



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

**A COMPARISON OF DEEP LEARNING
ARCHITECTURES FOR THE SEMANTIC
SEGMENTATION OF CHOROID PLEXUS FROM BRAIN
MRI: APPLICATION TO MULTIPLE SCLEROSIS**

Relatore:

Dott. Ing. Marco Castellaro

Laureanda:

Valentina Visani

ANNO ACCADEMICO 2021 – 2022

Data di laurea: 14 aprile 2022

A chi mi ha voluto bene

*“Sii paziente verso tutto ciò
che è irrisolto nel tuo cuore e...
cerca di amare le domande, che sono simili a
stanze chiuse a chiave e a libri scritti
in una lingua straniera.
Non cercare ora le risposte che possono esserti date
poiché non saresti capace di convivere con esse.
E il punto è vivere ogni cosa. Vivere le domande ora.
Forse ti sarà dato, senza che tu te ne accorga,
di vivere fino al lontano
giorno in cui avrai la risposta”*

Rainer Maria Rilke, *Lettera ad un giovane poeta*

ABSTRACT

The Choroid Plexus (ChP) is a vascular tissue located inside the brain ventricles. The increased interest on the ChP is related to the recent discoveries on its immunological function in inflammatory processes. ChP volume (ChPV) measured by T1-w MRI has been observed altered in several neurological disorders (i.e., Multiple Sclerosis, Major Depressive Disorder, Alzheimer's Disease, psychotic disorders). Therefore, ChPV can become a promising biomarker to improve the understanding of neurological diseases. However, the manual segmentation of ChP, that is the ground truth (GT), is time-consuming and affected by inter-operator variability due to the complexity of the task. FreeSurfer (FS) and the Gaussian Mixture Model Method (GMM) are the automatic methods proposed in literature.

The aim of this work is to propose a method for the completely automatic, accurate, reliable, and fast semantic segmentation of the ChP based on novel deep learning neural networks (DNN). The main goal is to find the combinations of parameters that maximize the performance indices with respect to the gold-standard manual segmentation depicted over T1-w MRI image with contrast injection, without the use of sequences with contrast agents to make this task less invasive for the patient.

The dataset analyzed is composed of 60 relapsing-remitting Multiple Sclerosis (RR-MS) patients, divided into a training set (45) and a validation set (15). The tested DNN are: 3D U-Net, V-Net, nnU-Net and UNETR. The training parameters and configurations that have been tested are: the input MRI sequence (T1-w, FLAIR, FLAIR+T1-w) and the GT segmentation (respectively, T1-w, FLAIR, T1-w or FLAIR; gold-standard cT1-w), the preprocessing with data augmentation transformations, the patch size (64x64x64, 96x96x96, 128x128x128) and the loss function (Cross-Entropy, Weighted Cross-Entropy, Dice, Dice-CE). The analyzed performance indices are: Dice Coefficient, Jaccard Index, 95% Hausdorff Distance, Percentage Volume Difference, Root Mean Squared Error (RMSE).

The preliminary analysis over the GT of T1-w and FLAIR sequences, and the two automatic segmentations of FS and GMM with respect to the gold-standard one has demonstrated that FS and GMM are quite inaccurate with respect to the manual segmentations. This consideration has corroborated the need to propose an alternative automatic method to segment the ChP. Moreover, on one hand, T1-w sequence is to be preferred to use the ChPV as a quantitative biomarker, because it has the lower Percentage

Volume Difference; on the other hand, FLAIR sequence lowers the variability of the resulted segmentation, as showed by the higher Dice Coefficient. The training analyses over all 672 possible combinations of DNNs have shown the better performances of nnU-Net and UNETR during the segmentation task. It was not possible to delineate the best combination of DNN parameters that could be equally suitable for each performance indices. Nevertheless, the most significant observations are that it is suggested avoiding the use of V-Net and Weighted Cross-Entropy. Making the comparison with the gold-standard segmentation, UNETR is slightly superior to nnU-Net and has brought to a Percentage Volume Difference on the validation set around 8% over T1-w images, trained both with cT1-w MSeg and T1-wMSeg. These results are remarkable and let its use on large clinical dataset, where the magnitude of Volume Difference between MS patients and healthy controls is around 21%.

To conclude, UNETR is a reliable tool for the segmentation of the ChP using the T1-w images and shows its promising usefulness to establish a new neuroimaging biomarker without the use of invasive techniques.

