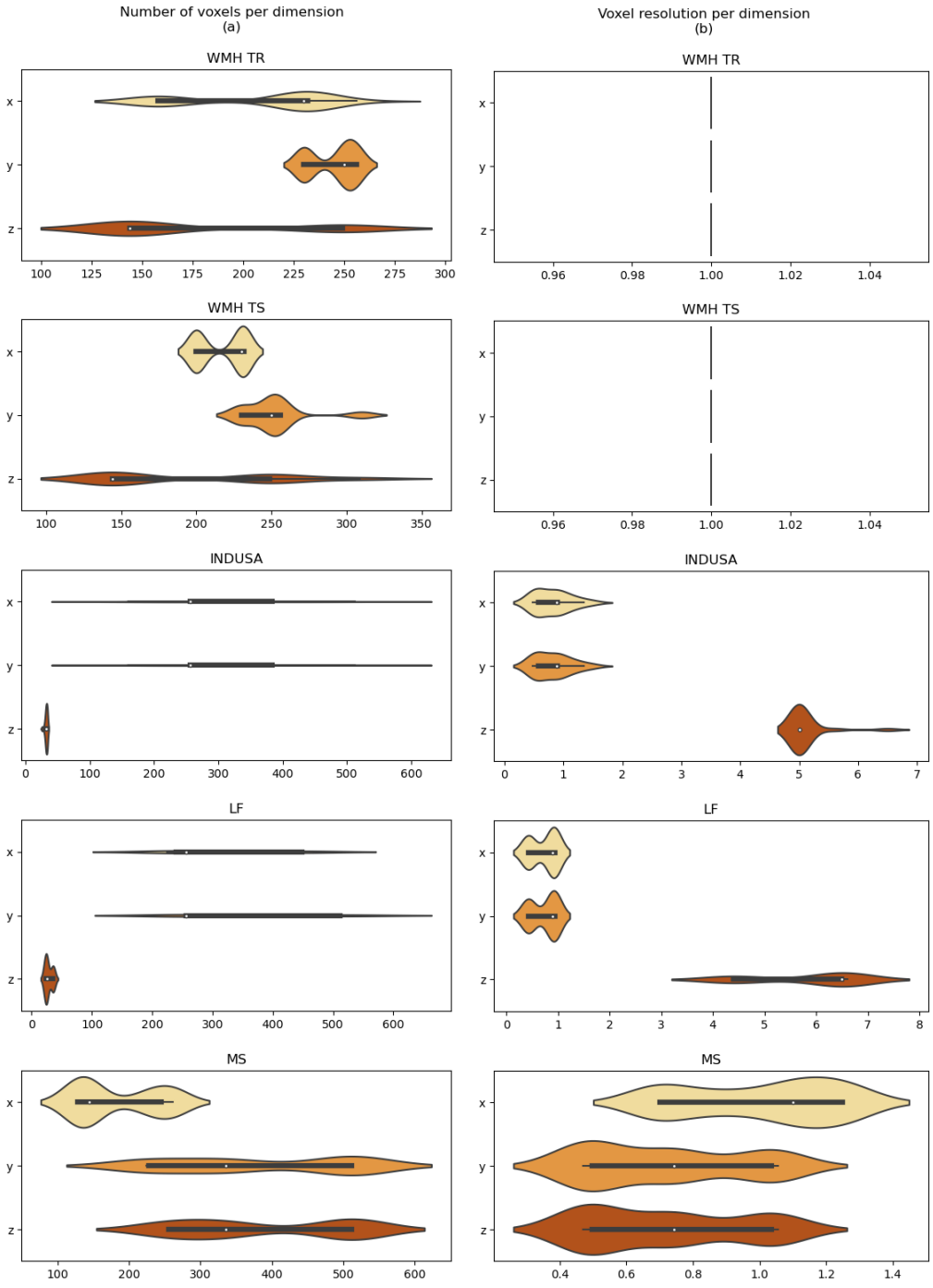


Figure 3.4: Distribution of number of voxels (a) and voxel resolution (b) of each dataset cases, divided into the three available dimensions.

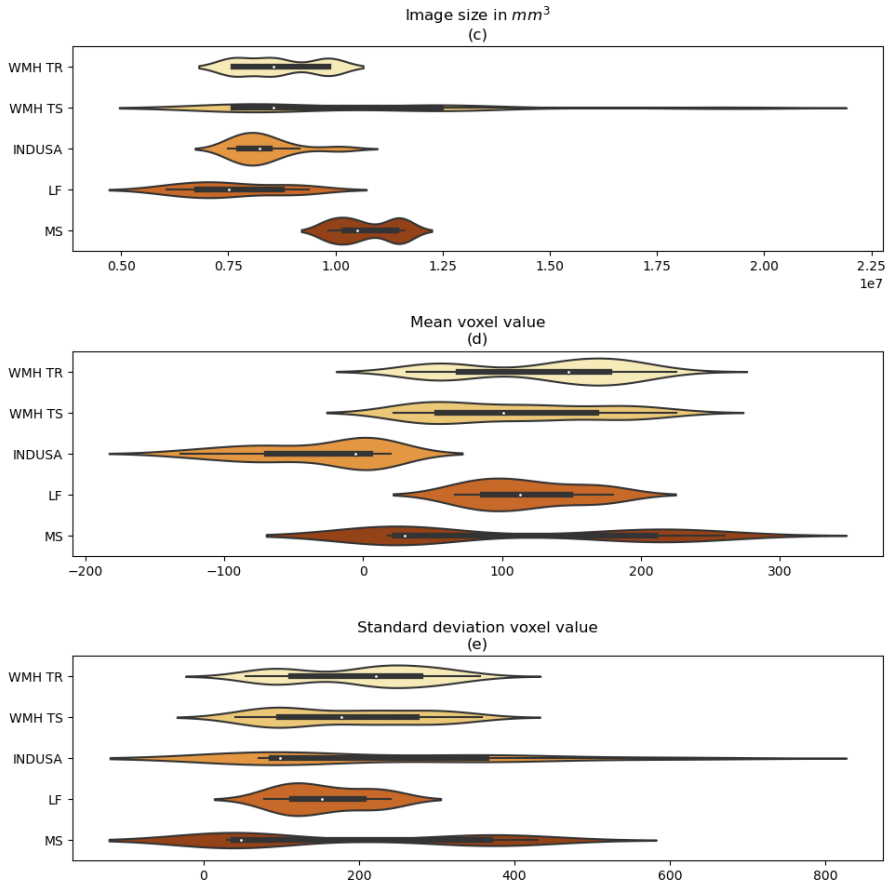


Figure 3.5: For each patient in each dataset, the plotted distribution of the size of the images (c), the mean of the voxels value (d) and the standard deviation of the voxels value (e).

In Figure 3.6, the distribution of the volume of WMH (f) shows that they are present almost equally in the datasets made for the segmentation challenges, but the value is much lower in INDUSA, with very few cases with a lot of lesions. The opposite happens if we consider the volume of other pathologies (g), as INDUSA clearly displays a lot of them, while the challenges datasets have been selected in a way that would exclude them. This fact alone indicates INDUSA as a very good testing set for robustness to "in the wild" conditions, as it is not biased towards WMH. The same can be said when looking at the volume ratio of present other pathologies (i), computed as the ratio between the whole image volume and WMH: INDUSA lesion ratio is much higher in "other pathologies" (i) than in WMH (h).

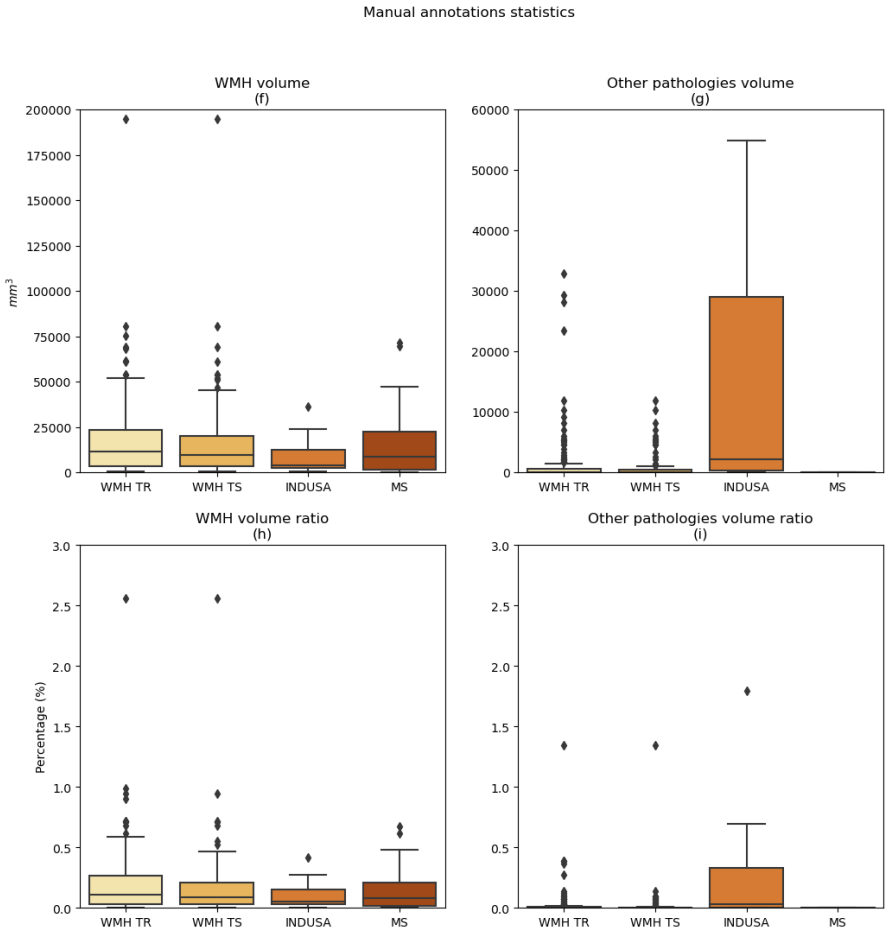


Figure 3.6: The volume of white matter hyperintesities (WMH) across annotated datasets (f), and the ratio between the volume of WMH and the whole image (h). The same information is plotted but with respect to the other pathologies present in the images (g,i).

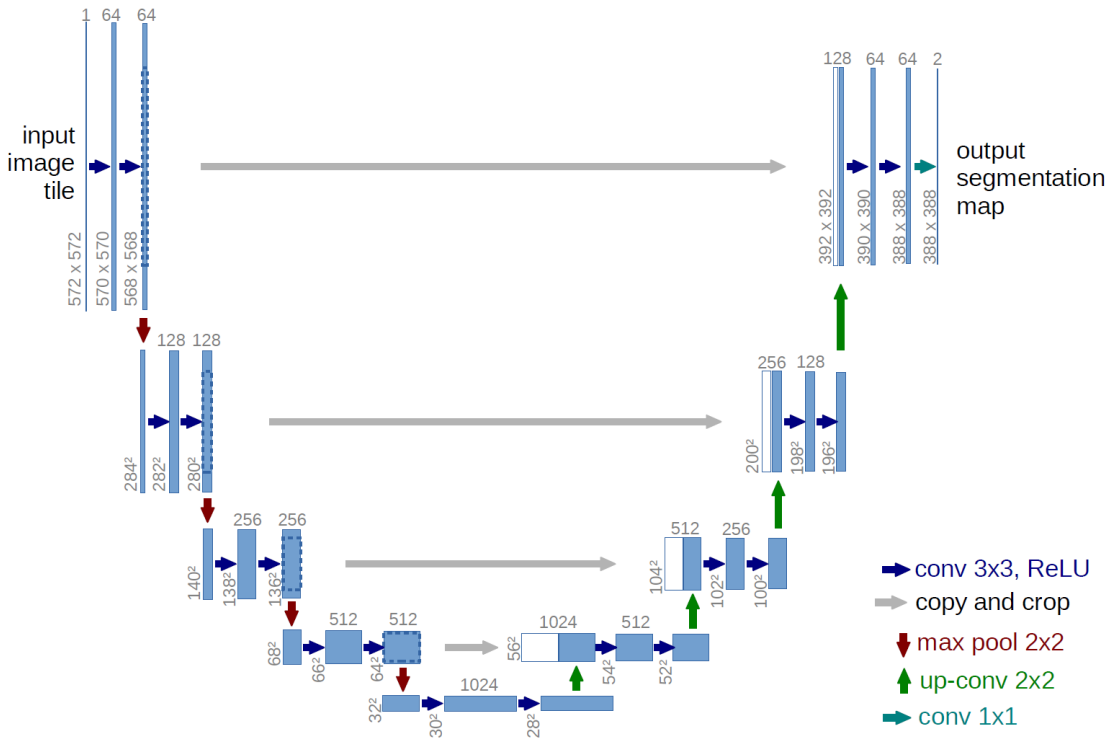


Figure 3.7: Classical U-net architecture [12]

3.2 | Algorithms

3.2.1 | U-Net

The U-Net architecture [65] is the standard de facto for segmentation tasks in medical imaging. It is a deep convolutional neural network that uses a contracting path to capture the context from the image, while a symmetric expanding path adds precise localization. Since 2015, the year the U-Net was published, the architecture has been reimplemented in many forms, by adding information in different ways. Nevertheless, the basic idea stays the same. The goal of U-Net is to be able to train on fewer images, while still yielding precise segmentations, with respect to the previous methods. The original model was developed to work on two-dimensional input, but the three-dimensional version was developed soon after.

The main limitation of standard convolutional networks is the requirement of a large enough training set to achieve good results on a model with a high number of parameters. These networks are usually implemented for classification tasks that do not output

the localization of the class label on each single pixel of the image, but only the class label itself. In order to keep the information that is lost in a common Convolutional Neural Network, researchers came up with the idea of skip connections and, consequently, of the U-Net. Many different U-Net architectures can be implemented and trained with different strategies, as discussed in Section 3.4.

The U-Net is a network made entirely of convolutional layers. It is symmetric, meaning it is made by an *encoder* and *decoder* to extract and reconstruct spatial features. The encoder path is a sequence of downsampling operations that double the number of filters in the convolutional layers each time. The downsampling is performed by two 3×3 convolutions and a max pooling operation with a pooling size of 2×2 and a stride size of 2. The decoder path up-samples the feature map with a 2×2 transposed convolution operation that halves the number of feature channels. This is then followed by two 3×3 convolutions. Again, this sequence is repeated four times. The final segmentation map is generated by performing a 1×1 convolution on the last layer, with a sigmoid activation function. All the other layers use the *ReLU* as the activation function.

$$\text{ReLU}(x) = max(0, x)$$

The *skip connections* are a clever addition that is used to propagate to successive later the output of the upsampling operation. When the encoder performs pooling operations, we lose spatial information. This data is recovered by the skip connections that are present in all four levels of the U-shaped architecture, which we can visualize as *bridges* between the encoder and the decoder path. Particularly, the connections concatenate the feature maps that are generated by the convolutions of the encoder levels, before pooling, and are concatenated with the output of the upsampling operations of the decoder path. The original architecture is presented in Figure 3.7.

To extend the U-Net for volumetric segmentation, the three-dimensional counterpart of convolution and pooling were used instead. However, this greatly increased the number of parameters. To reduce the computational complexity and avoid bottlenecks, the depth of the network was reduced by one and the number of filters was doubled before reaching the pooling layers.

3.2.2 | Fully Convolutional Network Ensembles (FCNE)

The challenge about the segmentation of white matter hyperintesities was won by a team, *sysu_media*, that implemented an ensemble of fully convolutional networks [13]. We can refer to this method as FCNE, which stands for Fully Convolutional Network Ensembles. The final evaluation on the testing dataset is a reported mean *DSC=0.80*

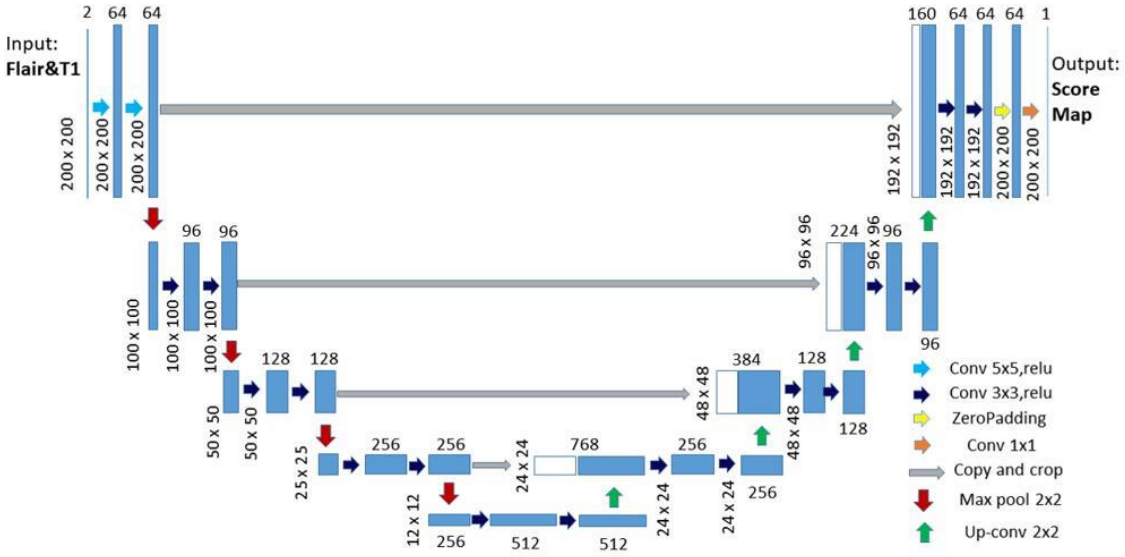


Figure 3.8: 2D Convolutional Network Architecture of the winning method of the MICCAI 2017 WMH Segmentation Challenge. [13]

(0.78-0.82, 95% confidence interval). This is considered the state-of-the-art result. The network architecture by *sysu_media* is a 2D convolutional network architecture. It combines fully convolutional networks [66] with U-nets [12]. The network (Figure 3.8) takes as input two image modalities (T1 and FLAIR) as a two-channel input. It shrinks the spatial dimensions with a down-convolutional part (left side) and expands the score maps with the up-convolutional part. In total, the network has 19 convolutional layers and its final layer is a 1×1 convolution used to map each feature vector to two classes (WMH or not WMH). In order to handle different transformations, this network replaces in the first two convolutional layers the 3×3 kernels, which were used by [67], with 5×5 kernels.

To train the model, the authors used the Dice loss [50], which is useful when the number of positives and negatives are highly unbalanced, as is the case in medical imagining segmentation tasks. The formulation is similar to the one reported in Section 2.5.1 (page 29).