Keywords: deep learning, choroid plexus, multiple sclerosis, MRI, segmentation

INDEX

1	INTRODUCTION.....	1
1.1	The Choroid Plexus.....	2
1.2	The Choroid Plexus role in Multiple Sclerosis Disease	4
1.3	The aim of the study.....	6
2	STATE OF THE ART OF IMAGE SEMANTIC SEGMENTATION.....	7
2.1	The gold standard for the ChP segmentation	7
2.2	The importance of an automatic segmentation.....	8
2.3	Overview of the State-of-the-Art automatic semantic segmentation techniques	9
2.3.1	FreeSurfer.....	9
2.3.2	Gaussian Mixture Model (GMM)	10
2.4	Deep Learning Neural Networks for Image Semantic Segmentation	14
2.4.1	Optimizer: Adam.....	15
2.4.2	DNN Architectures for Semantic Segmentation	16
2.5	U-Net.....	18
2.6	V-Net.....	20
2.7	UNETR: Transformers for Image Segmentation Tasks	21
2.7.1	UNETR structure.....	21
2.8	nnU-Net (Dyn U-Net): a self-configuring DNN	24
3	MATERIALS AND METHODS	26
3.1	Dataset: MRI scans and general description	26
3.2	Preliminary Analysis for the Dataset composition.....	27
3.2.1	Manual segmentation	30
3.2.2	Preliminary Analysis	30
3.3	MONAI	30
3.4	Preprocessing of the images.....	31
3.4.1	Data Augmentation	32
3.5	Training Parameters	32
3.6	Loss Function.....	34
3.6.1	Cross-Entropy Loss and Weighted Cross-Entropy Loss	34
3.6.2	Generalized Dice Loss	36
3.6.3	DiceCE Loss.....	37
3.7	Validation Procedure: Performance Evaluation	38
3.7.1	Dice Coefficient	38
3.7.2	Jaccard Coefficient.....	39
3.7.3	Hausdorff Distance.....	40
3.7.4	Percentage Volume Difference	40
3.7.5	RMSE and MSE.....	41
3.7.6	Pearson’s Correlation Analysis	41

3.7.7	Linear Regression and OLS	43
4	RESULTS	44
4.1	Preliminary Analysis on the sub-dataset	44
4.2	Preliminary Analysis on the whole dataset	47
4.3	DNN Validation set results	54
4.3.1	Legend.....	54
4.4	DNN Validation set results: training with ground truth with contrast	55
4.4.1	Input T1-w, GT cT1-w - reference cT1-w.....	55
4.4.2	Input FLAIR, GT cT1-w - reference cT1-w.....	58
4.4.3	Input T1-w+FLAIR, GT cT1-w - reference cT1-w.....	60
4.5	DNN Validation set results: training with ground truth without contrast, performance indices calculated with respect to the contrast ground truth	63
4.5.1	Input T1-w, GT T1-w – reference cT1-w.....	63
4.5.2	Input FLAIR, GT FLAIR – reference cT1-w.....	65
4.5.3	Input T1-w+FLAIR, GT T1-w – reference cT1-w.....	67
4.5.4	Input T1-w+FLAIR, GT FLAIR – reference cT1-w.....	69
4.6	Performance Analysis: comparison between models trained with GT cT1-w and those with GT without contrast (T1-w, FLAIR).....	71
4.6.1	Input T1-w, GT T1-w compared with Input T1-w, GT cT1-w	72
4.6.2	Input FLAIR, GT FLAIR compared with Input FLAIR, GT cT1-w.....	73
4.6.3	Input T1-w+FLAIR, GT T1-w and Input T1-w+FLAIR, GT FLAIR compared with Input T1-w+FLAIR, GT cT1-w	74
4.7	Comparison with the state-of-the-art automatic methods FS and GMM	75
4.8	Visual inspection of the predicted segmentations using the proposed DNN	84
5	DISCUSSION	87
5.1	Preliminary Analysis on the sub-dataset	87
5.2	Preliminary Analysis on the whole dataset	88
5.3	DNN validation set results	89
5.4	Performance Analysis: comparison between models trained with GT cT1-w and those with GT without contrast (T1-w, FLAIR).....	91
5.5	Comparison with the state-of-the-art automatic methods FS and GMM	93
5.6	Visual inspection of the predicted segmentations using the proposed DNN	94
6	CONCLUSION.....	95
6.1	Future directions.....	97
Appendix A		98
Appendix B		109
List of Acronyms.....		140
BIBLIOGRAPHY		142

1 INTRODUCTION

The Choroid Plexus (ChP) is a brain vascular tissue located in the ventricles of the brain. Recently, research studies have shown its important role in brain homeostasis and its relation to several diseases (Althubaity et al., 2022; Fleischer et al., 2021; Lizano et al., 2019; Marques et al., 2017; Ricigliano et al., 2021; Vercellino et al., 2008). In fact, the ChP produces the majority of the Cerebrospinal Fluid (CSF), as shown in *Figure 1.1*, that is a mediator of the brain clearance pathways.

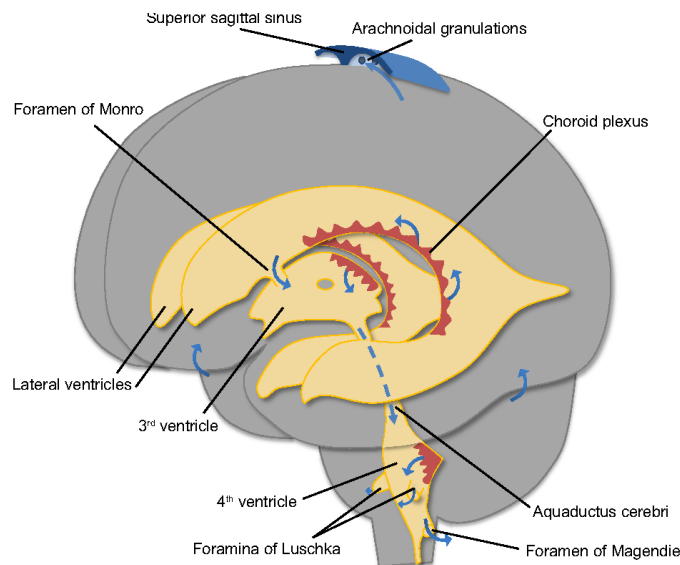


Figure 1.1: ChP inside the ventricles and CSF production, adapted from Cerebrospinal fluid secretion by the choroid plexus (Damkier et al., 2013).