FCNE [13] combines multiple models to "obtain better predictive performance than any of the learning constituent learning algorithms alone". The model ensemble approach has been discussed in more detail in Section 2.5.3.4. The authors trained n U-Net models with the same architecture but trained with different and randomly initialized parameters. Moreover, the data feed in batches was shuffled. At inference time, the n

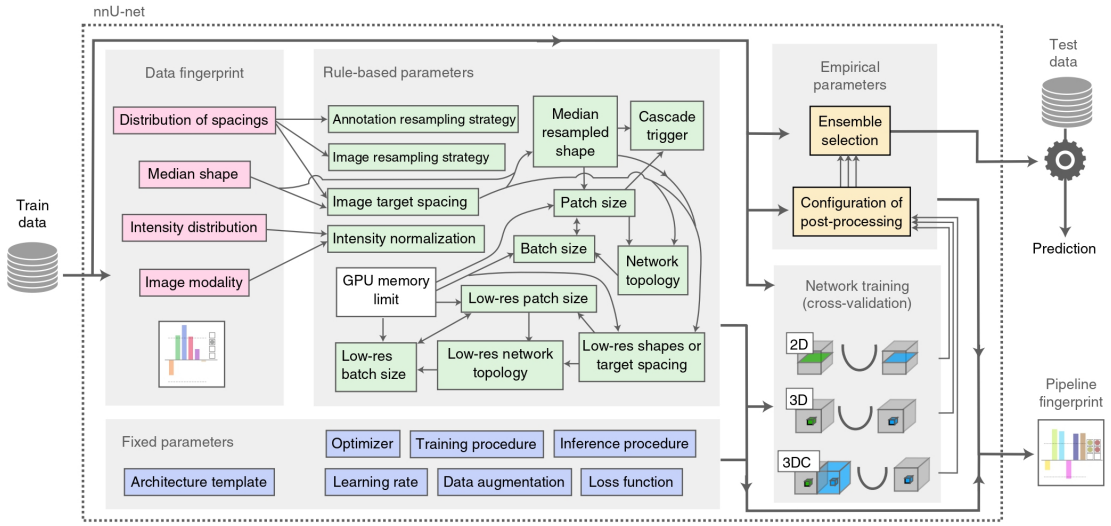


Figure 3.9: NnUnet, for any new given a new segmentation task, extracs dataset properties in the form of a 'dataset fingerprint' (pink). Then , a collection of heuristic rules represents relationships between parameters (depicted as thin arrows). These rules are applied to the fingerprint to deduce the 'rule-based parameters' (in green) that depend on the data. Additionally, there are 'fixed parameters' (in blue) that remain predefined and do not necessitate adjustment. The training involves up to three configurations in a five-fold cross-validation setup. [14]

U-Net models predicted a probability segmentation map per a given test image. The n resulting maps were then averaged to output the final prediction, which was then thresholded to transform the score maps into a binary segmentation map. The threshold was empirically picked.

The last step is post-processing, which reverses any crop or padding to match the original image, or removes any *anatomically unreasonable artifact*. The authors describe an unreasonable artifact as WMH that may not appear in the first or last few axial slices. They tackled this problem with a simple solution: WMH in the first m and last n were considered false positives, and so removed. The parameters m and n were empirically set at 10% of the number of axial slices in the z-direction.

3.2.3 | NnUNet: a deep learning framework

NnUNet [14] is a deep learning-based segmentation method that automatically configures itself. For any new task, the method configures how to perform preprocessing, and chooses a network architecture, a training strategy, and post-processing. NnUNet

stands for "no new U-Net", as it is based on the standard U-Net, but adapts it to the task at hand. NnUNet was presented for the Medical Segmentation Decathlon challenge, where methods were evaluated in ten disciplines with no manual adjustments between datasets allowed. These datasets had distinct entities, image modalities, image geometries and dataset sizes. NnUNet won across most classes and tasks, proving its robust nature.

This method is a comparatively simple U-Net model that contains only minor modifications to the original U-Net, as it adapts its own architecture based on the given image geometries. This is not the only step performed by the nnUNet framework. This method performs preprocessing on the input dataset, like resampling and normalization, then it configures its own training parameters, like loss and optimizer, and how to perform inference. Optionally, it can also perform some form of post-processing.

The training procedure includes five-fold cross-validation on the training set. The training loss is a combination of DSC and cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{DSC}}$$

For a multi-class segmentation task, the cross-entropy loss can be defined as

$$\mathcal{L}_{\text{CE}}\left(y, \hat{y}\right) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log\left(\hat{y}_{i,k}\right),$$

y employs a one-hot encoding method for representing ground truth labels, while $\hat{y}$ stands for a matrix containing predicted values for each class. The variables c and i iterate through all classes and pixels, respectively. The goal of the cross-entropy loss is to minimize the error on a per-pixel basis. However, when there is an imbalance in the distribution of classes, this can lead to an overemphasis on larger objects in the loss function, which in turn can lead to lower-quality segmentation of smaller objects. The DSC can be defined as

$$\mathcal{L}_{\text{DSC}} = -\frac{2}{|K|}\sum_{k\in K}\frac{\sum_{i\in I}\hat{y}_{\sigma}^{k} y_{i}^{k}}{\sum_{i\in I}\hat{y}_{\sigma}^{k} + \sum_{i\in I} y_{i}^{k}}$$

where $\hat{y}_{\sigma}$ is the softmax output of the network.

An epoch is defined as the iteration over 250 training batches, where the number of samples θ in a batch Θ is set based on how many samples can fit in GPU memory.

All experiments were run with the *Adam* optimizer [68] with initial learning rate of 3×10^{-4}. Whenever the performance on the validation set didn't improve by at least 5×10^{-3} within 30 epochs, the learning rate was reduced. If no improvement happened

within the last 60 epochs, the training was stopped, but not before the learning rate had a value smaller than 10^{-6}.

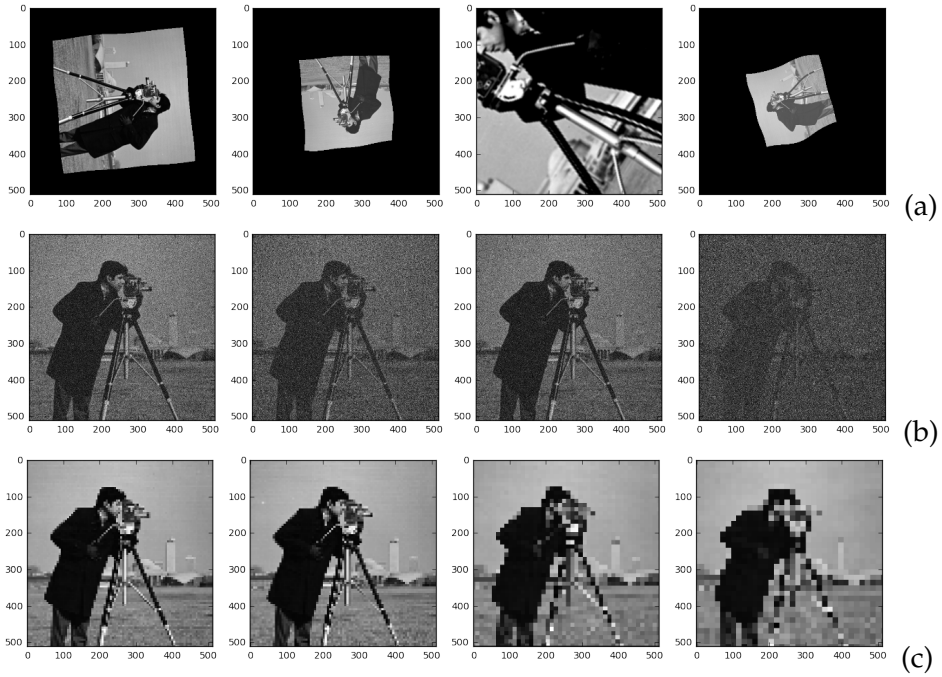


Figure 3.10: Examples of transformations applied by the nnUNet framework. [15] It includes deformations, scaling and rotations (a), noise addition (b), and resampling (c).

3.3 | Data Augmentations

3.3.1 | FCNE data augmentation

The authors of the Fully Convolutional Network Ensembles (FCNE) method point out that pre-processing plays an important role in their overall framework, aimed at uniforming the size of the data, so that it can be fed to the same network and normalize voxel intensity to reduce variation across subjects. To do so, they cropped or padded each axial slice to a uniform size (200×200) and applied Gaussian normalization on the brain voxel intensity. This pre-processing is applied before training and testing. To ensure robustness and invariance, they applied augmentations to the images during training. The augmentation and their parameter ranges are:

- Rotation $[-15°, 15°]$
- Shearing $[-18°, 18°]$
- Scaling $[0.9, 1.1]$

3.3.2 | NnUNet data augmentation

The nnUNet augmentation pipeline applies common image transformations on the fly during training, including:

- Random rotations
- Random scaling
- Random elastic deformations
- Gamma correction augmentation
- Mirroring.

The parameters are not different between datasets, but they differ if the task is trained with a 2D or 3D U-Net. Parameters are available in Appendix B. Examples of possible transformations are shown in Figure 3.10

3.3.3 | MRI-specific data augmentation

The presence of medical imaging artifacts can significantly impact the annotation of medical images, especially in the context of machine learning and computer-aided diagnosis. When training algorithms to recognize abnormalities or specific anatomical structures in MRI images, artifacts can confuse the best radiologists. It 's not surprising that they can also disrupt the learning process. For instance, if an algorithm is exposed to images with motion artifacts, it might incorrectly learn to associate these artifacts with actual anatomical features. Similarly, metallic artifacts could lead to false positives or negatives in the algorithm's predictions. To mitigate these challenges, preprocessing techniques such as artifact correction or removal are often applied to ensure that the algorithm focuses on the true underlying anatomy.

This section reports all the data augmentations used for most of the models trained to solve the task of WMH segmentation. The augmentations come from a different project where they were implemented for the segmentation of the human brain into 132 regions. [16] These data augmentations are tailored to generate realistic distortions or deformations on the MRI scans. In fact, they mimic the artifacts that commonly occur during MRI acquisition. The authors claim that the utilized transforms are 3D and fast so that they can be applied on the fly during training to the original scans, making the network eligible for online learning. The source codes, in MATLAB, of the developed tool are made publicly available at `https://github.com/Mostafa-Ghazi/MRI-Augmentation`.

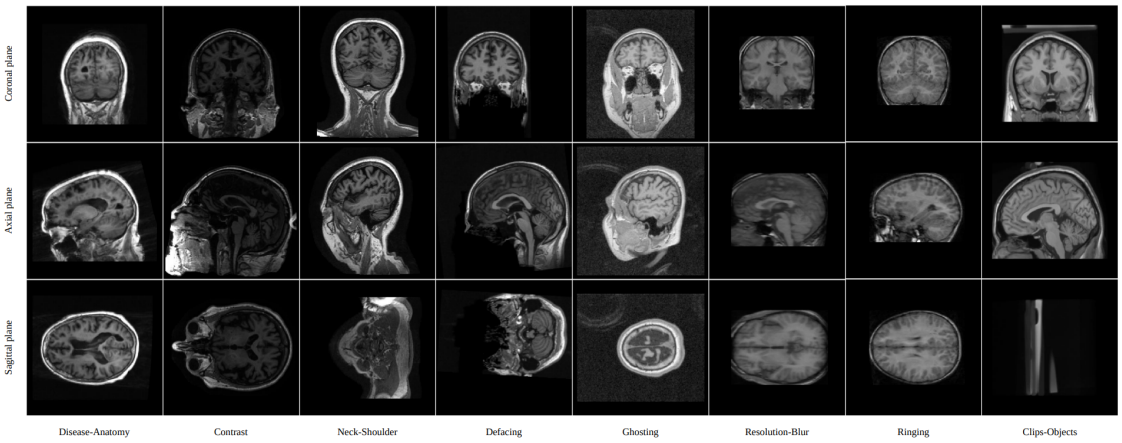


Figure 3.11: Multiview illustration of heterogeneous brain MRI data. [16]. The high variability of the scan parameters indicates the importance of data augmentation for training robust models on the data expanded with different realistic artifacts and changes.

3.3.3.1 | Head Orientation Correction

Using a standard head orientation for all brain scans can improve the training and prediction performance for brain segmentation. RAS orientation is a neurologically preferred convention where the coordinate system in [+X, +Y, +Z] directions is oriented towards the right, anterior, and superior of the head, respectively. The affine matrix stored in the headers of the scans is applied to transform all brain positions in the RAS orientation [69].

3.3.3.2 | Resolution Adjustment

Since different scans can have different isotropic/anisotropic voxel sizes, in order to train the brain segmentation network, we may need to resize the volume dimensions considering the spacing or slice thicknesses. Therefore, we can scale the brain volumes to the nearest even integers, after multiplying the original array dimensions by the corresponding spacing. The resampling can be done by using the linear and nearest-neighbor interpolations for the image and label volumes, respectively.

3.3.3.3 | Contrast Adjustment

Contrast adjustment is an image processing technique that remaps pixel intensities to a stretched display range by sharpening differences between low and high pixel values. To this end, we can normalize the intensity values of the volumes to [0, 1] using the

available dynamic range of each scan, which saturates the high and low intensity values and stretches the distribution to fill the entire intensity range. In addition, the gamma transform [70] is applied to the corrected image volumes with random values in [0.8, 1.2] to augment data with slightly different contrasts.

3.3.3.4 | Volume Rotation

The orientation of the brain can be slightly different for various scans even after the head position correction. Therefore, we can augment the available volumes with randomly rotated ones by allowing the volumes to be rotated about the three perpendicular coordinate axes with an angle randomly chosen in $[-1010$. Linear and nearest-neighbor interpolations are applied to the image and label volumes, respectively.

3.3.3.5 | Skull Stripping and Defacing

Skull stripping [71] and defacing [72] are pre-processing steps that aim at removing facial features or more areas surrounding the brain. These techniques can help with brain extraction and address clinical data privacy issues by securely training models on anonymized data. Hence, to enable the network to focus on learning representations from the brain regardless of the presence of the face, skull, ears, neck, or shoulders, we extract volumes of randomly cropped areas encompassing the brain.

3.3.3.6 | Noise Addition and Multiplication

The MRI images are usually prone to suffer from additive and multiplicative noises such as Gaussian and speckle [73]. Hence, to make the output predictions robust to these types of noises and improve the generalization accuracy, we augment the available intensity volumes with distorted ones by introducing a zero-mean Gaussian noise or a speckle noise with variances randomly chosen in $[0, 0.0001]$ mm^2.

3.3.3.7 | Intensity Inhomogeneity Distortion

Intensity variation or nonuniformity across the image is a common problem in MRI acquisition and can be due to several reasons such as the failure of the radio frequency coil, induced eddy currents, B1 field inhomogeneity, and scanning nonferromagnetic materials [74]. Although preprocessing techniques have been used to estimate the bias field to remove intensity inhomogeneity [75], deep learning methods can obtain superior segmentation results with data augmentation using synthetically introduced intensity inhomogeneity [76]. On this account, we can train the network using data containing

simulated intensity inhomogeneities by multiplying an elliptic gradient field with the brain volumes [77]. Assuming images of the same cubic dimension 256, the gradient field is calculated based on the equation of an ellipse in standard form using the points from a structured rectangular grid with integer values from 1 to 256, centers randomly chosen in [1, 256], and radiuses of 256.

3.3.3.8 | Ringing Artifact Augmentation

The ringing disturbance is a Gibbs phenomenon that occurs as oscillation at boundaries with high contrast transitions. It is caused by under-sampling or truncation of high-frequency components in the image [74]. We augment this artifact by applying the centralized fast Fourier transform (FFT) to the brain volumes in three orthogonal directions and cutting the edges of the k-space [78] at a random integer in [90, 120] along the three axes.

3.3.3.9 | Ghosting Artifact Augmentation

The ghosting noise is a phase-encoded motion that appears as repeated versions of the scanned object in the image [74]. It is caused by periodic movements of tissue or fluid during the scan, affecting data sampling in the phase-encoding direction. We augment this artifact by modulating the k-space lines of each axis differently; we can weight every n-th component of the k-space (FFT) per dimension by a random factor in [0.85, 0.95], where n is a random integer in [2, 4], representing the number of the repeated brains.

3.3.3.10 | Elastic Deformation

Elastic distortion is a state-of-the-art method [79] for expanding the training data by synthesizing plausible transformations of data, and hence, learning shape-invariant representations. Accordingly, we can apply the random elastic deformation algorithm to augment our training data. First, a random 3D uniform displacement field is generated along each axis. The obtained random fields are smoothed using a Gaussian filter with an elasticity coefficient σ randomly chosen in [20, 30] and a square kernel size of $2[2\sigma] + 1$. They are scaled then with a factor α randomly selected in [200, 500], which controls the intensity of the deformation. Finally, a structured rectangular grid with integer values from 1 to 256 is interpolated with the MRI volume to obtain a plausibly deformed volume. Linear and nearest-neighbor interpolations are applied to the image and label volumes, respectively.

3.4 | Architectures

Multiple network architectures were tested with the NnUNet framework. Some of them have also tested with different training strategies.

3.4.1 | 2D U-Net

This architecture is what we could call the *classical U-Net* as it has been implemented by 3.7. In the nnUNet framework, the architectural modifications are negligible. The method's focus is on automating the training pipeline on medical segmentation tasks, which are usually performed on three dimensions. For this reason, both 2D and 3D U-Nets are available. The standard U-Net ReLU activation functions are replaced by leaky ReLUs, defined as

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \text{negative, slope} \times x, & \text{otherwise} \end{cases}$$

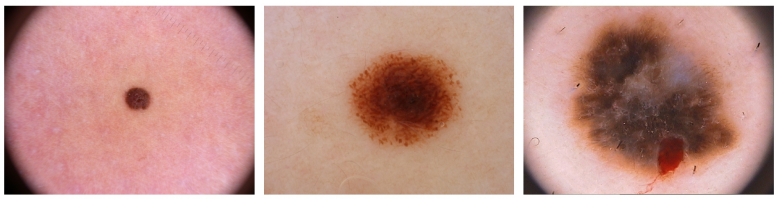
There are also layers that perform instance normalization. [80] Let y_{tijk} denote its $tijk$-th element, where k and j span spatial dimensions, i is the feature channel, and t is the index of the image in the batch. Then instance normalization can be defined as

$$y_{tijk} = \frac{x_{tijk} - \mu_{ti}}{\sqrt{\sigma_{ti}^2 + \epsilon}}, \quad \mu_{ti} = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} x_{tilm}, \quad \sigma_{ti}^2 = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} (x_{tilm} - \mu_{ti})^2.$$

At first glance, employing a 2D U-Net for 3D medical image segmentation seems less than ideal, as it fails to capture important information along the z-axis. Nevertheless, there is evidence indicating that traditional 3D segmentation approaches exhibit reduced performance when dealing with anisotropic datasets. [81]

3.4.2 | 2D MultiRes U-Net

The MultiResUNet [17] is a modification of the standard U-Net. It addresses some flaws of the classical U-Net presented in Section 3.2.1. In different types of medical image segmentation tasks, objects of interest are of irregular and different scales (Figure 3.12). A network should be able to generalize to different objects. The MultiResUNet architecture introduces a new skip connection block and convolution block to address this issue. (Figure 3.14).



(a) (b) (c)

Figure 3.12: Example of variation of scale in medical images from the MultiResUNet paper. [17] The images have been taken from the ISIC-2018 dataset, that shows lesions with small (a), medium (b) and large (c) size.

3.4.3 | 3D U-Net

An architecture that learns over all available dimensions is ideal. However, in reality, GPU memory is often not enough to train deep learning methods. To solve this problem, NnUNet trains on image patches, which is not a problem in most tasks, like the one we are discussing now. Image patch learning limits the field of view of the architecture, so in tasks that rely on contextual information, it's not the right choice. In the case of tasks involving liver images, for example, image patches are not enough to distinguish different liver parts from those of other organs.