Therefore, the functional and anatomical modification of the ChP can lead to alterations that characterize, for example, the inflammatory state in Multiple Sclerosis patients (Ricigliano et al., 2021; Vercellino et al., 2008) or protein accumulation in Alzheimer's disease (Tadayon, Pascual-Leone, et al., 2020) or worse symptoms in psychosis, like schizophrenia or bipolar disorder (Lizano et al., 2019). It is worth noting that the ChP volume calculated from both automatic or manual segmentation can be used as a biomarker to both improve the diagnosis or to stage the actual inflammatory activity of Multiple Sclerosis. However, the manual segmentation, that is the gold-standard approach for the ChP segmentation, is time-consuming and operator dependent. Therefore, the automatic segmentation of this structure can be a very promising tool to investigate ChP alterations in neurological disorders.

This chapter will report anatomical structure and function of the ChP. The last paragraph explains the aim of this work. The state of the art of the Choroid Plexus segmentation starting from brain Magnetic Resonance Imaging (MRI) will be detailed in the second chapter. In the following chapters the dataset used in this work will be described. The methods implemented, and the results obtained will be introduced and discussed in terms of qualitative as well as quantitative performances.

1.1 THE CHOROID PLEXUS

The Choroid Plexus (ChP) is a vascular tissue located inside the brain ventricular system (*Figure 1.2*). The ChP is present in all four ventricles: two laterals, the third and the fourth ventricles. This vascular tissue forms a major part of the Blood-CSF-Barrier (BCSFB), while a remainder part is composed by the arachnoid membrane and the circumventricular organs (Damkier et al., 2013).