NnUNet did not perform any significant architecture modification to standard 3D U-Net. (Figure 3.13)

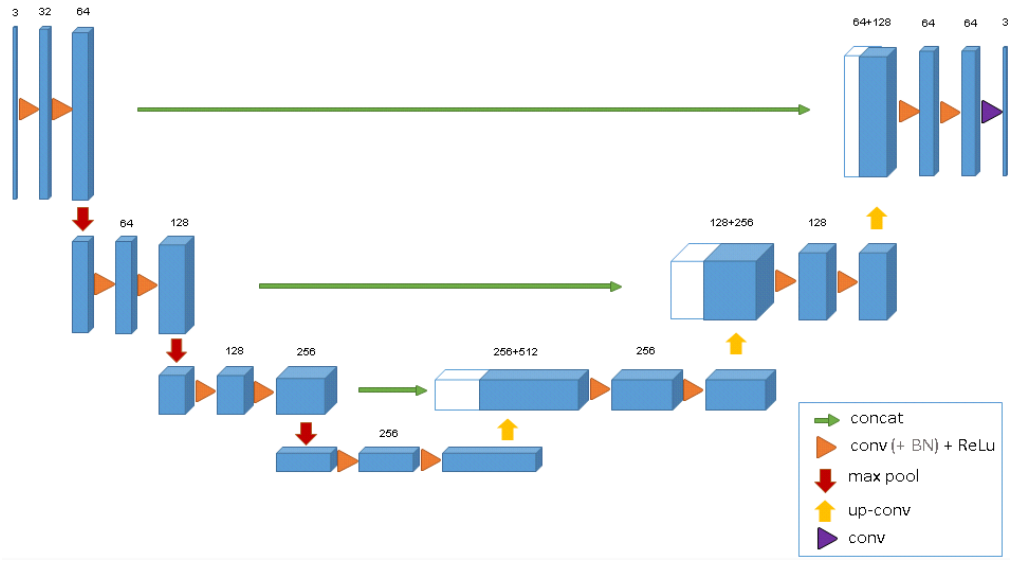


Figure 3.13: The 3D U-Net architecture. [18]

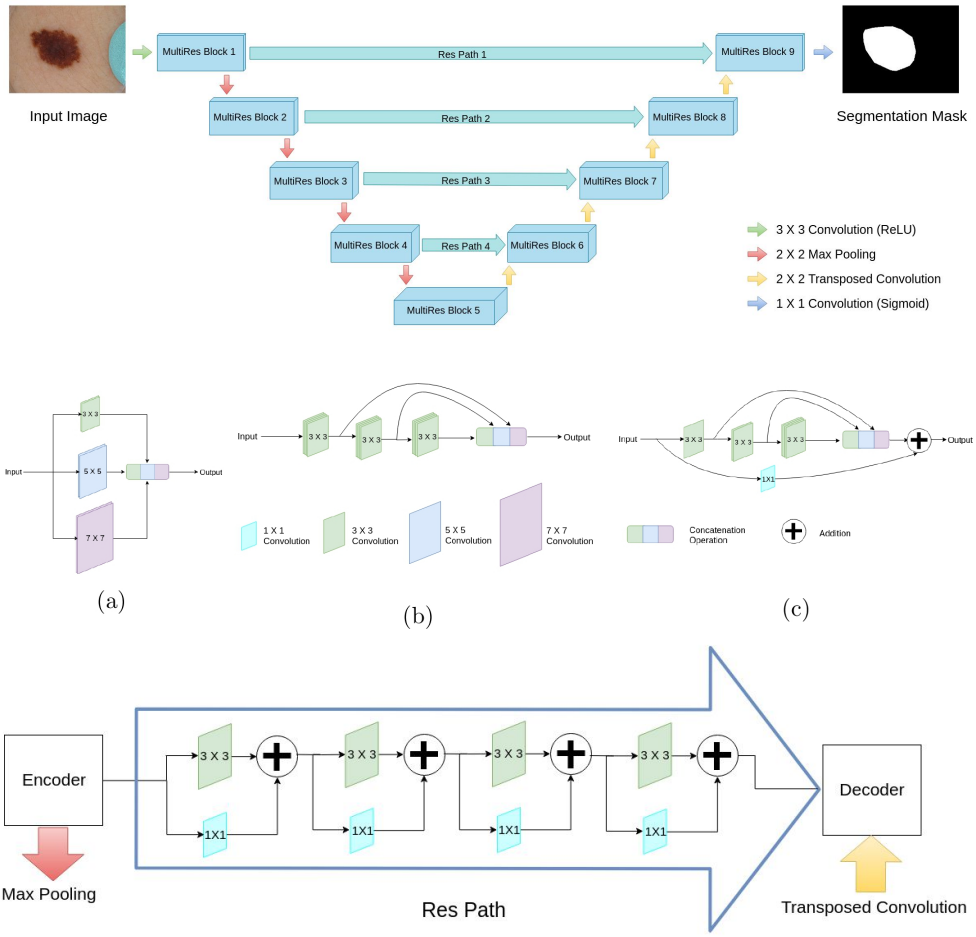


Figure 3.14: The MultiRes architecture and its building blocks. [17] The convolution block is made by 3×3, 5×5 and 7×7 convolutional filters in parallel and concatenating the generated feature maps (a). This allows us to reconcile spatial features from different context size. The bigger and more expensive 5 × 5 and 7 × 7 filters are factorized as a succession of 3 × 3 filters, instead of using the 3 × 3, 5 × 5 and 7 × 7 filters in parallel. In the MultiRes block (c) we we have increased the number of filters in the successive three layers gradually and added a residual connection (along with 1 × 1 filter for conserving dimensions). For the proposed *Res* path, the encoder features are passed through a sequence of convolutional layers. These additional non-linear operations are expected to reduce the semantic gap between encoder and decoder features.

3.5 | Experiments

All experiments were run through the NnUNet framework (see Section 3.2.3). Because NnUNet takes care of the tuning of most hyperparameters that one can choose when training a U-Net, experiments could be focused more on the type of model that could help solve the task rather than tuning parameters for one single architecture. The order in which experiments are presented is not chronological, but it's ordered logically by presenting the ones that are similar or improvement of the previous in a sequential order.

Experiments could be divided into two macro categories: models that had processed the input as a two-dimensional image, and models that did so as a three-dimensional image. The difference between the two approaches is significant. It mostly impacts the cost of training, but also has an influence on which type of information is extracted from the images.

One important remark is that all experiments, except one version of the *3D U-Net*, were performed with the aid of a specific set of data augmentations. These can be found explained in detail in Section 3.3.3. As we will discuss later, data augmentations play a very important role in this task.

Most of the models were trained on one 16GB Quadro RTX 5000 GPU, for at most 1000 epochs. The training time varied on the architecture and on the type of augmentation used, ranging from 2 to 8 days.

The training set is the one presented in Section 3.1.1. A total of 60 FLAIR images are used for training. The exclusion of T1-weighted images is a conscious choice: the final solution to the problem should include as little information as possible, as explained in Section 1.2.

3.5.1 | Experiments with 2D convolutions

The is a big difference in experiments with 2D or 3D convolutions. A three-dimensional kernel is capable of elaborating information in all three possible dimensions of a volume, extracting inter-slice information from adjacent frames.

The decision to sacrifice one dimension might be taken because of multiple reasons. The most common and relevant reason is the computational cost of 3D convolutions, which greatly increases the number of parameters needed to train the same network. For example, a 3D kernel of size $3 \times 3 \times 3$ has 81 parameters. Compared to its 2D equivalent of size 3×3 with only 9 parameters, the straightforward conclusion is that networks computing 2D convolutions have a much easier time computing training passes

and updating parameters. However, this comes at the expense of completely losing information along one axis. Because of the MRI acquisition process, the information along the axial plane is usually less. The two dimensions used for 2D convolutions are commonly the first two, resulting in elaborating the information on the axial slices. Nonetheless, this is not the only way in which we can apply 2D kernels.

3.5.1.1 | 2D U-Net with MRI-specific data augmentation

The first experiments with 3D convolutions and MRI-specific data augmentation already showed that this setup benefits greatly from the new augmentations. How much improvement do we get when including the third dimension in the training? Can we train a 2D U-Net to save resources and obtain the same performance? To answer these questions, the 2D U-Net was also trained with MRI-specific data augmentation. No baseline without this set of processing functions has been trained.

3.5.1.2 | 2D MultiRes U-Net with MRI-specific data augmentation

A 2D MultiRes U-Net as described in Section 3.5.1.2 was trained with the nnUNet framework. This network, like the others, was trained with MRI-specific data augmentation.

3.5.1.3 | 2.5D MultiRes U-Net

The 2.5D MultiRes U-Net is an experiment very similar to the *2D MultiRes U-Net* in Section 3.5.1.2, except that one model was trained for each of the views of the brain (see Section 2.1). In this way, even if each model was still trained on two-dimensional slices, the ensemble of the three models would give more information about the brain as a whole. The hope was that such a model would have been more robust to domain shift.

Unfortunately, it was hard to make converge the models trained on the coronal and the sagittal view at training time, with what was probably a problem of *gradient explosion*. We didn't spend more time trying to improve this three-view model ensemble because we preferred to work more on validation, and it didn't seem a worthwhile task to solve. This is also motivated by the fact that WMH are very hard to see on non-axial views even by trained radiologists, who usually only work on the axial view when annotating them. This model is not included in the final inference results evaluation, so it will not be discussed further.

3.5.2 | Experiments with 3D convolutions

3.5.2.1 | 3D U-Net

A U-Net with 3D convolutions was trained with the nnUNet framework, with its default data augmentations functions and parameters. We can assume this model is the baseline model. No other parameter was trained with the nnUNet base augmentation set because it was shown in early experiments that MRI-specific data augmentation yielded better results across every experiment.

3.5.2.2 | 3D U-Net with MRI-specific data augmentation

This model is the same as in Experiment 3.5.2.1, but with the added MRI-specific data augmentations. The set of parameters for both the 2D and 3D versions of augmentations is reported in Appendix B.

3.5.2.3 | 3D U-Net with MRI-specific data augmentation and enhanced resampling

Because the Experiment 3.5.2.2 failed mostly in under segmenting white matter hyperintesities around their surface, the same model was retrained with a change in the parameters regarding the resampling function. The motivation behind this choice is given by the fact that undersampling might be due to the model being conservative about its predictions, particularly on images with lower resolution. If the model is trained on high-resolution images, it will be careful when classifying voxels around the surface of a lesion, as those are the most difficult to assess. Increasing the resampling ranges to downgrade the image resolution is a strategy that aims at forcing the network to give more importance to voxels on the surface.

3.5.2.4 | 3D U-Net with MRI-specific data augmentation and contralateral information

Differently from other types of brain pathologies, white matter hyperintesities are very symmetrical. This means that if a lesion appears in a hemisphere, it typically appears in a similar capacity in the *contralateral* hemisphere. The word *contralateral* (Latin: contra, against;) can be used to refer to any part of the human body that has an opposite and equal part in the other side of the body, as is the case with the human brain. Because white matter hyperintesities appear symmetrical, we can define the *contralateral information* as the same image modality, but flipped along the horizontal axis. This flipped

image modality is then fed as a second, separate modality to the network. To make this possible, we need to register the two images, the original and the *flipped* or *contralateral* image, to each other. The ground truth segmentation map needs also to be transformed. We register the two images with a function that finds the minimum difference between two volumes when transposing and rotating along the first two dimensions, meaning we don't transform along the third dimension. When the minimum is found, the optimal transformation is applied to the original image and its label. The contralateral image is the transformed image but flipped along the first axis.

Name	Method
FCNE	Fully Convolutional Network Ensembles (FCNE)
2D-unet	2D U-Net, trained with nnUNet
2D-multires-unet	2D MultiResUNet, trained with nnUNet and MRI-specific data augmentation
3D-unet-noDA	3D U-Net, trained with nnUNet and standard nnUNet data augmentation
3D-unet	3D U-Net, trained with nnUNet and MRI-specific data augmentation
3D-unet-resampl	3D U-Net, trained with nnUNet and MRI-specific data augmentation, with enhanced resampling
3D-unet-contralat	3D U-Net, trained with nnUNet and MRI-specific data augmentation, with contralateral information as additional modality

Table 3.3: Summary with the experiments abbreviated name and their specific set up. For all experiments, the training set is the same as used for the MICCAI 2017 WMH Segmentation Challenge, excluding T1-weighted information and relying only on FLAIR, for a total of 60 cases.

4

Results & Discussion

4.1 | Results

The methods presented in Chapter 3 have been tested on the data presented in the same chapter. This section presents the outcomes of these methods through the use of tables or plots. To facilitate a comprehensive understanding of these findings, explanatory comments have been included to translate the presented material into natural language. The structure of this chapter is based on three sections presenting quantitative results for the test sets with an available ground truth annotation, namely MICCAI WMH, INDUSA, and MS. Single tested cases from these datasets and the LF dataset are then shown in Section 4.1.4 for visual inspection and qualitative evaluation.

In Figure 4.1 we can have a first overview of the models' performance on three datasets, where each point represents a tested case and its DSC and VS scores. The model whose cluster of cases is closer and more compact toward the top right corner of the plot is the model that performs best and has more consistent results.

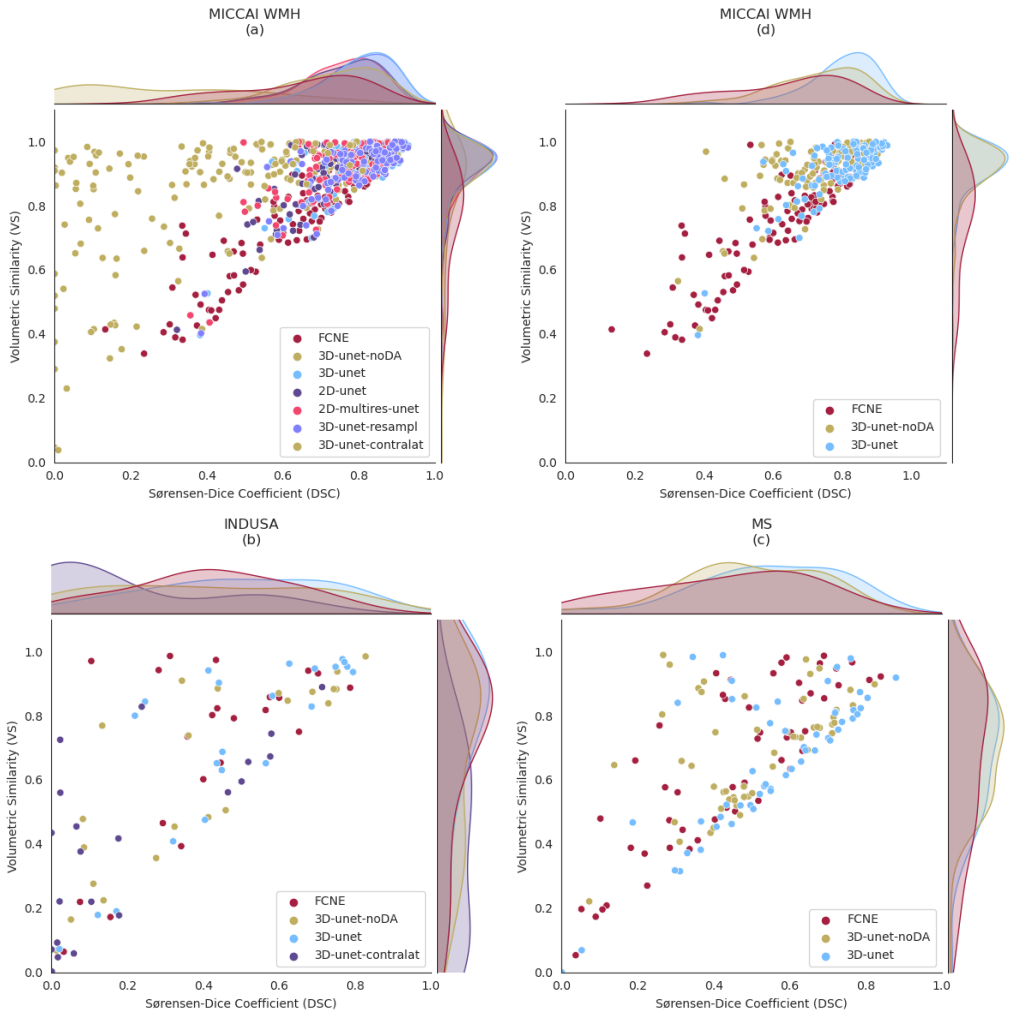


Figure 4.1: Comparison of Volumetric Similarity (VS) and Sørensen–Dice Similarity Coefficient (DSC) on all datasets, except LF, where each point represents a tested case.

4.1.1 | MICCAI WMH

As already explained in Section 3.2.2, the winners of the MICCAI 2017 WMH Segmentation Challenge did so with DSC= .80. The winner of the challenge was chosen based on four other metrics. However, for the reasons shown in Section 2.4.1, only Sørensen–Dice Similarity Coefficient (DSC), Volumetric Similarity (VS) and Hausdorff Distance (HDD) will be considered for evaluating the presented methods. Additionally, the winners' method, *FCNE*, won the challenge with the aid of an additional modality (T1-weighted) that has *not* been used to train any of the methods used for testing in this thesis, including *FCNE*.

	DSC (↑)	VS (↑)	HDD (↓)	Precision (↑)	Recall (↑)
FCNE	.64 (±.17)	.76 (±.17)	6.74 (±29.88)	.54 (±.20)	**.83** (±.**12**)
2D-unet	.75 (±.11)	.91 (±.11)	27.14 (±**1.84**)	.78 (±.11)	.74 (±.15)
2D-multires-unet	.75 (±.11)	**.92** (±.**09**)	32.16 (±2.49)	.74 (±.**10**)	.78 (±.14)
3D-unet-noDA	.73 (±.13)	.91 (±.**09**)	33.08 (±2.55)	.74 (±.13)	.74 (±.15)
3D-unet	**.80** (±.**10**)	**.92** (±.09)	28.32 (±16.23)	**.80** (±.**10**)	.81 (±.14)
3D-unet-resampl	.79 (±.**10**)	.91 (±.**09**)	28.95 (±16.68)	**.80** (±.**10**)	.81 (±.14)
3D-unet-contralat	.30 (±.24)	.81 (±.22)	**4.47** (±17.63)	.33 (±.22)	.30 (±.25)

Table 4.1: Results for each tested method on the MICCAI WMH dataset. In bold, the best mean and standard deviation for each metric.

From Table 4.1 we can see that *FCNE* scored DSC=.64 (.47-0.81) when using FLAIR only. All the other presented methods, except for *3D-unet-contralat*, achieved a better result. *FCNE* has a VS=.76 (.59-.93) which was also surpassed by all the other methods. Regarding HDD, *FCNE* scored 6.74 (0-36.62), which is the second best result. On voxel-level Recall *FCNE* is the best-performing method, but second-worst on voxel-level Precision, which means that in the case of *FCNE* the *DSC* score is greatly penalized by the PPV. FCNE is good at identifying a large portion of the relevant voxels, but it also tends to incorrectly classify a significant number of backgrounds as WMH.

The best-performing method based on Table 4.1 is the *3D-unet*. As explained previously in Section 3.5, *3D-unet-noDA* differs from *3D-unet* not because it doesn't apply data augmentation (DA), but because these are not the MRI-specific data augmentations presented in Section 3.3.3. All the methods have used MRI-specific data augmentation during training, except for *3D-unet-noDA* and *FCNE*. Given that *3D-unet* is the best-performing method, we most often compare it with *FCNE* and its counterpart *3D-unet-noDA*. If *3D-unet* is consistently better than *FCNE* and *3D-unet-noDA*, we can conclude that a nnUNEt model trained with MRI-specific data augmentation is the model, among the ones tested, that confirms our research hypothesis.

Visually, the results in Table 4.1 can be shown in the form of violin plots, such as in Figure 4.2. While the performance of *3D-unet* and *3D-unet-noDA* seems similar on VS and HDD, the great variability in DSC of *3D-unet-noDA* shows that *3D-unet* have much more consistent results. *FCNE* performance is far from the other two presented methods. A similar plot about the Sørensen–Dice Similarity Coefficient can be found in Appendix C, in Figure C.2.

4.1.1.1 | Site analysis

The MICCAI WMH test set cases were acquired from five different sites. (Section 3.1.1). Of these five, only three had acquired images that have been used to train the presented methods. To test the robustness of the models on out-of-distribution data, compare the performance of each site separately. From Table 4.2, we can see that *FCNE* performed best only in HDD on *UTR*. *3D-unet-noDA* performed better in *GE1* on VS. For all the other metrics and sites, *3D-unet* was the best-performing method. The complete table with all trained models can be found in Appendix C as Table C.1.

In Figure 4.3 we can visualize the distribution of the samples over DSC and VS. All methods exhibit great variability in the three scanners seen during training, namely *UTR*, *SIN*, and *GE1*. Most of the methods display the same variability over the unseen scanners, except for *3D-unet* and *3D-unet-noDA* on VS. Compared to the other methods, which can be found in Appendix C in Figure C.3, *3D-unet* and *3D-unet-noDA* performance is very similar. However, *3D-unet* is more consistent and yields better scores, overall. All methods seem to suffer greatly when the domain shifts to scanners unseen during training (in red), except for *3D-unet*.