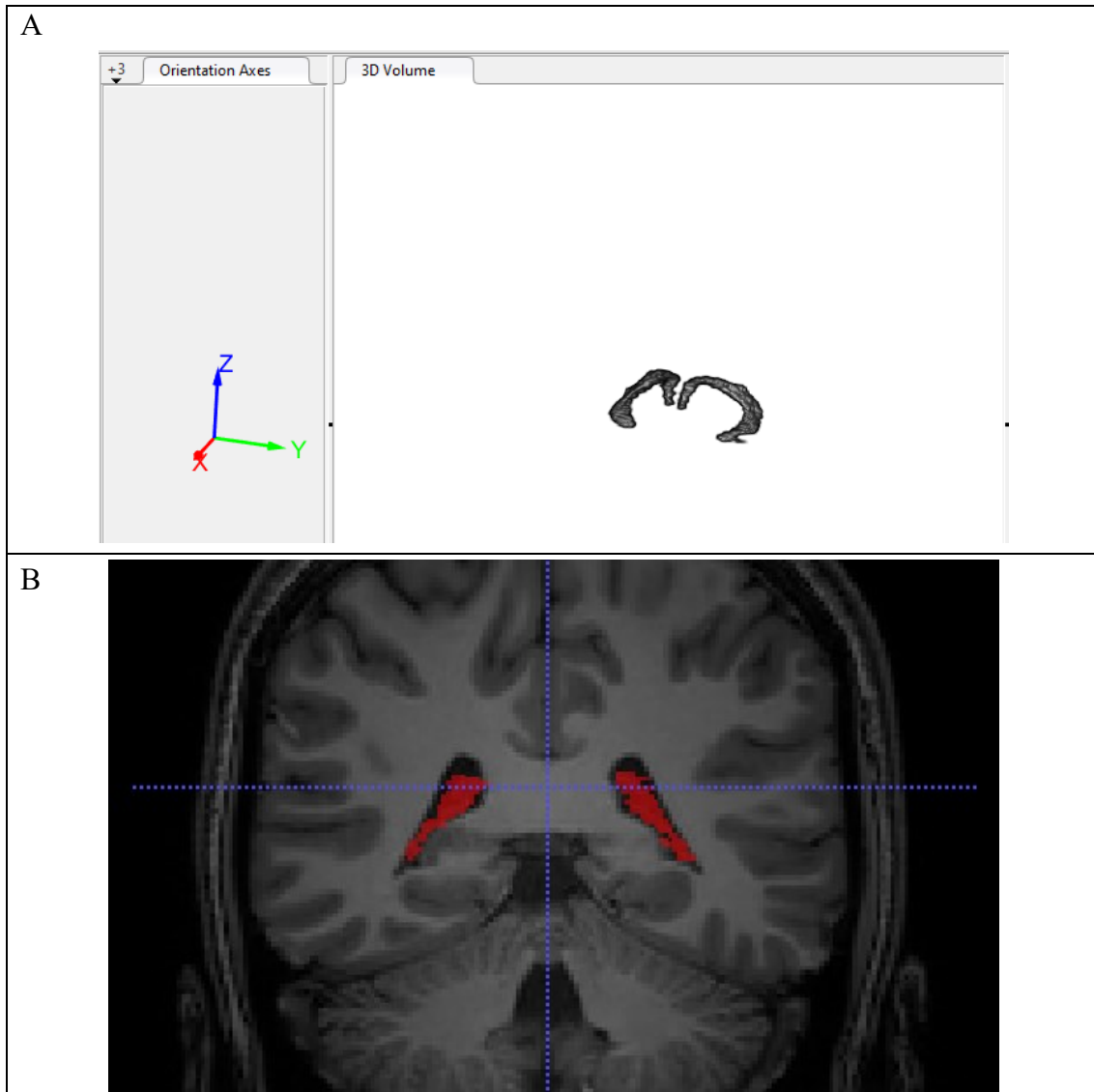


Figure 1.2: (A) 3D Volume of the ChP - segmented by cT1-w image with contrast, MATLAB VolumeViewer; (B) T1-w image with the segmentation of ChP - ITK-SNAP.

1.Introduction

The ChP is a monolayer of epithelial cells (basement membrane) (*Figure 1.3*), but it has also a stromal compartment (Balusu et al., 2016; Damkier et al., 2013; Lizano et al., 2019). In particular, the choroid plexus epithelial cells surround the capillaries and form the BCSFB. The tight junctions allow the connection between the Brain-blood-barrier (BBB) endothelial cells and the epithelial one of the BCSFB to regulate the passage of substances from the blood to the brain. Blood is perfused into the ChP through the internal carotid arteries and the vertebral artery with a perfusion rate of $4 \frac{mL}{min * gr_{tissue}}$ in healthy conditions: ten times higher than that for the brain parenchyma.

The main role of the ChP is the production of CSF with a rate of $0.4 \frac{mL}{min * gr_{tissue}}$ (Damkier et al., 2013). Moreover, the ChP supports further features (Balusu et al., 2016; Lassmann, 2019; Spector et al., 2015):

- active and passive transport of substances between BCSFB and BBB (tight junctions)
- intercellular communication through gap junctions
- maintenance of brain homeostasis through the CSF production
- clearance pathways in the brain for drugs, proteins, and waste products
- immunological function implicated in the inflammatory processes

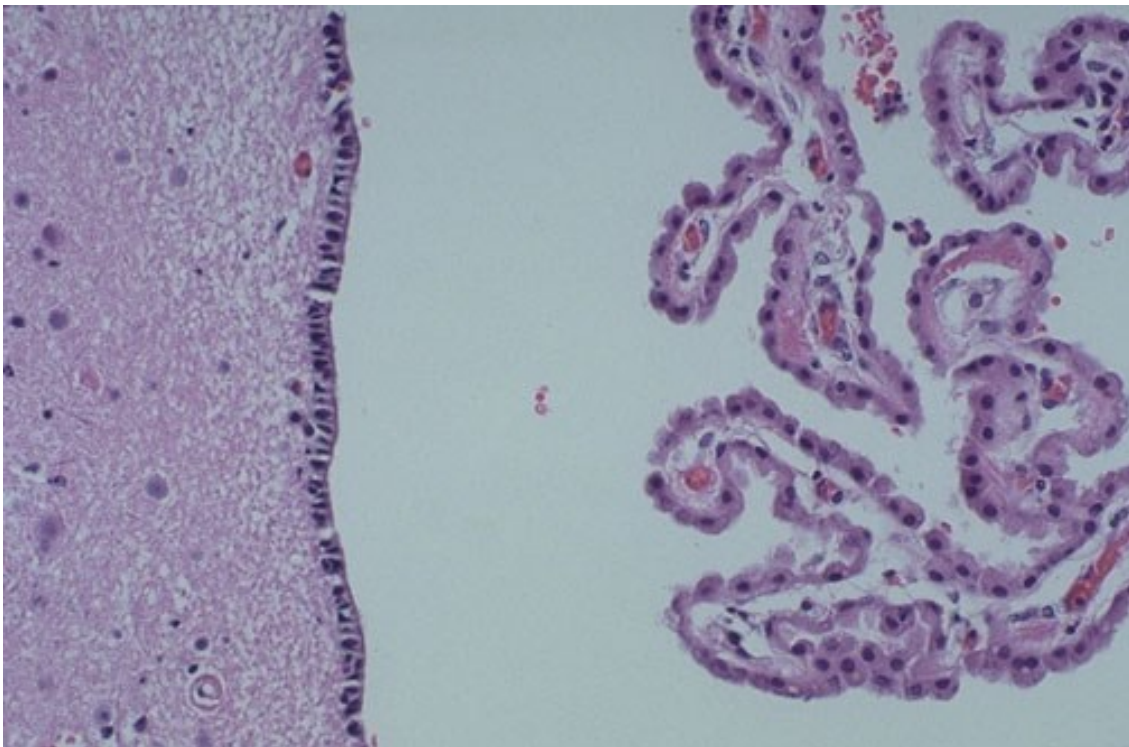


Figure 1.3: Normal choroid plexus, with an epithelial layer around a stroma with prominent blood vessels. This is where the cerebrospinal fluid is formed. The ependymal cells lining the ventricle are present at the left (<https://webpath.med.utah.edu/>).

As regards the immunological functions, several studies (Engelhardt et al., 2001; Lassmann, 2019; Vercellino et al., 2008) suggested the implication of the BCSFB, and the ChP as consequence, in inflammatory processes of the Central Nervous System (CNS), regulating the T and B- cells immunological response in some autoimmune diseases like Multiple Sclerosis, as explained in the paragraph 1.2.

It was theorized that an alteration of the ChP functions can be involved in neurodegenerative pathologies. To make an example, on the epithelial surface, the ChP expresses some receptors that are fundamental for the clearance of the Amyloid-Beta in the Alzheimer's Disease (Balusu et al., 2016; Tadayon, Moret, et al., 2020). Moreover, a recent study (Carloni et al., 2021) has described a vascular barrier inside the ChP that can be modified accordingly to the intestinal inflammation state thanks to molecules with bacterial origin.

Lastly, people affected by MDD (Major Depressive Disorder) show a higher ChP volume with respect to the controls and the ChP volume is higher in presence of CNS inflammation, measured through 3D MRI and Positron Emission Tomography (PET) imaging techniques (Althubaity et al., 2022).

Thanks to these discoveries, the interest on the ChP has grown faster, so arise the need to correctly delineate this anatomical structure within the cerebral ventricles.

As briefly explained before, the volume of the ChP can reflect the state of functioning of the ChP and this seems to be related to the brain homeostasis. For example, Lizano et al. had found that the ChP volume is larger in patients with psychotic disorders with respect to healthy control subjects, moreover, the larger is the volume, the worse is the cognition (Lizano et al., 2019). In addition, Zhou et al. (Zhou et al., 2020) noted a larger ChP volume, above all in the lateral ventricles, in patients affected by schizophrenia, even if the cause-and-effect relationship between this clinical parameter and the disease is not clear yet.

1.2 THE CHOROID PLEXUS ROLE IN MULTIPLE SCLEROSIS DISEASE

Multiple Sclerosis (MS) is an autoimmune neurodegenerative disease that affects the CNS. The Immune System (IS) promotes an inflammatory state that can damage both the myelin sheets and the oligodendrocytes. This process, called demyelination process, leads to lose the myelin (plaque) in defined areas like the optic nerves, the spinal cord, around the ventricles (periventricular plaques) and randomly in the whole White Matter (WM). These plaques can linger in an inflammatory state or can evolve to a scar tissue (sclerosis). The diagnosis is confirmed by a multi-factorial diagnostic criterion (McDonald) that

incorporates clinical, biological, and imaging based biomarkers, detected by MRI (Thompson et al., 2018).

Lassmann (Lassmann, 2019) classifies two types of MS inflammation: Relapsing Remitting MS (RRMS) and Progressive (primary and secondary) MS (PPMS, SPMS). These two types of inflammations are closely related even if they can progress independently of each other. The RRMS is characterized by a fast T and B-cells accumulation causing BBB damage and, as consequence, the WM lesions. On the contrary, the PMS implies a very slow accumulation of T and B-cells without big BBB damages. The RRMS is present in the acute phase of the disease, while the PMS could be present from the beginning of the disease, but it becomes predominant over the years (Lassmann, 2019).

Recent studies have hypothesized an association between the MS inflammatory state and the ChP. Vercellino et al. (Vercellino et al., 2008) had found, in a histopathological study, an inflammatory activity inside the ChP in MS patients. Years after, Ricigliano et al. (Ricigliano et al., 2021) had found an increased volume of the ChP and a higher inflammatory state in Multiple Sclerosis (MS) patients, so it was hypothesized the important role of the ChP in the evolution of the disease, even if it has not been verified yet. A recent study (Fleischer et al., 2021) highlighted the absence of an in vivo demonstration about the implication of the ChP in the MS inflammatory process and its aim was studying the link between the ChP morphology in MS patients and in two mice models for CNS demyelination. The discoveries about the ChP enlargement in both species and an association between this parameter and the disease severity suggest using the ChP volume as a biomarker to evaluate the progression of the disease and the effectiveness of the therapy, becoming an image correlate of the inflammation state (Fleischer et al., 2021; Manouchehri & Stüve, 2021). Even if these findings are very interesting, a commentary study (Manouchehri & Stüve, 2021) specified that there are still many open questions about the correlation between the ChP volume, the staging of the MS, and the inflammation biological mechanisms, that can be answered only verifying Fleischer et al. hypotheses.

Consequently, the ChP segmentation is crucial to allow the progress of scientific research in this field of medicine.

1.3 THE AIM OF THE STUDY

As a matter of fact, the main fence to the study of the ChP in a big cohort of patients is the time-consuming characteristic of the manual segmentation (Schmidt-Mengin et al., 2021; Tadayon, Moret, et al., 2020). Moreover, the manual segmentation is extremely radiologist-dependent since the contrast between the ChP and the ventricles is not always optimal in T1-w images. For these reasons, some studies (Zhao et al., 2020) used as ground truth an automatic segmentation provided by the free software FreeSurfer, that has a lower accuracy with respect to the manual segmentation, manually correcting them, while other studies (Tadayon, Moret, et al., 2020) considered as ground truth the segmentation resulted by the majority vote approach applied to the manual segmentations performed by two different researchers.