Looking at the distribution of HDD in Figure 4.4, we can notice that while all methods have the same score distribution on all seen scanners, *FCNE* fails more on *GE3*. Regarding *GE1*, we can see that it follows the same distribution of *GE3*: even if images from this set were unseen during training, they come from the same site (VU Amsterdam) and same scanner (GE Signa HDxt), but they were acquired at different resolutions (1.5T vs 3T). This similarity is not displayed when considering DSC and VS.

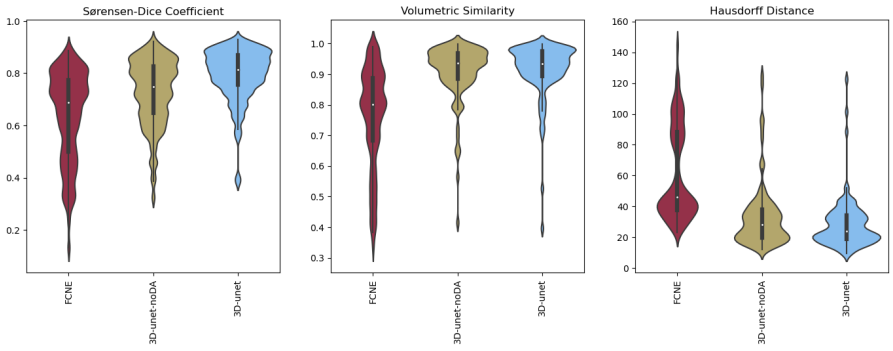


Figure 4.2: Violin plot on the MICCAI WMH dataset, showing the distribution of Sørensen–Dice Similarity Coefficient (DSC), Volumetric Similarity (VS), and Hausdorff Distance (HDD) on all cases. Figure C.1 in Appendix C shows the violin plots for all tested methods.

Site	n	Model	DSC (↑)	VS (↑)	HDD (↓)
UTR	30	FCNE	.67 (±.17)	.86 (±.13)	**4.57** (±8.85)
		3D-unet-noDA	.73 (±.11)	.88 (±.10)	34.73 (±**1.97**)
		3D-unet	**.80** (±**.09**)	**.93** (±.08)	31.28 (±9.87)
SIN	30	FCNE	.71 (±.14)	.80 (±.14)	39.49 (±7.44)
		3D-unet-noDA	.80 (±.13)	**.92** (±.12)	26.14 (±**1.07**)
		3D-unet	**.82** (±**.12**)	.91 (±.12)	**24.47** (±9.98)
GE3	30	FCNE	.55 (±.19)	.66 (±.18)	93.96 (±13.72)
		3D-unet-noDA	.68 (±.12)	**.91** (±**.08**)	25.27 (±**8.48**)
		3D-unet	**.78** (±**.11**)	**.91** (±.09)	**23.53** (±**8.48**)
GE1	10	FCNE	.59 (±.17)	.69 (±.15)	93.96 (±37.64)
		3D-unet-noDA	.64 (±.11)	**.93** (±**.03**)	43.87 (±37.37)
		3D-unet	**.78** (±**.09**)	.92 (±.04)	**39.48** (±39.40)
PHI	10	FCNE	.64 (±.15)	.73 (±.13)	52.14 (±22.64)
		3D-unet-noDA	.74 (±.12)	**.93** (±**.06**)	61.62 (±36.91)
		3D-unet	**.79** (±**.08**)	**.93** (±**.06**)	**34.19** (±**2.95**)

Table 4.2: Results for each tested method on the MICCAI WMH dataset, divided by the tested site. In bold, the best mean and standard deviation for each metric. The complete table with all trained models can be found in Appendix C as Table C.1.

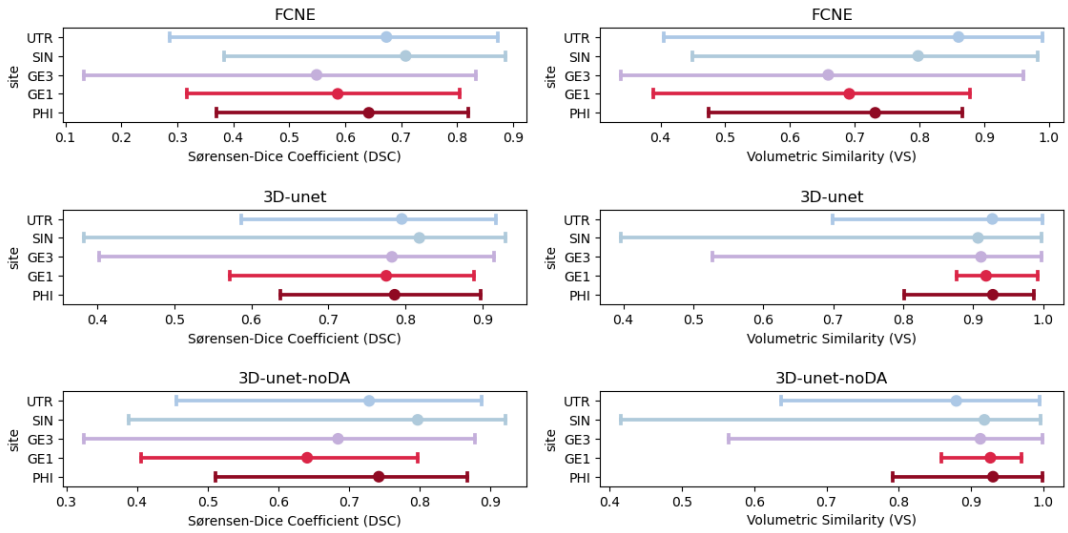


Figure 4.3: Distribution of Sørensen–Dice Similarity Coefficient (DSC) and Volumetric Similarity (VS) for each tested model on sites from the MICCAI WMH dataset. In red, sites unseen during training. The complete figure with all trained models can be found in Appendix C in Figure C.3.

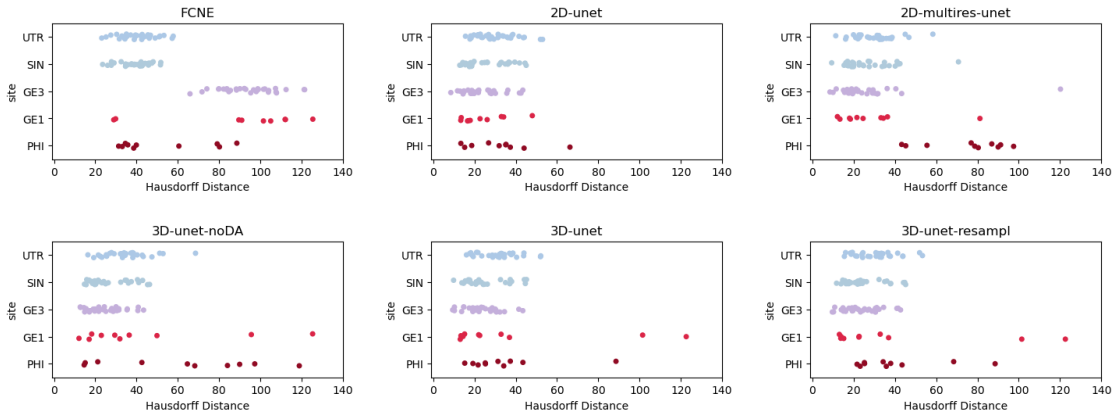


Figure 4.4: Hausdorff Distance distribution for each tested model on the MICCAI WMH test set, divided by site. In red, the sites unseen during training. For each point representing a case, its HDD is better the closer it is to the left axis.

4.1.1.2 | White matter hyperintesities and other abnormalities volume analysis

While evaluating the performance of *3D-unet* on the MICCAI WMH test set, one important question was asked: which are the samples that got a very low DSC in the "tail" of the violin plot in Figure 4.2? From the image is clear that most cases got at least a DSC=.60, while just a handful of them got very low scores. The same can be said about VS and HDD, which the phenomenon is even more advanced.

To answer this question, we need to introduce two new concepts, the *presence of other abnormalities* and *white matter hyperintesities load*.

The nature of white matter hyperintesities is relevant to the task. They are a common finding in most aging patients and are usually not symptomatic. However, a doctor requests MRI for their patient when there is a suspected case that requires imaging. For this reason, it is often the case that along WMH there may be *other abnormalities* in the brain tissues. We can divide the test set into those presenting abnormalities different than WMH, along the WMH themselves, and those presenting little or no abnormalities. We will say that cases with one or more abnormalities with a volume equal to or greater than $1mL$ belong to the *WMH+OA* subset. The other cases belong to the *WMH* subset.

Additionally, white matter hyperintesities may be a common finding, but their distribution is not equal on the whole human population. Many patients may present just a few *mm* wide lesions, while others may be affected more extensively. To measure this concept, most clinicians use the Fazekas scale, described in Section 2.1.2. We can have a more precise measurement by using the voxels in our segmentation maps and mea-

	WMH	WMH + OA
high WMH load	x WMM load $>= 10mL$ and x OA volume $< 1mL$	x WMM load $>= 10mL$ and x OA volume $>= 1mL$
low WMH load	x WMM load $< 10mL$ and x OA volume $< 1mL$	x WMM load $< 10mL$ and x OA volume $>= 1mL$

Table 4.3: Given a test case x, we can assign it to one of four subsets based on some conditions, where *OA* stands for *other abnormalities* that are not WMH.

		n	WMH volume (range, mL)	OA volume (range, mL)
WMH+OA	high WMH load	14	$[10.82 - 195.05]$	$[1.06 - 14.92]$
	low WMH load	7	$[0.9 - 9.45]$	$[1.03 - 10.34]$
WMH	high WMH load	39	$[10.85 - 69.27]$	$[0.0 - 0.88]$
	low WMH load	50	$[0.8 - 9.47]$	$[0.0 - 0.86]$

Table 4.4: Number and volume ranges for the MICCAI WMH subsets created by conditions defined in Table 4.3

suring the lesions' volume. Then, given a threshold, we can divide the tested images into two sets, those with a *high load* of WMH and those with a *low load*. For the purpose of this thesis, the threshold has been set to $10mL$, based on a visual examination of the samples.

Based on the two previous facts, the MICCAI test set has been divided into two subsets, *WMH+OA* and *WMH*. Each of these has been further divided into two other subsets, *high* and *low* WMH load. This subdivision is summarized in Table 4.3. The resulting average volume of WMH and OA per each set is presented in Table 4.4

In Table 4.5 we can see how the three main methods that we are analyzing perform on WMH only, with little or no other abnormalities. *3D-unet* is the best-performing method in mean and standard deviation, both on high WMH load and low WMH load subsets. However, on the low load set, the performance drop is significant, going from a DSC= .87 (.83-.91) on the high load set to DSC= .74 (.65-.83).

When considering only samples that present some sort of other abnormality, *3D-unet* is still the overall best method. There is also a drop from DSC= .87 (.83-.91) on the high WMH only set to DSC= .81 (.67-.95). The respective low load set also presents a significant drop in DSC.

	High WMH load		Low WMH load	
Methods	DSC (↑)	VS (↑)	DSC (↑)	VS (↑)
FCNE	.78 ± .06	.88 ± .08	.50 ± .15	.65 ± .15
3D-unet-noDA	.82 ± .06	.94 ± .05	.65 ± .12	.89 ± .10
3d-unet	**.87** ± **.04**	**.95** ± **.04**	**.74** ± **.09**	**.90** ± **.09**

Table 4.5: Results on the *WMH* subset of the MICCAI WMH test set. In bold, the best mean and standard deviation for each metric.

	High WMH load		Low WMH load	
Methods	DSC (↑)	VS (↑)	DSC (↑)	VS (↑)
FCNE	.75 ± **.12**	.84 ± **.13**	.63 ± .10	.76 ± .14
3D-unet-noDA	.77 ± .13	.89 ± .15	.72 ± .07	**.89** ± **.08**
3d-unet	**.81** ± .14	**.89** ± .16	**.75** ± **.05**	.88 ± .09

Table 4.6: Results on the *WMH+OA* subset of the MICCAI WMH test set. In bold, the best mean and standard deviation for each metric.

Table 4.7: Considering *WMH+OA* the cases where at least another abnormality is present and exceeds a volume of $1mL$, we can divide the MICCAI WMH dataset in subset *WMH+OA* and subset *WMH*. Additionally, each subset can be further divided into *high* and *low load* of WMH volume, where a case is considered *high load* if it exceeds $10mL$.

From Figure 4.5 we can visualize these observations about DSC. While all methods suffer when the load of WMH is low, *FCNE* is the one that shows the greatest variability, with some samples achieving very low scores. *3D-unet* is very consistent on samples with high WMH load, even in the presence of OAs, except for a few cases.

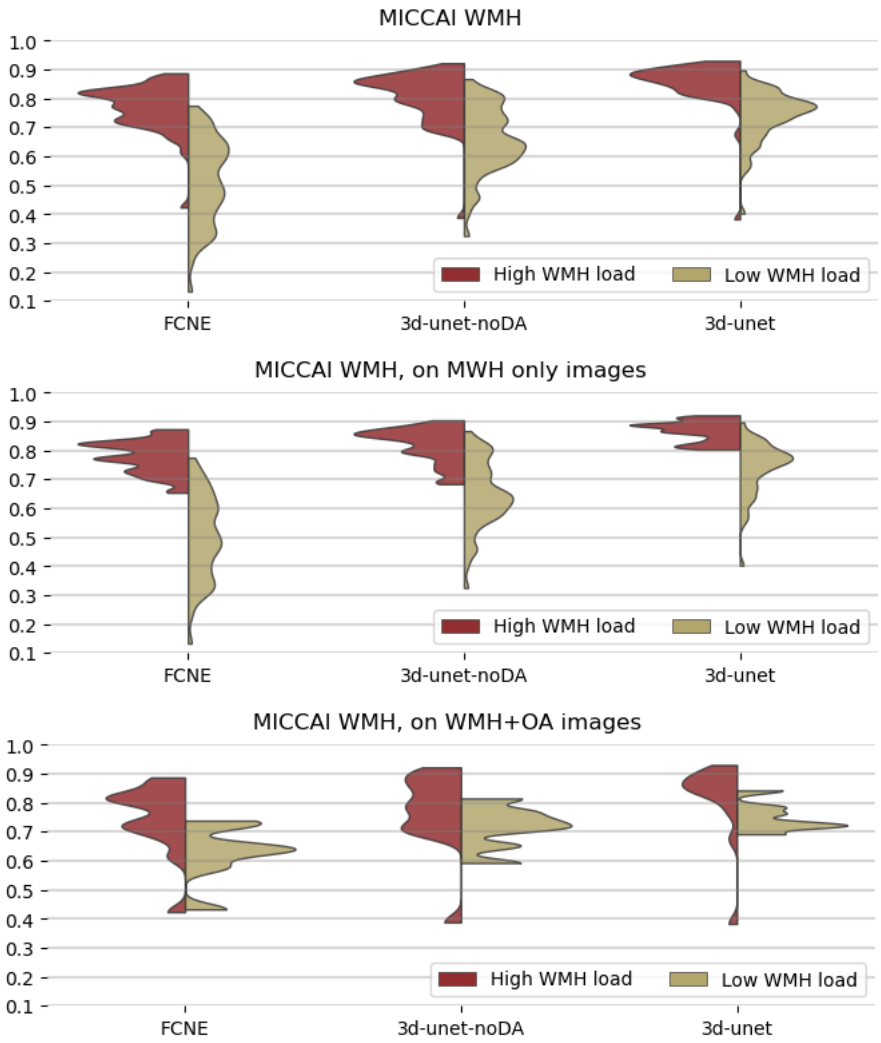


Figure 4.5: Sørensen–Dice Similarity Coefficient distribution based on white matter hyperintesities (WMH) load. In this thesis, a case is considered to be *high WMH load* if the volume of the lesions exceeds $10mL$.

	DSC (↑)	VS (↑)	HDD (↓)	Precision (↑)	Recall (↑)
FCNE	0.42 (±**0.21**)	47.32 (±15.08)	**0.41** (±**0.23**)	**0.54** (±0.23)	0.71 (±**0.28**)
3D-unet-noDA	0.41 (±0.27)	37.53 (±**13.11**)	0.51 (±0.34)	0.43 (±0.22)	0.63 (±0.29)
3D-unet	**0.49** (±0.23)	33.72 (±14.33)	0.58 (±0.25)	0.52 (±0.25)	**0.72** (±0.28)
3D-unet-contralat	0.20 (±0.24)	**51.94** (±23.95)	0.44 (±0.36)	0.16 (±**0.21**)	0.40 (±0.29)

Table 4.8: Results for each tested method on the INDUSA dataset. In bold, the best mean and standard deviation for each metric.

4.1.2 | INDUSA

We can follow the same reasoning, as we did for the MICCAI WMH test set, for the INDUSA test set. In this set, the number of samples is much lower ($n = 22$), but it is completely out-of-distribution and the best scenario for testing the robustness of the presented methods.

In Table 4.8 we can see that the average scores are much lower than the MICCAI WMH test set. Here, there is no clear "winner" method: *3D-unet* has the best DSC, with an increase of 14.3% with respect to *FCNE*. On the MICCAI WMH test set, the increase in DSC was 20.0%. However, on MICCAI WMH *3D-unet* was much more confident, with only .10 in standard deviation, compared to the .23 on INDUSA.

FCNE is the best or second-best performing across all other metrics, but this method has a huge range in variability, indicating that the performance is not robust over all 22 samples.

4.1.2.1 | White matter hyperintesities and other abnormalities volume analysis

As discussed in Section 3.1.2, when presenting the INDUSA test set, this set has a very high number of cases with abnormalities other than WMH. The subdivision in subsets is important, given the results on the MICCAI WMH, to see if the methods have much worse outcomes because they have many OAs and low WMH load, or because they are not robust to change in distribution. The number of samples and their WMH load ranges are reported in Table 4.9.

From Table 4.10, we see that all methods have improved performance on the high WMH load set with no OAs. However, the set is made of only two samples. When considering samples with a low load of WMH, and *3D-unet* in particular, the performance drop in DSC is 23 points, confirming what was already said for MICCAI WMH. The same conclusion can be made by looking at Table 4.11, but here the performance drop in DSC is huge: from DSC=.72 to DSC=.31, a staggering 41 points difference.

		n	WMH volume (range, mL)
WMH+OA	high WMH load	2	$[19.71 - 23.76]$
	low WMH load	10	$[0.4 - 4.63]$
WMH	high WMH load	3	$[15.74 - 36.15]$
	low WMH load	7	$[1.83 - 12.66]$

Table 4.9: Number and volume ranges for the MICCAI WMH subsets created by conditions defined in Table 4.3.

	High WMH load		Low WMH load	
Methods	DSC (↑)	VS (↑)	DSC (↑)	VS (↑)
FCNE	.70 ± .09	.89 ± .04	.46 ± **.12**	**.86** ± **.10**
3D-unet-noDA	.77 ± .05	.92 ± .06	.50 ± .17	.69 ± .23
3D-unet	**.78** ± **.01**	**.96** ± **.02**	**.55** ± .16	.72 ± .24

Table 4.10: Results on the *WMH* subset of the MICCAI WMH test set. In bold, the best mean and standard deviation for each metric.

	High WMH load		Low WMH load	
Methods	DSC (↑)	VS (↑)	DSC (↑)	VS (↑)
FCNE	.53 ± .18	.68 ± .10	.27 ± .18	.56 ± .34
3D-unet-noDA	.69 ± .09	.89 ± **.06**	.17 ± **.15**	.44 ± **.28**
3D-unet	**.72** ± **.04**	**.89** ± .09	**.31** ± .18	**.61** ± .33

Table 4.11: Results on the *WMH+OA* subset of the MICCAI WMH test set. In bold, the best mean and standard deviation for each metric.

Table 4.12: Using the same assumptions as for the MICCAI WMH dataset and considering *WMH+OA* the cases where at least another abnormality is present and exceeds a volume of $1mL$, we can divide INDUSA in subset *WMH+OA* and subset *WMH*. Additionally, each subset can be further divided into *high* and *low load* of WMH volume, where a case is considered *high load* if it exceeds $10mL$.