To conclude, all the factors listed above demonstrate the need to use a new approach for the segmentation of the ChP. This work is focused on the automatic segmentation of the ChP. A cohort of Multiple Sclerosis patients was used as a case study, considering the increased attention on this inflammatory disease, however it can be considered a good proxy for the application of automated methods also to other pathological conditions or in case of healthy subjects. MRI acquired with a standard protocol was used to investigate which is the best deep learning neural network (DNN) structure, its configuration and learning parameters and strategies, to perform the best automatic ChP segmentation that can make the analysis of a huge cohort of MS patients faster, reliable, and reproducible. The DNN performance will be compared to the manual segmentation performed by a trained neuroradiologist with ITK-SNAP (this will be considered as ground truth) and to other available automated segmentation methods. The dataset is composed by several MRI sequences. Therefore, in addition, MRI sequences will be used as separate channel or combined in a multi-channel approach and the sequence mixture that provide the best results will also be investigated.

2 STATE OF THE ART OF IMAGE SEMANTIC SEGMENTATION

Currently, there is only one publication (Tadayon, Moret, et al., 2020) on the automatic segmentation of the ChP starting from MRI images. A single conference paper exploiting Deep Learning Network has been published (Zhao et al., 2020), and several further pre-prints are available which, however, have not been officially reviewed and published yet.

2.1 THE GOLD STANDARD FOR THE CHP SEGMENTATION

The structure and the function of the ChP have been investigated using different quantitative imaging modalities, like Diffusion Weighted Imaging (DWI) (Maekawa et al., 2019), perfusion imaging and PET (Althubaity et al., 2022; Schubert et al., 2019). However, for all these options there is the need of the segmentation derived by the anatomical MRI because of its higher abundance, contrast, and resolution compared to both quantitative MRI and PET.

The gold standard to estimate the volume of the ChP is to manually segment the ChP using T1-weighted MRI after the injection of contrast agent (cT1-w). The contrast agent uptake (usually Gadolinium) reveals the ChP that is a vasculature structure and enhances its contrast from the ventricle image intensity allowing a better delineation of the ChP. It is important to underline that the segmentation is performed only over the lateral ventricles. Indeed, the ChP is present also in the third and fourth ventricles, however it is not routinely segmented because of its size, that is much lower than that of ChP in the lateral ventricles, and therefore cannot always be detected and segmented from MRI images.

However, the main issue in the use of the contrast agent is that in lot of cases the risks outweigh the benefits, since the injection of the contrast agent is an invasive procedure that is always better to avoid if not strictly necessary. On the contrary, the sequence T1-weighted acquired without contrast injection (T1-w) can always be non-invasively acquired (Tadayon, Moret, et al., 2020). An MRI exam is composed by several sequences that allow to acquire images with different contrasts; for this reason, several mixtures of sequences will be tested with the aim to find which mixture provide the best performances when compared to the cT1-w image (gold-standard).

ITK-SNAP is an open-source software that have been used for the manual segmentation of the ChP from MRI images (<http://www.itksnap.org>) (Yushkevich et al., 2006, 2016). This tool requires an expert that manually paint the anatomical district of interest, in this case the ChP, using any MRI sequence. This procedure is time consuming and there is a high probability to deliver errors because the ChP is only 0,1% of the entire brain (Zhao et al., 2020), so the reliability and the robustness of the task depend on the skills and on the experience of the person responsible for the segmentation.

2.2 THE IMPORTANCE OF AN AUTOMATIC SEGMENTATION

Manual segmentation result depends on the ability of the operator, moreover the main concern that arises with MRI images is that they have a high resolution with good contrast ($\leq 1\text{mm}^3$), however they can be affected by the partial volume effect. ChP dimension can be lower than the resolution achievable with MRI and, therefore, blurred edges can occur, and this can hamper the segmentation of the ChP in the brain (Zhao et al., 2020). Moreover, a big cohort of patient is difficult to be studied using the manual segmentation due to the heavy workload.

The most used method to automatically segment the ChP is implemented in the open-source FreeSurfer suite (paragraph 2.3). Despite its large use, the segmentation of the ChP obtained with FreeSurfer is often poor and mistaken. This automatic segmentation method probably fails because the ChP is a highly variable anatomical structure from patient to patient (Zhao et al., 2020). As consequence, there is the needing of more accurate methods.

It has been verified that for medical image segmentation tasks, like brain tumors or multi organ segmentation or cell tracking-challenge (Hatamizadeh et al., 2021; Milletari et al., 2016; Ronneberger et al., 2015; Zhao et al., 2020), the Deep Neural Networks (DNN) have been successful. Other pre-print studies (Schmidt-Mengin et al., 2021; Tolstikhin et al., 2021) propose the use of the multi-layer perceptrons (Axial-MLPs); in particular, Schmidt-Mengin et al. find good results for the ChP segmentation task (Schmidt-Mengin et al., 2021).

As these new hypotheses have been advanced for the construction of algorithms for automatic segmentation, the manual segmentation is needed and used as the ground truth to train and validate the new segmentation approaches.