4.1.2.2 | Other abnormalities analysis by pathology type

Seeing that the presence of abnormalities affects performance, we want to see if different pathologies affect the methods performance in the same way. We can categorize them in two sets, *hyperintense* and *not hyperintense* pathologies. (Table 4.13) The last set is made by abnormalities that show in black on FLAIR, or in the same color of healthy brain tissue, that is gray in FLAIR. Hyperintense signal on FLAIR is white, so the same as white matter hyperintesities. If the method learns the spatiality and shape of lesions,

it will not be confused by the presence of another hyperintense signal. If, instead, the model learns only to segment white shapes, it means that is doing something more similar to thresholding.

Pathology	FLAIR Appearance
White matter hyperintesities	Hyperintense
Gliosis	Hyperintense
Edema	Hyperintense
Infarct, Acute	Hyperintense
Tumor Cyst	Hyperintense or hypointense
Infarct, Hyper-Acute	Isointense or slightly hyperintense
Infarct, Chronic	Hypointense
Tumor Solid	Hypointense, isointense, or hyperintense

Table 4.13: FLAIR MRI Appearances of other abnormalities present in INDUSA.

Based on the type of abnormality that is present in INDUSA cases along WMH, we can establish if the methods are learning from the data the semantic meaning behind WMH lesions, and not simply their color. Looking at Table 4.14, the first thing that the reader might notice is that the *n* columns do not add up to 22. This is because multiple pathologies may appear in a single case. Nonetheless, we consider them separately.

We will not discuss the *not in use* label, as it is present in the dataset but we don't know which abnormalities it may represent. Starting from hyperintense pathologies, we can notice that, generally speaking, models have a really low performance, with *3D-unet* generally performing better. As these lesions appear white as WMH, this result is not unexpected.

On hypointense or isointense lesions the performance is slightly higher for all methods, with *3D-unet* still generally performing best. We will not discuss hyperacute infarcts, as there is only a single case in the test set.

It is interesting to see that both *edema* and *tumor cyst* are labels that perform really poorly. The reason is probably due to a single test case that has a big tumor cyst, which is surrounded by a related edema, as it often happens with these types of pathologies. In addition, the annotated segmentation map reports a very tiny WMH volume, of $0.4mL$. This single sample scored a DSC=0.03 with *FCNE*, as it is affected by the worst conditions for this task, namely a poor WMH load and the presence of a very big hyperintense pathology. The sample can be visualized in Section 4.1.4, in Figure 4.7g.

Abnormality	n	Model	DSC (↑)	VS (↑)	HDD (↓)
Gliosis	7	3D-unet	**.49** (±**.21**)	**.77** (±**.11**)	31.33 (±12.23)
		3D-unet-noDA	.41 (±.27)	.66 (±.29)	34.87 (±**1.08**)
		FCNE	.44 (±.28)	.70 (±.28)	**4.52** (±17.16)
Edema	3	3D-unet	**.35** (±.26)	**.59** (±.42)	**44.16** (±21.53)
		3D-unet-noDA	.24 (±.25)	.45 (±.33)	49.96 (±18.87)
		FCNE	.32 (±**.24**)	.42 (±**.30**)	46.76 (±**15.42**)
Infarct, acute	5	3D-unet	.49 (±.28)	.64 (±.35)	**28.74** (±14.16)
		3D-unet-noDA	.43(±.30)	.60 (±.24)	33.10 (±**9.99**)
		FCNE	**.53** (±**.18**)	**.89** (±**.06**)	49.31 (±16.59)
Tumor cyst	3	3D-unet	**.30** (±.35)	**.40** (±.47)	**5.40** (±27.99)
		3D-unet-noDA	.28 (±.36)	.39 (±**.43**)	54.00 (±23.73)
		FCNE	.21 (±**.21**)	**.40** (±.50)	59.56 (±**16.03**)
Infarct, hyper.	1	3D-unet	.56	.65	22.24
		3D-unet-noDA	.46	.51	38.86
		FCNE	.56	.82	34.90
Infarct, chronic	6	3D-unet	**.53** (±**.20**)	.76 (±**.12**)	**27.90** (±8.98)
		3D-unet-noDA	.47 (±.24)	.74 (±.21)	32.16 (±7.78)
		FCNE	.50 (±.24)	**.78** (±.20)	34.61 (±**7.75**)
Tumor solid	4	3D-unet	**.51** (±.19)	.67 (±.27)	28.20 (±**4.06**)
		3D-unet-noDA	.39 (±.27)	.51 (±.30)	**3.69** (±7.13)
		FCNE	.40 (±**.03**)	**.74** (±.10)	49.43 (±8.71)
Not in use	3	3D-unet	**.53** (±**.26**)	**.89** (±**.07**)	**31.47** (±18.40)
		3D-unet-noDA	.30 (±.39)	.44 (±.39)	43.63 (±**6.47**)
		FCNE	.40 (±.31)	.66 (±.38)	59.75 (±15.37)

Table 4.14: Results for each tested method on the *WMH+OA* subset of the INDUSA dataset, divided by pathology label. On the upper part, are the abnormalities that are hyperintense on FLAIR, and in the lowest part are abnormalities that are iso or hypointense. In bold, the best mean and standard deviation for each metric.

4.1.3 | Multiple Sclerosis (MS)

Applying the same methods on the Multiple Sclerosis (MS) dataset is not straightforward. While MS lesions may look similar to white matter hyperintesities, there are still clinically relevant differences between the two. Nonetheless, it is interesting to see how the methods perform on a such different distribution.

Referencing Table 4.15, we can see that *3D-unet* is, overall, still the best-performing method on most metrics, except VS and Recall. However, the performance is still much lower compared to that on the MICCAI WMH, but a bit better to that of INDUSA. The variability in all metrics is also an issue, for example, the best DSC belongs to *3D-unet*, with DSC=.55 (.36-.74).

	DSC (↑)	VS (↑)	HDD (↓)	Precision (↑)	Recall (↑)
FCNE	.44 (±.22)	.64 (±.27)	71.83 (±**35.22**)	.55 (±.31)	**.49** (±.19)
3D-unet-noDA	.50 (±**.18**)	**.68** (±**.19**)	95.08 (±42.52)	.70 (±.26)	.42 (±**.15**)
3D-unet	**.55** (±.19)	.66 (±.22)	**69.66** (±43.29)	**.82** (±**.22**)	.45 (±.17)

Table 4.15: Results for each tested method on the MS dataset. In bold, the best mean and standard deviation for each metric.

Lesion load	Model	DSC (↑)	VS (↑)	HDD (↓)
high (n=25)	FCNE	.57 (±.15)	**.70** (±.20)	**59.73** (±**22.12**)
	3D-unet-noDA	**.61** (±**.13**)	.69 (±**.14**)	95.32 (±33.63)
	3D-unet	.60 (±.20)	.64 (±.20)	82.57 (±48.81)
low (n=28)	FCNE	.33 (±.21)	.59 (±.31)	83.03 (±41.36)
	3D-unet-noDA	.39 (±**.16**)	.68 (±.24)	94.87 (±**5.02**)
	3D-unet	**.50** (±.17)	**.69** (±**.23**)	**57.70** (±34.16)

Table 4.16: Results for each tested method on the MS dataset, divided by lesion load load. In bold, the best mean and standard deviation for each metric.

While it might seem that the result of *3D-unet* is not good enough, we can compare it to the best-performing method on the challenge where the MS dataset comes from. On that challenge, the highest average DSC was of .59. [11] *3D-unet* reaches a close outcome with no training on MS data or any data from the same distribution.

By dividing the MS dataset by lesion load, with the same criteria defined in Table 4.3, we can see if Multiple Sclerosis (MS) segmentation is also affected by the volume of the lesions. (Table 4.16)

On high lesion load, *3D-unet* DSC seems to improve compared to the average, but it is not the best method on VS and HDD, where *FCNE* is doing better. On low lesion load *3D-unet* is the best-performing method across all metrics, with a .10 decrease in performance compared to its high load counterpart.

The other methods are much more affected in DSC and HDD by low lesion load. For example, *FCNE* sees a .24 points drop in DSC and *3D-unet-noDA*, which has the best DSC on high load volumes, drops by .22 points when testing on low lesion loads.

It seems that the load only slightly affects the *3D-unet* method while improving its performance on HDD.

4.1.4 | Visual inspection of automatic segmentation

While qualitative evaluation measures may be useful to compare the performance of a wide range of methods quickly, it is also important to visualize the automatic segmentations to see if high scores translate into a good prediction. In medical imaging, it is particularly relevant to evaluate the quality of predictions. While in other imaging tasks, we might prefer to get the highest accuracy when classifying pixels, in medical imaging we can sacrifice pixel-wise accuracy if it means that the segmentation will not miss important details. Consequently, we will look at specific examples from each dataset to compare the performance of the three most relevant models (*FCNE, 3D-unet-noDA*, and 3D-unet). Per each presented case, the first image on the left is a single axial slice from the original image input, fed to each model for inference. The three images to its right side are the models' predictions. The *true positives* (TP), meaning the correct WMH predictions, are highlighted in green. In yellow, *false positives* display were the model is wrongly predicting the presence of WMH: this could be irrelevant if the false positives are a simple *oversegmentation* of a correctly identified prediction, or could be problematic if another abnormality (OA) is wrongly flagged as WMH. The latter case, *false positives on OA* is identified by the color red, as it is the outcome that we want to avoid as much as possible in this particular task. In blue, the original *ground truth* segmentation shows where the model failed to identify the presence of WMH, meaning the models' *false negatives* (FN). The ground truth annotation of the other abnormalities present in some datasets (OA) is colored purple.

4.1.4.1 | MICCAI WMH

In Figure 4.6 we can see four cases from the MICCAI WMH test set. In 4.6a, a case from the *GE3* site subset shows that *3D-unet* is the model that better adapts to this out-of-distribution case. In particular, we can notice that, on this particular slice, *FCNE*, and even more *3D-unet-noDA*, tend to over-segment the images. As previously discussed, this is not necessarily a big issue in this task, however, it explains why the DSC computation might be lower. The map predicted by *3D-unet* is very precise and fails mostly by having slightly differently shaped lesions. The Volumetric Similarity (VS) is high both in *3D-unet-noDA* and *3D-unet*, meaning that the volume predicted by both is close to the reference segmentation, however, *3D-unet-noDA* is less precise in delineating the shapes.

The case in 4.6a had a high load of WMH and little to no other abnormality (OA). In Figure 4.6b, we can see a case from the unseen scanner *SIN* with little to no OA, but a low WMH load. All models have a much lower DSC, but for different reasons. *FCNE* is greatly affected by the high signal that shows around the cortex on the left side and fails to predict the WMH around the right ventricle. *3D-unet-noDA* doesn't wrongly classify the cortex as WMH, but its map is very poor because of undersegmentation. *3D-unet* is able to improve *3D-unet-noDA* by segmenting the WMH around the ventricles with more accuracy.

In Figure 4.6c, we can see a case with a high WMH load and the presence of OAs. Additionally, we can easily notice a motion artifact. This artifact is wrongly predicted as WMH by *FCNE*. *3D-unet-noDA* doesn't pick up the artifact as positive but is still affected by undersegmentation. *3D-unet* is slightly better, but all three models have wrongly predicted part of the OA as WMH. Arguably, the OA is very small and the annotation accuracy is dependent on the voxel size.

A low load WMH ($0.9mL$) with the presence of OAs ($1.03mL$) is presented in Figure 4.6d. All models' predictions look very similar. The hyperintense signal around the ventricles might be an imaging artifact or WMH, so an over-segmentation around this area is expected and might also be a matter of disagreement. Concerning the OA, all models predict the small hyperintense lesion as WMH, but it is probably a small edema related to the other hypointense lesion close to it. *3D-unet* has the best prediction, with a $+.29$DSC ($1.41mL$ predicted volume) over *FCNE* ($2.89mL$), but it is only slightly more precise with respect to the other models. This shows that for very small lesions, it is hard to evaluate different models based on quantitative metrics alone.

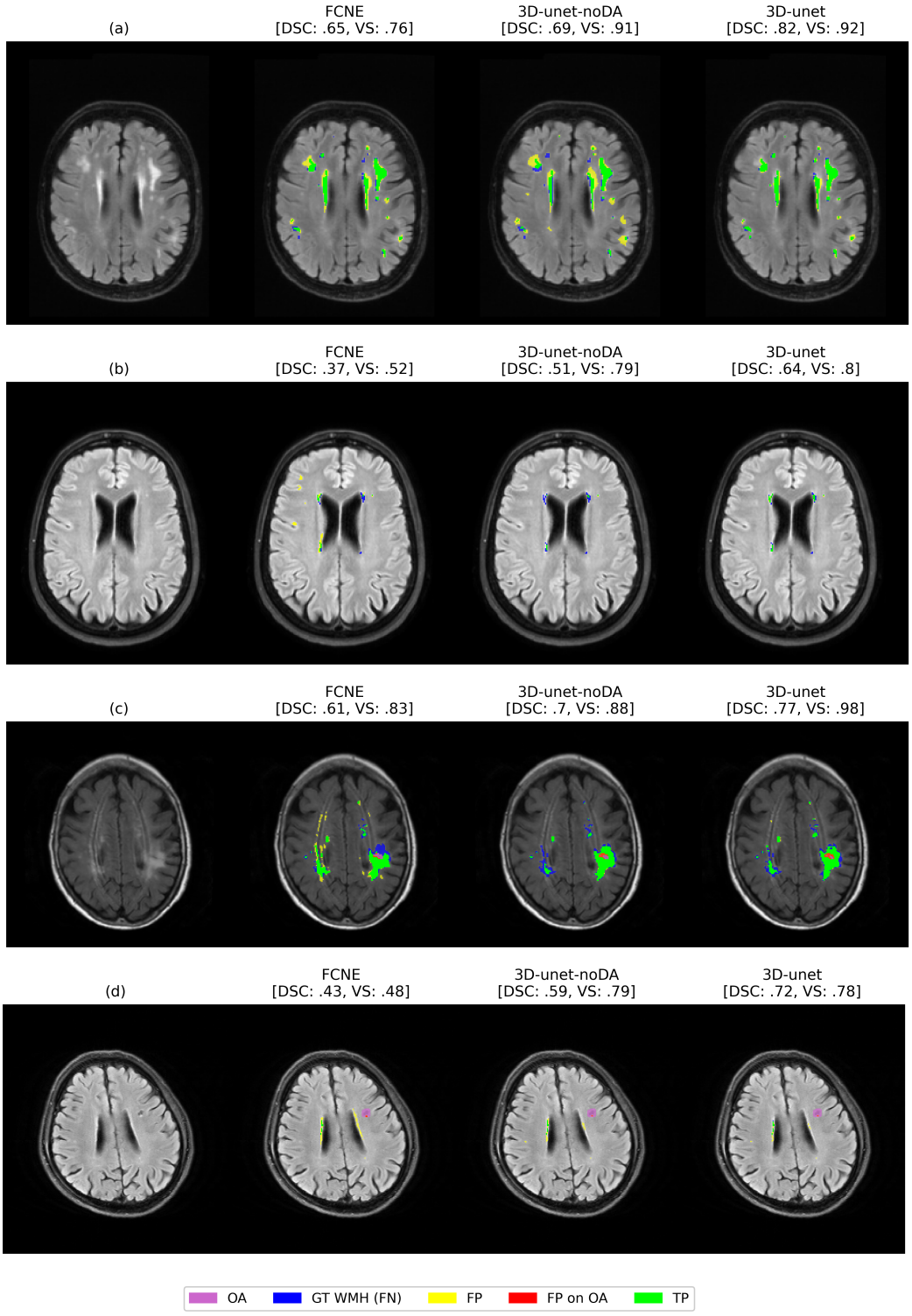


Figure 4.6: Predictions on images from the MICCAI WMH test set.

4.1.4.2 | INDUSA

On the INDUSA dataset, we are more interested in seeing how the models perform in difficult cases, that is to say cases with OAs and low WMH load.

First, by looking at Figure 4.7e, we can see how the model generalizes on a high load WMH with no OAs, to be sure that it is able to adapt to simple test samples. *FCNE* is undersegmenting lesions around the ventricles, while wrongfully segmenting some cortex and brain anatomy high intensity signals. *3D-unet-noDA* doesn't pick up wrong areas, but its segmentations around the ventricles are even poorer than *FCNE*. We seem to almost solve both issues with *3D-unet*.

What happens when a large hyperintense OA is co-present with WMH? In Figure 4.7f, an acute infarct is very bright and close to some WMH. FCNE is almost completely missing all WMH, which explains its poor DSC. *3D-unet-noDA*, on the other hand, is correctly segmenting the WMH close to the ventricles but is also fooled by the presence of the artifact, resulting in a big false positive area. *3D-unet* has the same DSC and close VS as *3D-unet-noDA*, but the wrong prediction area is much smaller. This brings us back to the concept discussed previously: two images may have the same quantitative results, but one may have a better segmentation than the other based on what is actually flagged as WMH and what is not.

In Section 4.1.2.2 we discussed a case that had a really poor performance, with the highest DSC= .03. We can see this case in Figure 4.7g, where we have the presence of both hyperintense and hypointense abnormalities, a tumor and an edema surrounding it. Additionally, the load of WMH in this case is very low, only $0.40mL$. For comparison, the tumor and edema have a volume of $37.78mL$ and $17.15mL$, respectively. All models failed in predicting the edema as WMH, but didn't with the tumor.

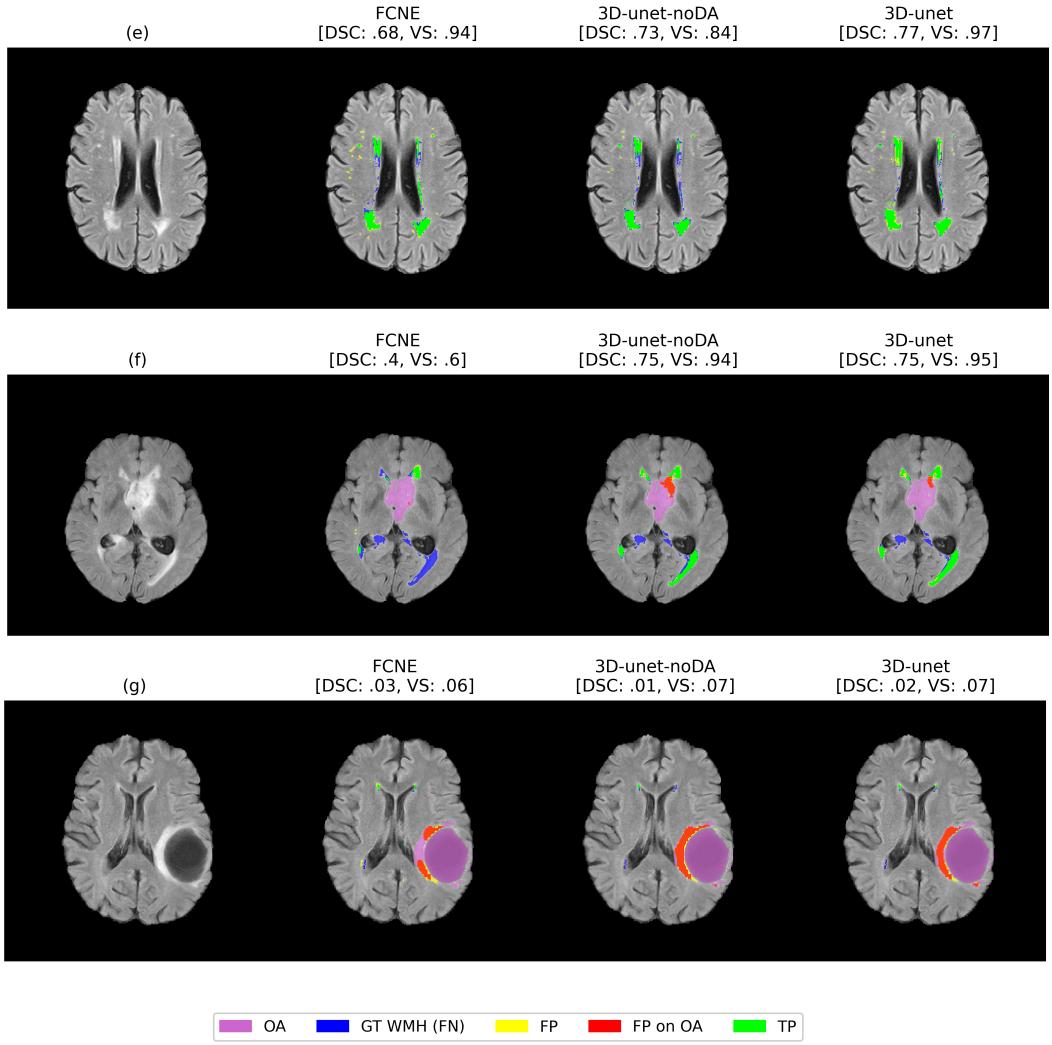


Figure 4.7: Predictions on images from the INDUSA test set.

4.1.4.3 | MS

In Figure 4.8 we can see how the models trained on WMH do on a completely different domain, Multiple Sclerosis (MS). The anatomy of lesions in MS is very different, but they still show up as hyperintesities on FLAIR.

All models reach a good DSC on Figure 4.8h, with *3D-unet* having the best DSC and *3D-unet-noDA* the best VS. On this slice, all models, but less for *3D-unet-noDA*, delineate the contours of the lesions precisely. The false positives in *FCNE* are less present in *3D-unet*.

It is interesting to see that in Figure 4.8i, *FCNE* fails to pick up very bright intensities but has a lot of less intense false positives. The latter does not seem to be a problem for *3D-unet-noDA* and *3D-unet*, with *3D-unet* yielding the more precise segmentation.

On a image with a low load of lesions like in Figure 4.8l, *FCNE* seems to be again confused by intensities on the cortex, but still segments the WMH around the ventricles correctly. While these don't affect the outcome of *3D-unet-noDA* and *3D-unet*, their DSC scores are still very low, but with a high VS, indicating again that it's hard to evaluate small lesions based on DSC.

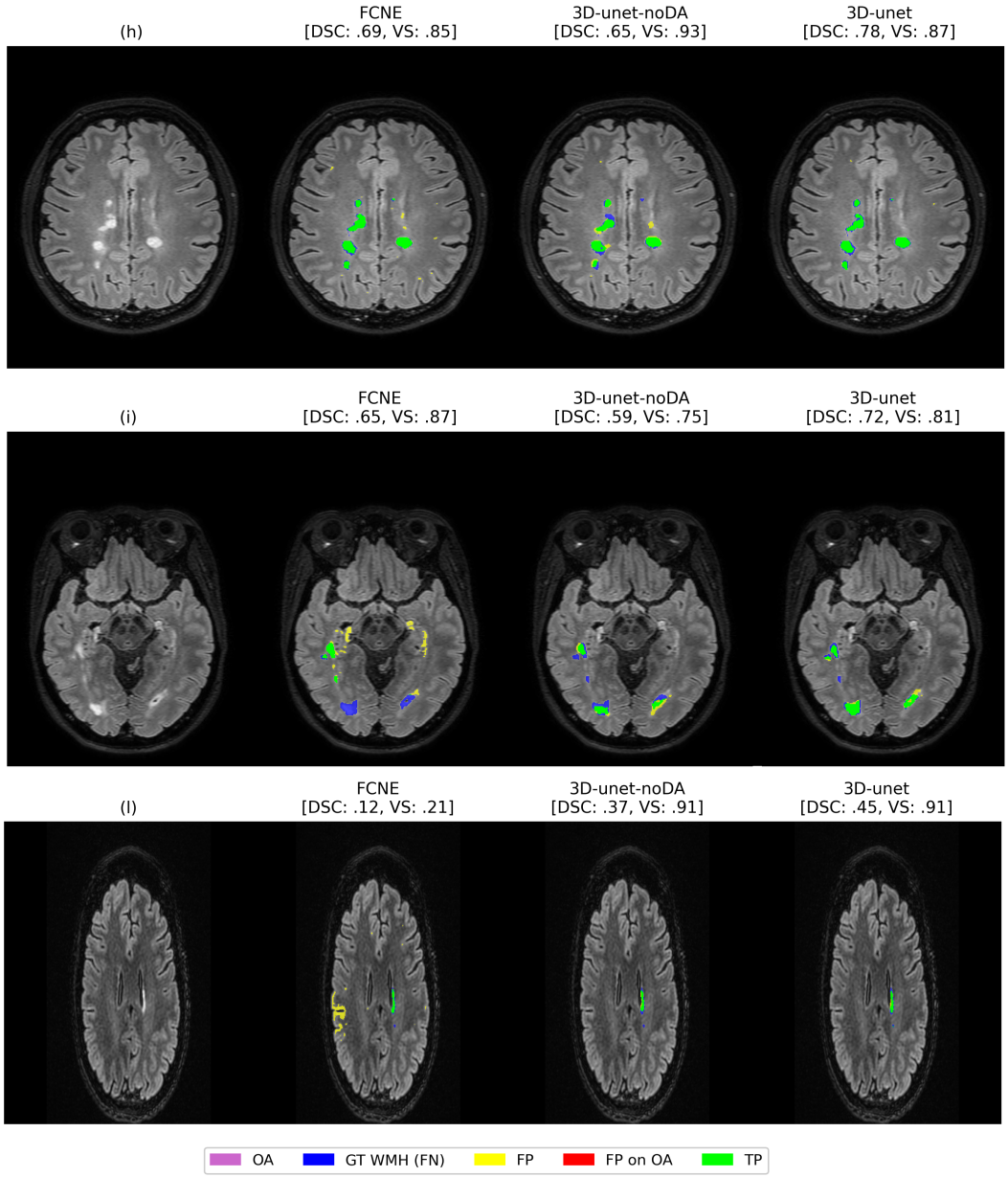


Figure 4.8: Predictions on images from the MS test set.

4.1.4.4 | LF

Most of the images in the Low (Magnetic) Field MRI (LF) test set do not have a high WMH load. This condition has been assessed visually, with no expert radiologist input. Given that no manual ground truth annotation is available for this test set, the following discussion is not to be considered correct in medical terms, but it is to be interpreted as a set of inexperienced observations. The model predictions are drawn in ocean green, to distinguish them from the true positives of previous visualizations.

In Figure 4.9m, we can assess the quality of the segmentation in a case with a high WMH load. *FCNE* fails almost completely at segmenting the proposed slice. The performance of *3D-unet-noDA* and *3D-unet* is similar. However, it is unclear if the big lesion on the right side is to be considered WMH or OA.

Regarding OAs, in Figure 4.9n, *3D-unet-noDA* is the only model that segments as WMH which probably isn't, judging from a single slice only. *FCNE* and *3D-unet* are more conservative in their outputs, which result very similar to each other.

Lastly, in Figure 4.9o, *FCNE* is the only model that, once again, finds WMH where is not anatomically possible to have them. In this case, positive voxels are present in the left eye and in the back of the skull bone. *3D-unet-noDA* and *3D-unet* are both correct in their absence of positives. *FCNE* failure is more visible in Figure 4.10.

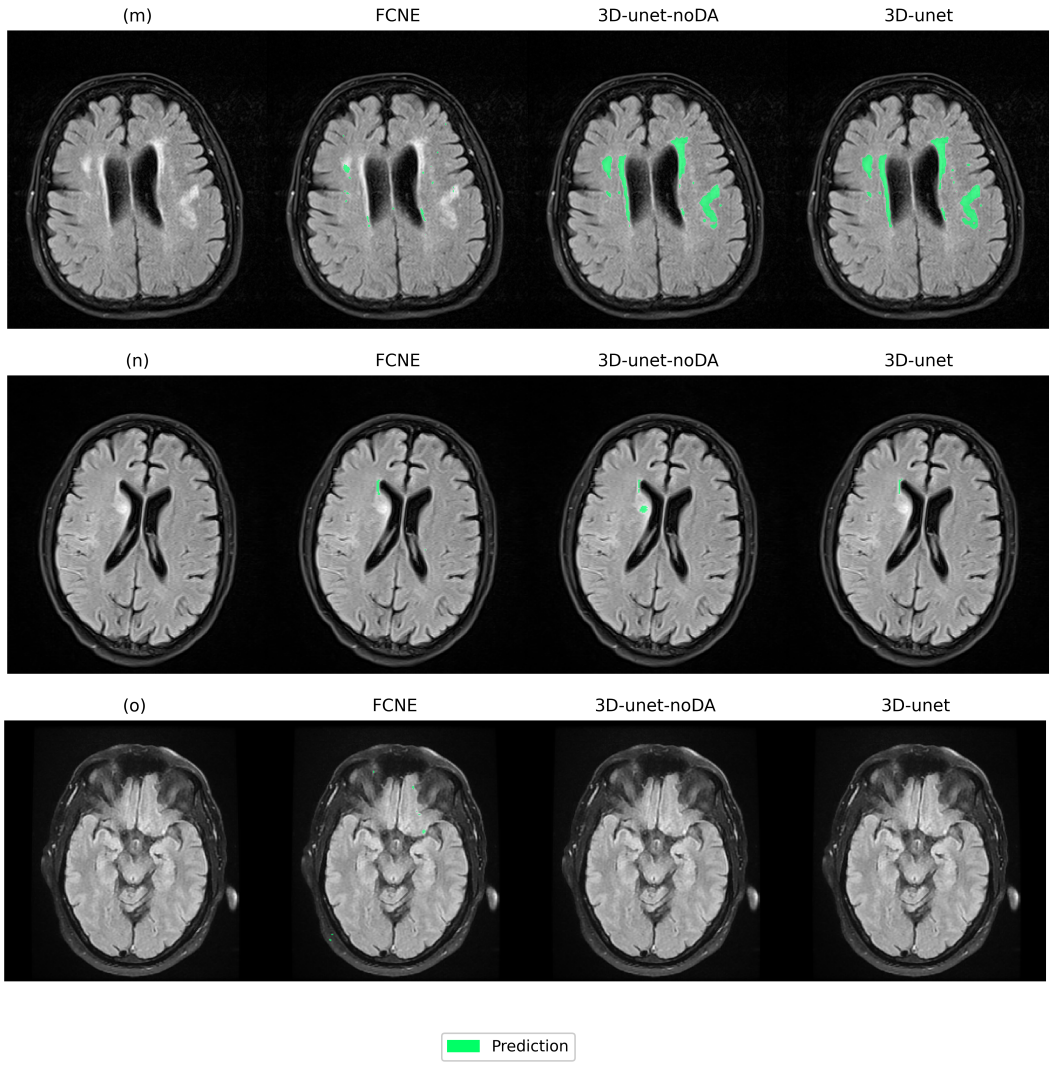


Figure 4.9: Predictions on images from the LF test set.

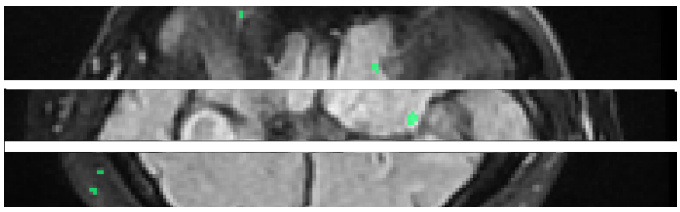


Figure 4.10: Prediction from Figure 4.9o, zoomed on the *FCNE* model output.

4.2 | Discussion

In the previous section, we presented our research findings. Now, in this discussion section, we'll interpret these results, explore their implications, and consider their limitations to better understand their significance on the research hypothesis of this thesis.

All evaluated methods were trained on the training set ($n = 60$) of the MICCAI WMH 2017 Segmentation Challenge, using only one image modality (FLAIR) of the two available. The *FCNE* was the winner of said challenge, back in 2017, but was trained on both FLAIR and T1-weighted MRI sequences. As we want to develop a method that is as robust as possible in a wide range of clinical settings, we want to achieve the same or better results with a model that requires only FLAIR as its input.

Evaluating all the methods developed for the task on the MICCAI WMH test set, we can conclude that *3D-unet*, a modification of the nnUNet framework with the addition of MRI-specific data augmentation, is the model that performs best. We infer this from quantitative findings and visual inspection of the images. *FCNE* performance was not consistent through all samples, displaying a wide variability in performance.

Methods trained on two-dimensional inputs, like axial slices instead of whole volumes, improved the performance over *FCNE*, but they did not reach the same quantitative metrics results as their three-dimensional counterparts. *3D-multires-unet* did not show any advantage when using a MultiRes architecture over the simple *2D-unet* architecture.

Of the 3D models, *3D-unet-contralat* was the only method that performed worse than *FCNE*. The explanation could be given by the fact that learning with two modalities instead of one may require more training resources, which stayed all the same during the training of the other methods. We leave to future work the exploration of the idea of adding contralateral information when learning how to segment WMH.

3D-unet-resampl is the same method as in *3D-unet*, but with a wider range of parameters in the resampling functions during data augmentation. It seems that, in this particular case, this single intervention was not enough to improve under segmentation. However, this could simply be due to random initialization of the models' weights, and more experiments are needed to be certain.

It seems that MRI-specific data augmentation plays a big role in improving the performance of *3D-unet-noDA*. The same model but with the additional augmentation functions, which we called *3D-unet*, performs better on average, especially on out-of-distribution within the MICCAI WMH test set, but also on datasets like INDUSA and MS.

All methods were challenged by the presence of abnormalities different than WMH.

It was especially clear when evaluating INDUSA, which is a dataset used for training models for the segmentation of pathologies like tumors and infarcts. The performance was poor when hyperintense abnormalities were present, but *3D-unet* is the model that yields a better performance, as it predicts a lower number of false positives on OAs.

Multiple Sclerosis (MS) segmentation was not solved by any of the models tested, however, *3D-unet* results were close to the best method on the challenge based on the tested data. This outcome suggests that fine-tuning or including MS lesions in the training of the proposed model may bring good results even on MS.

Low magnetic field scanners are becoming more popular in clinical practise. Methods that are trained on 1.5T or 3T images must be robust to changes in resolution and overall image appearance. We can conclude, based only on visual examination of a few cases, that *3D-unet* is the best method for the segmentation of WMH on low-field images.

Regarding quantitative metrics, we have seen how Sørensen–Dice Similarity Coefficient (DSC) is a good metric to evaluate the method performance on average, but that it is heavily influenced by the segment size: two lines that fall close to each other, but with no overlap, yield the same DSC score that two lines that lie far apart do. The Hausdorff Distance (HDD) has a similar use case, that is to compare models and choose the best performing, however being unbounded on one end makes it unintuitive for single case examination, or when shifting domain. Volumetric Similarity (VS) is a good reference metric to have along DSC, as it provides more information about the segmented volume without being affected by the lesions' shape and size. Visual evaluation is the best way to perform analysis on performance, but it is more time-consuming and not easy to deploy over a large number of tested models.

5

Conclusions

5.1 | Achieved Aims and Objectives

In summary, this project focused on the segmentation of white matter hyperintesities (WMH) using deep learning techniques. We trained and tested a multitude of models, searching for the one that was more robust to domain shift. We successfully achieved our original aims and objectives, effectively utilizing deep learning models and a single MRI sequence, fluid attenuated inversion recovery (FLAIR), to accurately identify and segment WMH regions from medical imaging data. Training a 3D U-Net with the nnUNet framework and the addition of MRI-specific data augmentation yielded the best performance over a variety of test sets, using different quantitative and qualitative measures, including individual case visual inspection.

5.2 | Critique and Limitations

Like other medical imaging tasks, this project was limited in the number of training and testing resources. The dataset used for training is too *clean* to be the only source for a method that aims at being deployed in clinical practice, meaning that it lacks the variety of other abnormalities that are often occurring along white matter hyperintesities (WMH).

Access to annotated data from different healthcare centers would greatly improve the study of white matter hyperintesities (WMH) and other brain imaging segmentation tasks, enabling models to encounter real-world scenarios.

Additionally, segmentation results can be very different based on the expertise of trained radiologists. Because annotated data coming from different clinicians may look

different, this variance must be expected also by learning models. The evaluation of the shape of lesions is greatly affected by the metrics and human raters' opinions, meaning that chasing a perfect score is not the ultimate goal of this type of task. Achieving a segmentation that is able to distinguish white matter hyperintesities (WMH) from other abnormalities is more important than yielding the perfect lesion contour.

5.3 | Future Work

A simple suggestion for future work is that of training more models with MRI-specific segmentation, not only on the task of white matter hyperintesities (WMH) analysis but also on other brain or different organ conditions.

Additionally, what effect can we expect by fine-tuning the best-performing method on a subset of the target distribution? Other research suggests that transferring the original domain knowledge to a new one generally improves the inference outcomes. In particular, adding cases with more hyperintense abnormalities may drastically improve the performance on new test sets.

From a technical perspective, the Sørensen–Dice Similarity Coefficient (DSC) based training loss might not be the best choice when learning how to segment very small lesions. Further efforts trying to solve this matter are required.

To assess the robustness of the proposed method, more testing on a large cohort of different sites is needed. The annotations in this cohort must be aimed at white matter hyperintesities (WMH) segmentation so that quantitative measures can be as precise as possible.

5.4 | Final Remarks

In conclusion, the proposed approach for the segmentation of white matter hyperintesities (WMH) using deep learning has yielded significant progress to existing methods, while offering a solid foundation for further research and innovation. Given the importance of white matter hyperintesities (WMH) on patients affected by cognitive disorders and other medical conditions, it is crucial to develop tools that can aid clinicians and researchers in their pursuit of answers and solutions in this critical area. Despite the challenges and limitations, this thesis underscores the potential of deep learning to make a meaningful impact on healthcare, which can be a valuable tool in the quest for more precise diagnosis and treatment strategies.

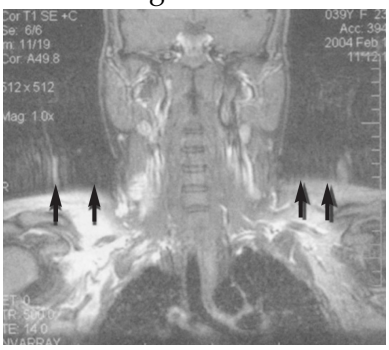
A

Overview of imaging artifacts

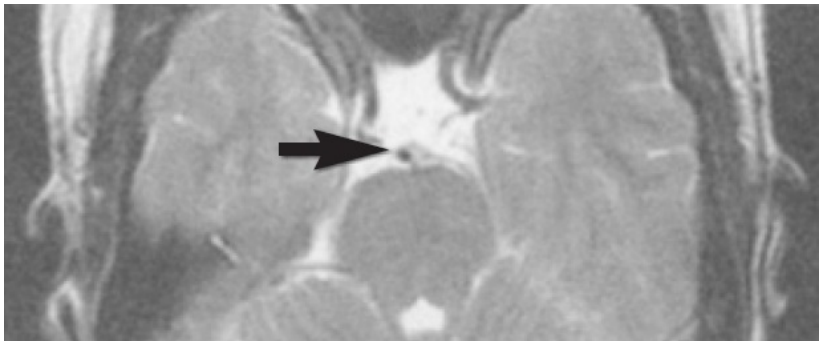
In this section we will display and describe a few of the most common MRI artifacts, as presented by [74]. They define an artifact as *a feature appearing in an image that is not present in the original object*. We can classify artefacts, depending on their origin, as patient-related, signal processing-dependent or hardware (machine)-related.

A.1 | Patient-related MR artefacts

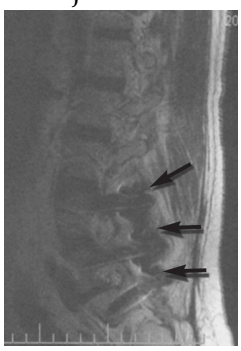
- Motion artifact: blurring or distortion of the image caused by the patient's movement during the scan.