However, the main issue of these options is that they are not validated specifically for Multiple Sclerosis patients. The only study that is specific for Multiple Sclerosis is that

of Schmidt-Mengin et al. (Schmidt-Mengin et al., 2021), and this is the main asset of this work. This study considers 144 patients: 44 healthy controls, 61 with relapsing-remitting Multiple Sclerosis (RR-MS), 36 with progressive Multiple Sclerosis (P-MS). The manual segmentation over T1-w images (derived from two different MRI scanners) was performed by two neurologists over the whole dataset. The dataset is split into a training set, used for the training step with 5-fold cross-validation, and an independent test set, used to evaluate the performances at the end of the study.

Moreover, Zhao et al. have used as ground truth the results of the segmentations obtained with FreeSurfer and manually corrected by a radiologist and the dataset was limited to 10 healthy subjects (Zhao et al., 2020), while Tadayon et al. have used the ground truth derived by the manual segmentation only for some patients: 19 patients from the Siena Dataset, 20 subjects over 1067 from the Human Connectome Project dataset (HCP), 20 subjects with AD (Alzheimer’s Disease) over 509 from the Alzheimer’s Disease Neuroimaging Initiative dataset (ADNI). Nevertheless, they have used lot of data (derived from online dataset) and they have validated the results using AV1451 (tau) PET (Tadayon, Moret, et al., 2020).

2.3 OVERVIEW OF THE STATE-OF-THE-ART AUTOMATIC SEMANTIC SEGMENTATION TECHNIQUES

In this part we present the automatic segmentation techniques proposed in the literature to automatically segment the ChP. Only two automatic options are published in the literature: one based on FreeSurfer (Fischl, 2012), one on the Gaussian Mixture Model (Tadayon, Moret, et al., 2020). All other proposals are pre-prints, except the conference paper of Zhao et al. (Zhao et al., 2020), and they will be discussed in the last paragraphs.

2.3.1 FreeSurfer

FreeSurfer (FS) is an open-source software for the automatic segmentation of brain images (Fischl, 2012). FreeSurfer and Gaussian Mixture Model are the only two methods published in the literature that can automatically segment the Choroid Plexus. This is since this anatomical region had never aroused particular interest before this time.

FreeSurfer pipeline is mostly automatic, and the volume-based subcortical segmentation pipeline is the sequence of the following steps (<https://surfer.nmr.mgh.harvard.edu/>) (Fischl et al., 2002). This method gives the probability of having a certain label in a certain voxel, the probability of having a label in a voxel looking at the neighboring voxel labels and the corresponding probability distribution function (pdf, usually a normal

distribution) of each class in each voxel. In this way, passing to the test dataset, a new subject is taken to the normalized space and the voxel intensities of the new image are included in that of the common space to find the segmentation that maximizes the probability of having a certain label in a certain input voxel.

On the contrary, Schmidt-Mengin et al. applied FS without register the images in MNI305 space. Moreover, they also used FastSurfer, a deep-learning method that obtains the same results of FS, but it is quicker, and they considered both algorithms when making the final comparison between their proposal and the state-of-the-art automatic techniques (Schmidt-Mengin et al., 2021).

The main problem of FS method, as said before, is that it is quite inaccurate with respect to the manual segmentation performed with ITK-SNAP (*Figure 2.1*).