- Flow-related signal loss: phenomenon where the signal intensity is reduced or distorted in areas with high flow or motion, leading to reduced image contrast.

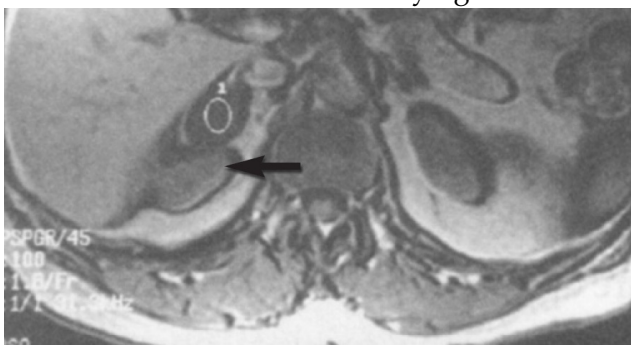


- Metal-related artifact: distortion or signal void observed in the presence of metallic objects within or near the imaging area.

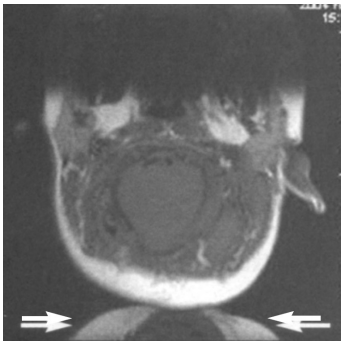


A.2 | Signal processing-dependent artefacts

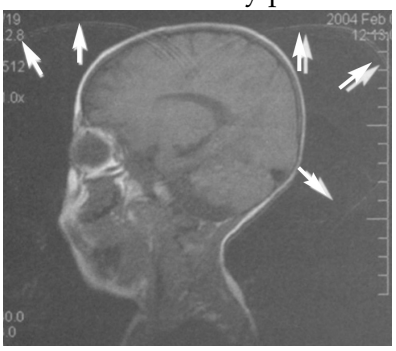
- Chemical shift artifact: occurs when there is a mismatch in the resonance frequencies of different tissues with varying chemical compositions.



- Partial volume artifact: occurs when objects smaller than the voxel dimensions lose their identity due to signal averaging, leading to a loss of detail and spatial resolution.
- Wrap-around artifact: occurs when the field of view is smaller than the body part being imaged, which is then is projected onto the other side of the image.

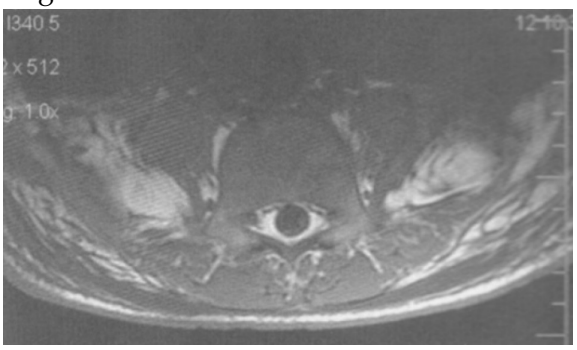


- Gibbs, or ringing artifact: caused by undersampling of the signal, resulting in oscillations or wavy patterns at sharp tissue boundaries.

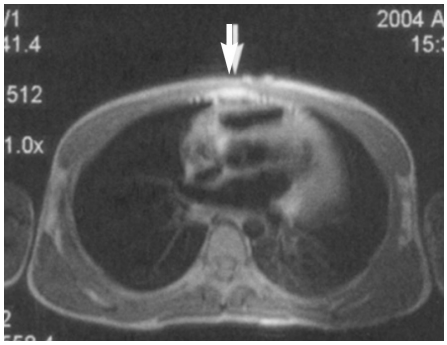


A.3 | Machine or hardware-related artifacts

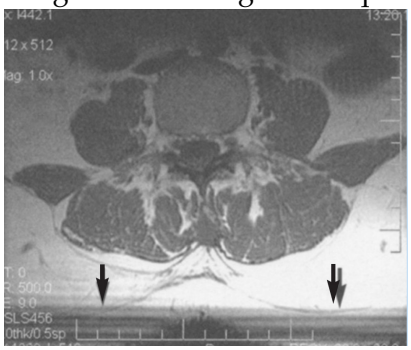
- Radiofrequency quadrature, or ghosting, artifact: it is due to improper detector channel operation, resulting in object ghosting.
- B0 inhomogeneity: variations in the main magnetic field (B0) can cause signal loss or geometric distortions.



- Surface coil artifact: is caused by using a specialized coil close to the body surface, leading to non-uniform signal intensity and localized image distortions.



- Asymmetrical brightness: a uniform decrease in signal intensity leads to rejection of part of the signal.
- Slice-to-slice intereference: occurs when there is contamination between adjacent image slices during data acquisition.



B

Data augmentation parameters

The configuration parameters for the data augmentation process carried during experiments are reported in this appendix.

Default 2D parameters

Transformation	**Parameters**
do elastic	True
elastic deform alpha	$(0., 200.)$
elastic deform sigma	$(9., 13.)$
p eldef	0.2
do scaling	True
scale range	$(0.85, 1.25)$
independent scale factor for each axis	False
p independent scale per axis	1
p scale	0.2
do rotation	True
rotation x	$(-180./360 \times 2 \times \pi, 180./360 \times 2 \times \pi)$
rotation y	$(-0./360 \times 2 \times \pi, 0./360 \times 2 \times \pi)$
rotation z	$(-0./360 \times 2 \times \pi, 0./360 \times 2 \times \pi)$
random crop	False
random crop dist to border	None
do gamma	True
gamma retain stats	True
gamma range	$(0.7, 1.5)$
p gamma	0.3
do mirror	True
mirror axes	(0, 1)
dummy 2D	False
mask was used for normalization	None
border mode data	"constant"
do additive brightness	False
additive brightness p per sample	0.15
additive brightness p per channel	0.5
additive brightness mu	0.0
additive brightness sigma	0.1

Default 3D parameters

Transformation	**Parameters**
do elastic	True
elastic deform alpha	$(0., 900.)$
elastic deform sigma	$(9., 13.)$
p eldef	0.2
do scaling	True
scale range	$(0.85, 1.25)$
independent scale factor for each axis	False
p independent scale per axis	1
p scale	0.2
do rotation	True
rotation x	$(-15./360 \times 2 \times \pi, 15./360 \times 2 \times \pi)$
rotation y	$(-15./360 \times 2 \times \pi, 15./360 \times 2 \times \pi)$
rotation z	$(-15./360 \times 2 \times \pi, 15./360 \times 2 \times \pi)$
random crop	False
random crop dist to border	None
do gamma	True
gamma retain stats	True
gamma range	$(0.7, 1.5)$
p gamma	0.3

New configuration parameters (2D)

Transformation	**Parameters**
selected data channels	None
selected seg channels	[0]
additiveNoise p per sample	0.33
additiveNoise mean	0
additiveNoise sigma	$1 \times 10^{-4} \times$ random.uniform()
biasField p per sample	0.5
elasticDeform p per sample	0.33
elasticDeform alpha	$(200, 600)$
elasticDeform sigma	$(20, 30)$
gibbsRinging p per sample	0.33
gibbsRinging cutFreq	random.randint$(96, 129)$
gibbsRinging dim	random.randint$(0, 3)$
motionGhosting p per sample	0.33
motionGhosting alpha	random.uniform$(0.85, 0.95)$
motionGhosting numReps	random.randint$(2, 11)$
motionGhosting dim	random.randint$(0, 3)$
multiplicativeNoise p per sample	0.33
multiplicativeNoise mean	0
multiplicativeNoise sigma	$1 \times 10^{-3} \times$ random.uniform()
rotation p per sample	0.33
rotation p per axis	0.66
rotation x	$(-30./360 \times 2 \times \pi, 30./360 \times 2 \times \pi)$
rotation y	$(-0./360 \times 2 \times \pi, 0./360 \times 2 \times \pi)$
rotation z	$(-0./360 \times 2 \times \pi, 0./360 \times 2 \times \pi)$
blurring per channel	False
blurring sigma	$(0., 1.)$
blurring p per sample	0.
blurring per axis	False
blurring p isotropic	0.
gamma retain stats	True
p gamma	0.
mirror axes	(0, 1, 2)
scale range	$(1, 1)$
scale p per sample	0.33
mask was used for normalization	None
border mode data	"constant"
num threads	6
num cached per thread	2

New configuration parameters (3D)

Transformation	**Parameters**
additiveNoise p per sample	0.33
additiveNoise mean	0
additiveNoise sigma	$1 \times 10^{-4} \times$ random.uniform()
biasField p per sample	0.5
elasticDeform p per sample	0.33
elasticDeform alpha	$(200, 600)$
elasticDeform sigma	$(20, 30)$
do gibbsRinging	True
gibbsRinging p per sample	0.33
gibbsRinging cutFreq	random.randint$(96, 129)$
gibbsRinging dim	random.randint$(0, 3)$
do motionGhosting	True
motionGhosting p per sample	0.33
motionGhosting alpha	random.uniform$(0.85, 0.95)$
motionGhosting numReps	random.randint$(2, 11)$
motionGhosting dim	random.randint$(0, 3)$
do multiplicativeNoise	True
multiplicativeNoise p per sample	0.33
multiplicativeNoise mean	0
multiplicativeNoise sigma	$1 \times 10^{-3} \times$ random.uniform()
rotation p per sample	0.33
rotation p per axis	0.66
rotation x	$(-30./360 \times 2 \times \pi, 30./360 \times 2 \times \pi)$
rotation y	$(-30./360 \times 2 \times \pi, 30./360 \times 2 \times \pi)$
rotation z	$(-30./360 \times 2 \times \pi, 30./360 \times 2 \times \pi)$
random crop	False
random crop dist to border	None
blurring per channel	True
blurring sigma	$(0., 1.)$
blurring p per sample	0.33
blurring per axis	True
blurring p isotropic	0.33
gamma retain stats	True
gamma range	None
p gamma	0.
mirror axes	(0, 1, 2)
scale range	$(1, 1.5)$
scale p per sample	0.33

C

More results

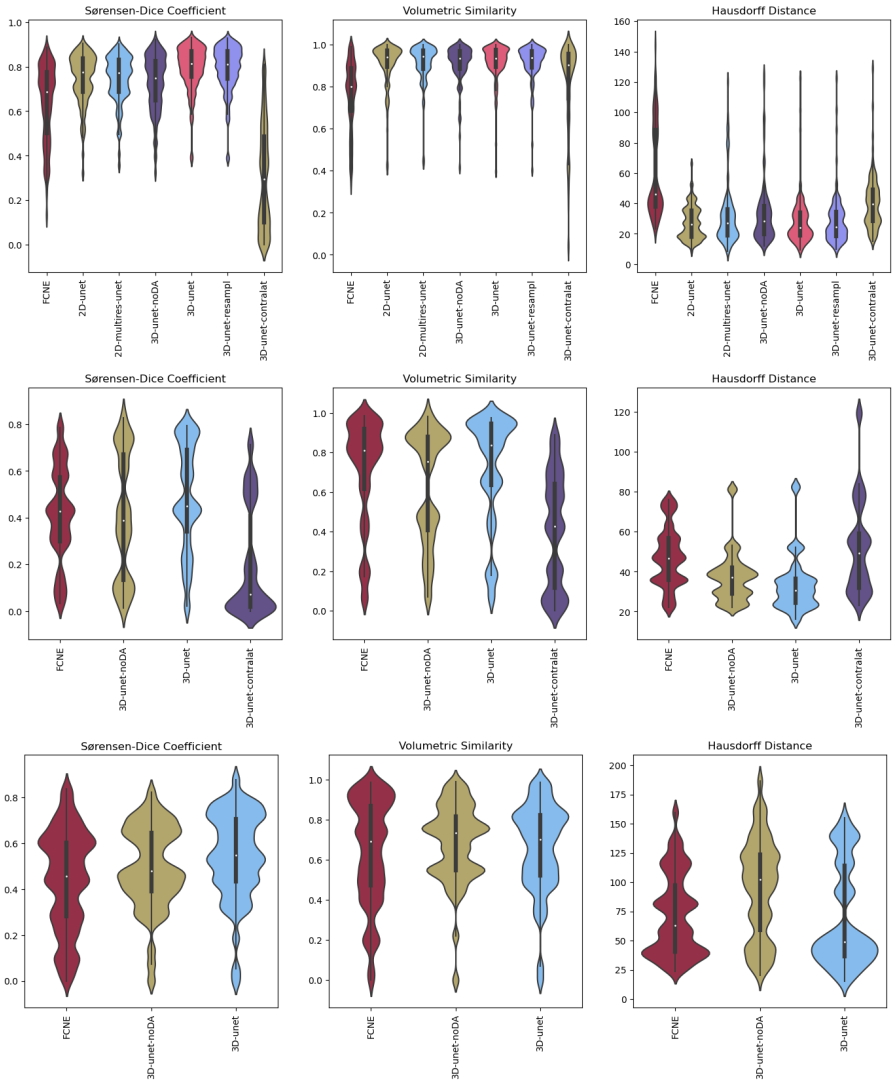


Figure C.1: Violin plots on the tested datasets, showing the distribution of Sørensen–Dice Similarity Coefficient (DSC), Volumetric Similarity (VS), and Hausdorff Distance (HDD) on all cases.

Site	n	Model	DSC (↑)	VS (↑)	HDD (↓)
UTR	30	FCNE	.67 (±.17)	.86 (±.13)	4.57 (±8.85)
		3D-unet-resampl	.79 (±**.09**)	.92 (±**.08**)	31.18 (±9.84)
		3D-unet-noDA	.73 (±.11)	.88 (±.10)	34.73 (±**1.97**)
		3D-unet-contralat	.29 (±.22)	.79 (±.23)	**45.16** (±11.57)
		3D-unet	**.80** (±.09)	**.93** (±.08)	31.28 (±9.87)
		2D-unet	.74 (±.10)	.89 (±.10)	3.64 (±9.97)
		2D-multires-unet	.75 (±.10)	**.93** (±.08)	3.34 (±9.95)
SIN	30	FCNE	.71 (±.14)	.80 (±.14)	**39.49** (±7.44)
		3D-unet-resampl	**.82** (±.12)	.91 (±.12)	24.37 (±9.99)
		3D-unet-noDA	.80 (±.13)	**.92** (±.12)	26.14 (±**1.07**)
		3D-unet-contralat	.34 (±.24)	.84 (±.19)	38.69 (±14.38)
		3D-unet	**.82** (±.12)	.91 (±.12)	24.47 (±9.98)
		2D-unet	.78 (±.12)	.91 (±.12)	26.18 (±1.27)
		2D-multires-unet	.78 (±**.11**)	.90 (±**.11**)	26.96 (±12.36)
GE1	30	FCNE	.59 (±.17)	.69 (±.15)	**93.96** (±37.64)
		3D-unet-resampl	.77 (±**.09**)	.92 (±.04)	39.52 (±39.38)
		3D-unet-noDA	.64 (±.11)	.93 (±**.03**)	43.87 (±37.37)
		3D-unet-contralat	.34 (±.23)	.86 (±.13)	33.16 (±14.03)
		3D-unet	**.78** (±.09)	.92 (±.04)	39.48 (±39.40)
		2D-unet	.73 (±.11)	**.96** (±.03)	24.27 (±11.19)
		2D-multires-unet	.73 (±.10)	.93 (±.04)	29.35 (±**2.17**)
GE3	10	FCNE	.55 (±.19)	.66 (±.18)	**93.96** (±13.72)
		3D-unet-resampl	**.78** (±.11)	**.91** (±.09)	23.97 (±9.15)
		3D-unet-noDA	.68 (±.12)	**.91** (±.08)	25.27 (±**8.48**)
		3D-unet-contralat	.31 (±.26)	.80 (±.25)	38.74 (±24.67)
		3D-unet	**.78** (±.11)	**.91** (±.09)	23.53 (±**8.48**)
		2D-unet	.73 (±.12)	**.91** (±.11)	23.80 (±9.21)
		2D-multires-unet	.72 (±**.11**)	.90 (±.10)	25.99 (±19.69)
PHI	10	FCNE	.64 (±.15)	.73 (±.13)	52.14 (±22.64)
		3D-unet-resampl	**.79** (±.08)	.93 (±.06)	4.41 (±21.82)
		3D-unet-noDA	.74 (±.12)	.93 (±.06)	61.62 (±36.91)
		3D-unet-contralat	.19 (±.16)	.79 (±.23)	44.16 (±18.49)
		3D-unet	**.79** (±.08)	.93 (±.06)	34.19 (±**2.95**)
		2D-unet	.73 (±.11)	.92 (±.11)	32.43 (±15.62)
		2D-multires-unet	.74 (±.09)	**.95** (±.05)	**74.58** (±19.58)

Table C.1: Results on each site from the MICCAI WMH test set, for each tested model. *GE1* and *PHI* are the site unseen during training.

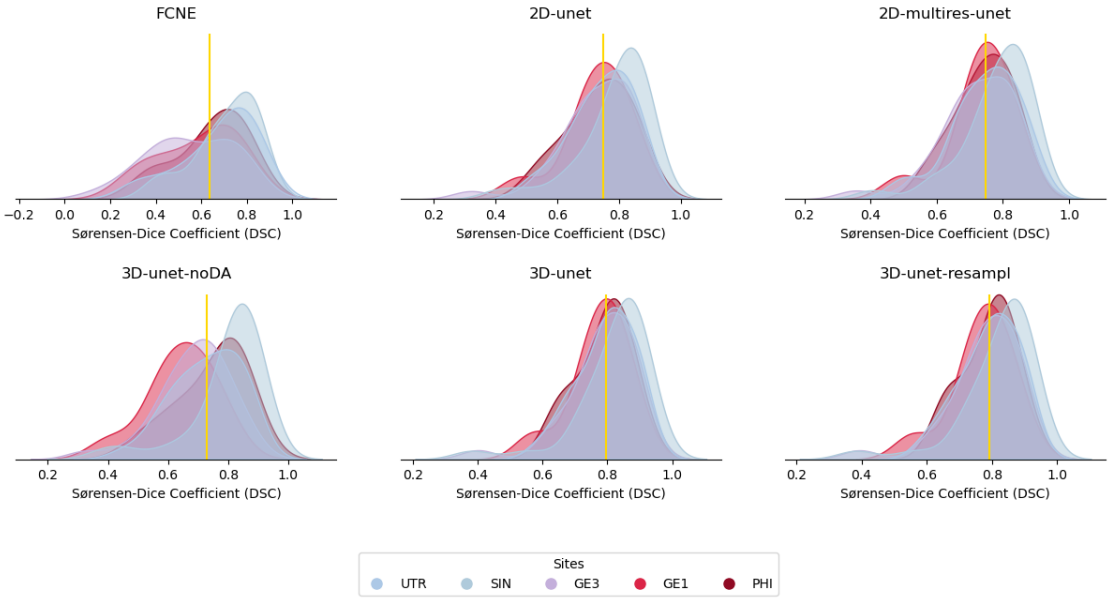


Figure C.2: Sørensen–Dice Similarity Coefficient distribution for each tested model on the MICCAI WMH test set, divided by site. In red, the sites unseen during training. In yellow, the average DSC value computed over all tested cases.

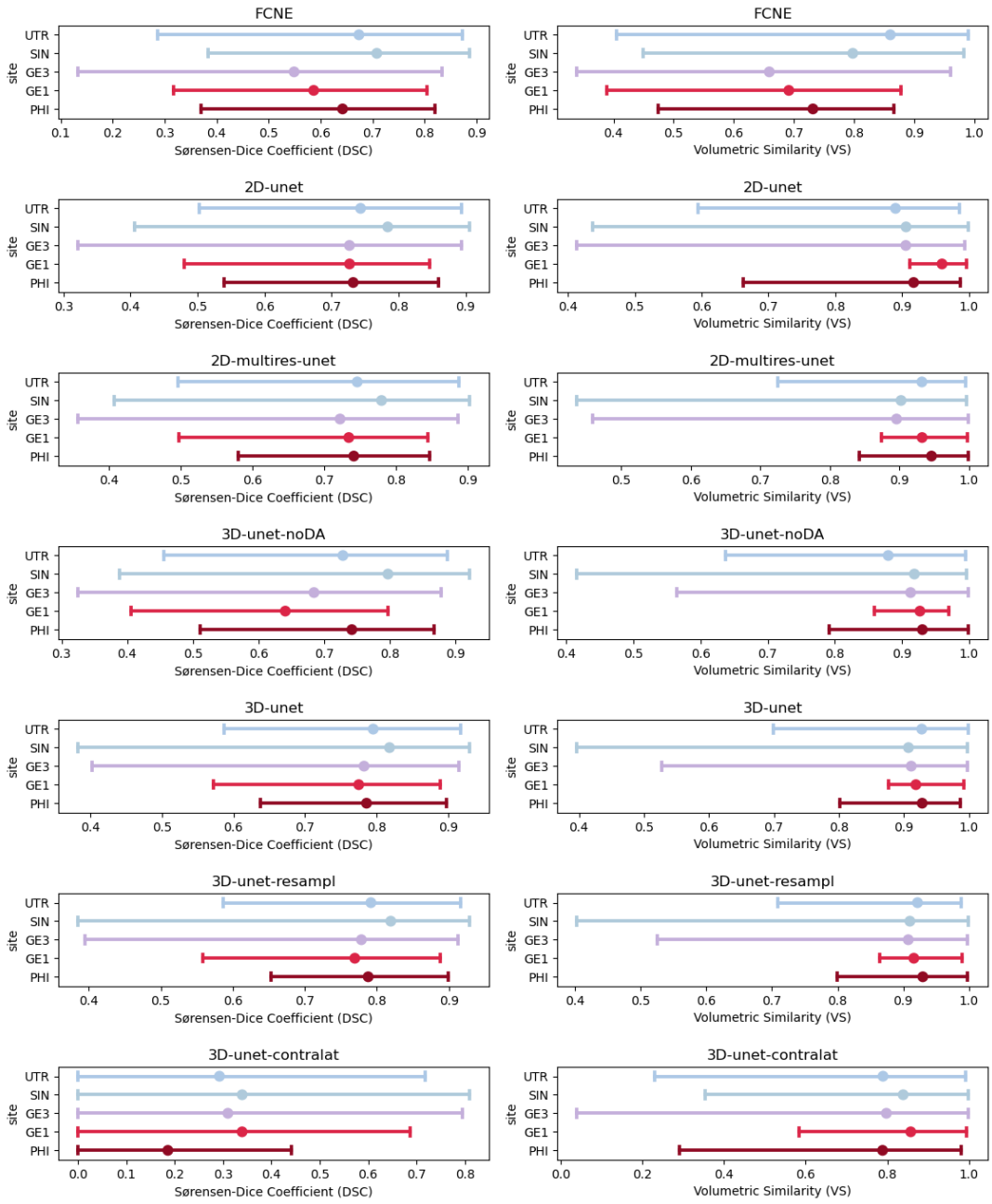


Figure C.3: Distribution of Sørensen–Dice Similarity Coefficient (DSC) and Volumetric Similarity (VS) for each tested model on sites from the MICCAI WMH dataset. In red, sites unseen during training.

References

[1] M. Omidi, M. Zibaii, and N. Granpayeh, "Simulation of nerve fiber based on anti-resonant reflecting optical waveguide," *Scientific Reports*, vol. 12, 11 2022.

[2] D. C. Preston, "Mri basics," 2006.

[3] A. Hawkins-Daarud, R. C. Rockne, A. R. A. Anderson, and K. R. Swanson, "Modeling tumor-associated edema in gliomas during anti-angiogenic therapy and its impact on imageable tumor," *Frontiers in Oncology*, vol. 3, 2013.

[4] A. D. Elster, "Slice cross-talk - questions and answers in mri." `https://mriquestions.com/cross-talk.html`, 2000. Accessed on July 20, 2023.

[5] A. Kavitha and C. Chellamuthu, *Advanced Brain Tumour Segmentation from MRI Images*. 03 2018.

[6] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, Aug. 2015.

[7] Anh Nguyen, "Understanding neural networks via feature visualization: A survey - scientific figure on researchgate," 2019.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[9] A. Elster, "Building blocks of a CNN (modified from lundervold under cc by)," 2023.

[10] H. J. Kuijf, A. Casamitjana, D. L. Collins, M. Dadar, A. Georgiou, M. Ghafoorian, D. Jin, A. Khademi, J. Knight, H. Li, X. Llado, J. M. Biesbroek, M. Luna, Q. Mahmood, R. McKinley, A. Mehrtash, S. Ourselin, B.-Y. Park, H. Park, S. H. Park, S. Pezold, E. Puybareau, J. D. Bresser, L. Rittner, C. H. Sudre, S. Valverde, V. Vilaplana, R. Wiest, Y. Xu, Z. Xu, G. Zeng, J. Zhang, G. Zheng, R. Heinen, C. Chen, W. van der Flier, F. Barkhof, M. A. Viergever, G. J. Biessels, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, and M. J. Cardoso, "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 2556–2568, Nov. 2019.

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

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[13] Q. Li, Y. Yang, C. Reis, T. Tao, W. Li, X. Li, and J. H. Zhang, "Cerebral small vessel disease," *Cell Transplantation*, vol. 27, pp. 1711–1722, Sept. 2018.

[14] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, Dec. 2020.

[15] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Schock, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, and K. Maier-Hein, "batchgenerators - a python framework for data augmentation," 2020.

[16] M. M. Ghazi and M. Nielsen, "Fast-aid brain: Fast and accurate segmentation tool using artificial intelligence developed for brain," 2022.

[17] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, Jan. 2020.

[18] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016.

[19] A. Schiavone, S. N. Llambias, J. Johansen, S. Ingala, A. Pai, M. Nielsen, and M. M. Ghazi, "Robust identification of white matter hyperintensities in uncontrolled settings using deep learning," in *Medical Imaging with Deep Learning, short paper track*, 2023.

[20] R. D. Fields, "White matter matters," *Scientific American*, vol. 298, pp. 54–61, Mar. 2008.

[21] Y. Wang, Y. Pan, and H. Li, "What is brain health and why is it important?," *BMJ*, p. m3683, Oct. 2020.

[22] M. Duering, G. J. Biessels, A. Brodtmann, C. Chen, C. Cordonnier, F.-E. de Leeuw, S. Debette, R. Frayne, E. Jouvent, N. S. Rost, A. ter Telgte, R. Al-Shahi Salman, W. H. Backes, H.-J. Bae, R. Brown, H. Chabriat, A. De Luca, C. deCarli, A. Dewenter, F. N. Doubal, M. Ewers, T. S. Field, A. Ganesh, S. Greenberg, K. G. Helmer, S. Hilal, A. C. C. Jochems, H. Jokinen, H. Kuijf, B. Y. K. Lam, J. Lebenberg, B. J. MacIntosh, P. Maillard, V. C. T. Mok, L. Pantoni, S. Rudilosso, C. L. Satizabal, M. D. Schirmer, R. Schmidt, C. Smith, J. Staals, M. J. Thrippleton, S. J. van Veluw, P. Vemuri, Y. Wang, D. Werring, M. Zedde, R. O. Akinyemi, O. H. Del Brutto, H. S. Markus, Y.-C. Zhu, E. E. Smith, M. Dichgans, and J. M. Wardlaw, "Neuroimaging standards for research into small vessel disease—advances since 2013," *The Lancet Neurology*, vol. 22, no. 7, pp. 602–618, 2023.

[23] L. Pantoni, "Pathophysiology of age-related cerebral white matter changes," *Cerebrovascular Diseases*, vol. 13, no. Suppl. 2, pp. 7–10, Mar. 2002.

[24] J. P. Appleton, L. J. Woodhouse, A. Adami, J. L. Becker, E. Berge, L. A. Cala, A. M. Casado, V. Caso, H. K. Christensen, R. A. Dineen, J. Gommans, P. Koumellis, S. Szatmari, N. Sprigg, P. M. Bath, and J. M. W. and, "Imaging markers of small vessel disease and brain frailty, and outcomes in acute stroke," *Neurology*, vol. 94, pp. e439–e452, Dec. 2019.

[25] M. K. Georgakis, M. Duering, J. M. Wardlaw, and M. Dichgans, "WMH and long-term outcomes in ischemic stroke," *Neurology*, vol. 92, pp. e1298–e1308, Feb. 2019.

[26] F. Fazekas, F. Barkhof, L. Wahlund, L. Pantoni, T. Erkinjuntti, P. Scheltens, and R. Schmidt, "CT and MRI rating of white matter lesions," *Cerebrovascular Diseases*, vol. 13, no. Suppl. 2, pp. 31–36, 2002.

[27] N. D. Prins and P. Scheltens, "White matter hyperintensities, cognitive impairment and dementia: an update," *Nature Reviews Neurology*, vol. 11, pp. 157–165, Feb. 2015.

[28] F. Fazekas, J. Chawluk, A. Alavi, H. Hurtig, and R. Zimmerman, "MR signal abnormalities at 1.5 t in alzheimer's dementia and normal aging," *American Journal of Roentgenology*, vol. 149, pp. 351–356, Aug. 1987.

[29] L. Pantoni, M. Simoni, G. Pracucci, R. Schmidt, F. Barkhof, and D. Inzitari, "Visual rating scales for age-related white matter changes (leukoaraiosis)," *Stroke*, vol. 33, pp. 2827–2833, Dec. Dec. 2002.

[30] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*. 02 2007.

[31] N. Sharma, A. Ray, K. Shukla, S. Sharma, S. Pradhan, A. Srivastva, and L. Aggarwal, "Automated medical image segmentation techniques," *Journal of Medical Physics*, vol. 35, no. 1, p. 3, 2010.

[32] L. Zhou, M. Fan, C. Hansen, C. R. Johnson, and D. Weiskopf, "A review of three-dimensional medical image visualization," *Health Data Science*, vol. 2022, Jan. 2022.

[33] R. B. Buxton, *Introduction to Functional Magnetic Resonance Imaging*. Cambridge University Press, Aug. 2009.

[34] M. Larobina and L. Murino, "Medical image file formats," *Journal of Digital Imaging*, vol. 27, pp. 200–206, Dec. 2013.

[35] R. Cárdenes, R. de Luis-García, and M. Bach-Cuadra, "A multidimensional segmentation evaluation for medical image data," *Computer Methods and Programs in Biomedicine*, vol. 96, pp. 108–124, Nov. 2009.

[36] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, July 1945.

[37] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review," *Neuroinformatics*, vol. 13, pp. 261–276, Feb. 2015.

[38] C. R. Jack, P. C. O'Brien, D. W. Rettman, M. M. Shiung, Y. Xu, R. Muthupillai, A. Manduca, R. Avula, and B. J. Erickson, "FLAIR histogram segmentation for measurement of leukoaraiosis volume," *Journal of Magnetic Resonance Imaging*, vol. 14, no. 6, pp. 668–676, 2001.

[39] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," *NeuroImage*, vol. 21, pp. 1037–1044, Mar. 2004.

[40] F. Kruggel, J. S. Paul, and H.-J. Gertz, "Texture-based segmentation of diffuse lesions of the brain's white matter," *NeuroImage*, vol. 39, pp. 987–996, Feb. 2008.

[41] M. Gaubert, A. Dell'Orco, C. Lange, A. Garnier-Crussard, I. Zimmermann, M. Dyrba, M. Duering, G. Ziegler, O. Peters, L. Preis, J. Priller, E. J. Spruth, A. Schneider, K. Fliessbach, J. Wiltfang, B. H. Schott, F. Maier, W. Glanz, K. Buerger, D. Janowitz, R. Perneczky, B.-S. Rauchmann, S. Teipel, I. Kilimann, C. Laske, M. H. Munk, A. Spottke, N. Roy, L. Dobisch, M. Ewers, P. Dechent, J. D. Haynes, K. Scheffler, E. Düzel, F. Jessen, and M. W. and, "Performance evaluation of automated white matter hyperintensity segmentation algorithms in a multicenter cohort on cognitive impairment and dementia," *Frontiers in Psychiatry*, vol. 13, Jan. 2023.

[42] L. Griffanti, G. Zamboni, A. Khan, L. Li, G. Bonifacio, V. Sundaresan, U. G. Schulz, W. Kuker, M. Battaglini, P. M. Rothwell, and M. Jenkinson, "BIANCA (brain intensity AbNormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities," *NeuroImage*, vol. 141, pp. 191–205, Nov. 2016.

[43] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, and M. Mühlau, "An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis," *NeuroImage*, vol. 59, pp. 3774–3783, Feb. 2012.

[44] P. Schmidt, *Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging*. PhD thesis, Ludwig-Maximilians-Universität München, Fakultät für Mathematik, Informatik und Statistik, 2016.

[45] G. Park, J. Hong, B. A. Duffy, J.-M. Lee, and H. Kim, "White matter hyperintensities segmentation using the ensemble u-net with multi-scale highlighting foregrounds," *NeuroImage*, vol. 237, p. 118140, Aug. 2021.

[46] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation* (D. E. Rumelhart, J. L. McClelland, and the PDP research group, eds.), MIT Press, 1986.

[47] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[48] David H. Hubel, "Single unit activity in striate cortex of unrestrained cats," *the Journal of Physiology*, p. 147, 1959.

[49] Kunihiko Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybernetics 36*, p. 193–202, 1980.

[50] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016.

[51] Y. Lecun; L. Bottou; Y. Bengio; P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE (Volume: 86, Issue: 11, Nov. 1998)*, pp. 2278 – 2324, 1998.

[52] SPM12, "Spm12 documentation manual," 2023.

[53] T. C. Arnold, C. W. Freeman, B. Litt, and J. M. Stein, "Low-field mri: Clinical promise and challenges," 1 2023.

[54] M. Sarracanie, C. D. LaPierre, N. Salameh, D. E. Waddington, T. Witzel, and M. S. Rosen, "Low-cost high-performance mri," *Scientific Reports*, vol. 5, p. 15177, 2015.

[55] J. Turpin, P. Unadkat, J. Thomas, N. Kleiner, S. Khazanehdari, S. Wanchoo, K. Samuel, B. O. Moclair, K. Black, A. R. Dehdashti, R. K. Narayan, R. Temes, and M. Schulder, "Portable magnetic resonance imaging for ICU patients," *Critical Care Explorations*, vol. 2, p. e0306, Dec. 2020.

[56] Y. Anzai and L. Moy, "Point-of-care low-field-strength MRI is moving beyond the hype," *Radiology*, vol. 305, pp. 672–673, Dec. 2022.

[57] M. M. Yuen, A. M. Prabhat, M. H. Mazurek, I. R. Chavva, A. Crawford, B. A. Cahn, R. Beekman, J. A. Kim, K. T. Gobeske, N. H. Petersen, G. J. Falcone, E. J. Gilmore, D. Y. Hwang, A. S. Jasne, H. Amin, R. Sharma, C. Matouk, A. Ward, J. Schindler, L. Sansing, A. de Havenon, A. Aydin, C. Wira, G. Sze, M. S. Rosen, W. T. Kimberly, and K. N. Sheth, "Portable, low-field magnetic resonance imaging enables highly accessible and dynamic bedside evaluation of ischemic stroke," *Science Advances*, vol. 8, Apr. 2022.

[58] T. M. V. Runge and J. T. Heverhagen, "The clinical utility of magnetic resonance imaging according to field strength, specifically addressing the breadth of current state-of-the-art systems, which include 0," 2021.

[59] A. M. Prabhat, A. L. Crawford, M. H. Mazurek, M. M. Yuen, I. R. Chavva, A. Ward, W. V. Hofmann, N. Timario, S. R. Qualls, J. Helland, C. Wira, G. Sze, M. S. Rosen, W. T. Kimberly, and K. N. Sheth, "Methodology for low-field, portable magnetic resonance neuroimaging at the bedside," *Frontiers in Neurology*, vol. 12, Dec. 2021.

[60] S. S. Bhat, T. T. Fernandes, P. Poojar, M. S. Ferreira, P. C. Rao, M. C. Hanumantharaju, G. Ogbole, R. G. Nunes, and S. Geethanath, "Low-field MRI of stroke: Challenges and opportunities," *Journal of Magnetic Resonance Imaging*, vol. 54, pp. 372–390, Aug. 2020.

[61] M. L. D. L. D. Bouter, G. Ippolito, T. P. A. O'reilly, R. F. Remis, M. B. V. Gijzen, and . A. G. Webb, "Deep learning-based single image super-resolution for low-field mr brain images," 2022.

[62] N. Koonjoo, B. Zhu, G. C. Bagnall, D. Bhutto, and M. S. Rosen, "Boosting the signal-to-noise of low-field mri with deep learning image reconstruction," *Scientific Reports |*, vol. 11, p. 8248, 2021.

[63] J. D. Rudie, T. Gleason, M. J. Barkovich, D. M. Wilson, A. Shankaranarayanan, T. Zhang, L. Wang, E. Gong, G. Zaharchuk, and J. E. Villanueva-Meyer, "Clinical assessment of deep learning-based super-resolution for 3d volumetric brain mri," *Radiology: Artificial Intelligence*, vol. 4, 2022.

[64] M. Hori, A. Hagiwara, M. Goto, A. Wada, and S. Aoki, "Low-field magnetic resonance imaging its history and renaissance," 2021.

[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[66] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2015.

[67] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[69] J. Shen, "Tools for nifti and analyze image," 2014.

[70] C. Chen, W. Bai, and D. Rueckert, "Multi-task learning for left atrial segmentation on ge-mri," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 292–301, Springer, 2018.

[71] J. E. Iglesias, C. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *IEEE transactions on medical imaging*, vol. 30, pp. 1617–1634, 2011.

[72] A. E. Theyers, M. Zamyadi, M. O'Reilly, R. Bartha, S. Symons, G. M. MacQueen, S. Hassel, J. P. Lerch, E. Anagnostou, R. W. Lam, B. N. Frey, R. Milev, D. J. M"uller, S. H. Kennedy, C. J. Scott, and S. C. Strother, "Multisite comparison of mri defacing software across multiple cohorts," *Frontiers in Psychiatry*, vol. 12, p. 617997, 2021.

[73] H. M. Ali, "Mri medical image denoising by fundamental filters," in *High-Resolution Neuroimaging-Basic Physical Principles and Clinical Applications*, pp. 111–124, IntechOpen, 2018.

[74] L. Erasmus, D. Hurter, M. Naude, H. Kritzinger, and S. Acho, "A short overview of mri artefacts," *SA Journal of Radiology*, vol. 8, 2004.

[75] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *Neuroimage*, vol. 7, no. 4, p. S163, 1998.

[76] N. Khalili, N. Lessmann, E. Turk, N. Claessens, R. de Heus, T. Kolk, M. A. Viergever, M. J. Benders, and I. Isgum, "Automatic brain tissue segmentation in fetal mri using convolutional neural networks," *Magnetic Resonance Imaging*, vol. 64, pp. 77–89, 2019.

[77] C. Hui, Y. Zhou, and P. Narayana, "Fast algorithm for calculation of inhomogeneity gradient in magnetic resonance imaging data," *Journal of Magnetic Resonance Imaging*, vol. 32, no. 5, pp. 1197–1208, 2010.

[78] D. Moratal, A. Valles-Luch, L. Mart'ı-Bonmat'ı, and M. E. Brummer, "k-space tutorial: an mri educational tool for a better understanding of k-space," *Biomedical Imaging and Intervention Journal*, vol. 4, no. 1, p. e5, 2008.

[79] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition*, Citeseer, 2003.

[80] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," *CoRR*, vol. abs/1701.02096, 2017.

[81] M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds., *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*. Springer International Publishing, 2018.