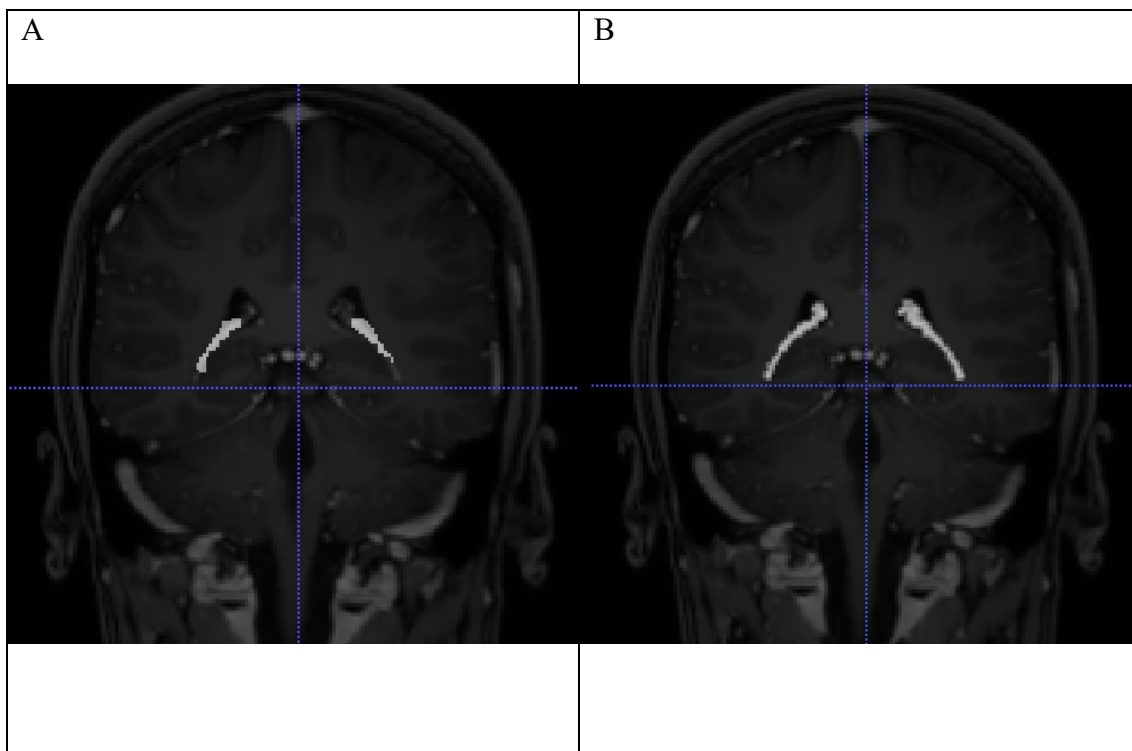


Figure 2.1: Coronal slice of cT1w sequence of a representative patient (#2060) Panel A) FreeSurfer segmentation B) Ground Truth. Images from ITK-SNAP.

2.3.2 Gaussian Mixture Model (GMM)

The Gaussian Mixture Model algorithm (Balafar, 2014; Greenspan et al., 2006; Tadayon, Moret, et al., 2020) is an unsupervised machine learning method. The hypothesis on which the GMM is based is that the intensities of the voxels derive from a mixture of a finite number of Gaussian Distributions (GD) (each class is embodied by a Gaussian). This algorithm considers both the voxel intensity and the spatial localization inside the brain to achieve a robust and accurate segmentation (Greenspan et al., 2006). An MRI image can be modelled as:

$$f(v_t|\theta) = \sum_{i=1}^n \alpha_i f_i(v_t|\mu_i, \Sigma_i)$$

v_t : feature vector for the t-th voxel

n : number of Gaussian functions f_i

θ : GMM parameter set

μ_i, Σ_i : mean and covariance of the i-th Gaussian function f_i

α_i : i-th Gaussian function weight

The feature vector v_t for each voxel is a 4D vector containing v^I (voxel intensity) and $v^{X,Y,Z} = (v^X, v^Y, v^Z)$ (spatial information). The set of feature vectors has dimension T, where T is the number of voxels. It is assumed that a single Gaussian describes a single tissue thanks to the lower intra-variability of v^I for a single voxel with respect to the intensity inter-variability among voxels. In this way, all Gaussian referred to the same tissues have the same intensity parameters. The grouping function performs this, passing from the Gaussians to the tissues:

$$\pi: \{1, \dots, n\} \rightarrow \{1, \dots, k\}$$

For the whole brain segmentation task, a voxel can belong only to three tissues: gray matter, white matter, and CSF. It is worth to underline that the spatial and the intensity features for each Gaussian are uncorrelated, so the mean and covariance are described as (Greenspan et al., 2006):

$$\mu_i = \begin{pmatrix} \mu_i^{XYZ} \\ \mu_{\pi(i)}^I \end{pmatrix} \quad \Sigma_i = \begin{pmatrix} \Sigma_i^{XYZ} & 0 \\ 0 & \Sigma_{\pi(i)}^I \end{pmatrix}$$

$\pi(i)$: tissue of the i-th Gaussian

μ_j^I, Σ_j^I : mean and variance parameters of all Gaussian related to the j-th tissue

Through the Expectation Maximization (EM) algorithm (Dempster et al., 1977) it is possible to derive the model parameters for the Gaussian functions (Greenspan et al., 2006):

$$w_{it} = p(i|v_t) = \frac{\alpha_i f_i(v_t|\mu_i, \Sigma_i)}{\sum_{i=1}^n \alpha_i f_i(v_t|\mu_i, \Sigma_i)}, \quad i = 1, \dots, n \quad t = 1, \dots, T$$

$$n_i = \sum_{t=1}^T w_{it}, \quad i = 1, \dots, n$$

$$k_j = \sum_{i \in \pi^{-1}(j)} n_i, \quad j = 1, \dots, k$$

2.State of the art of image semantic segmentation

n_i : number of voxels expected to be related to the i -th Gaussian

$\pi^{-1}(j)$: set of Gaussian linked to the j -th tissue

k_j : number of voxels expected to be in that tissue

At each step of the EM algorithm, the constrains are given to the intensity parameters as follows (Greenspan et al., 2006):

$$\alpha_i = \frac{n_i}{n}, \quad i = 1, \dots, n$$

$$\mu_i^{XYZ} = \frac{1}{n_i} \sum_{t=1}^T w_{it} v_t^{XYZ}$$

$$\Sigma_i^{XYZ} = \frac{1}{n_i} \sum_{t=1}^T w_{it} (v_t^{XYZ} - \mu_i^{XYZ})(v_t^{XYZ} - \mu_i^{XYZ})^T$$

$$\mu_j^l = \frac{1}{k_j} \sum_{i \in \pi^{-1}(j)} \sum_{t=1}^T w_{it} v_t^l, \quad j = 1, \dots, k$$

$$\Sigma_j^l = \frac{1}{k_j} \sum_{i \in \pi^{-1}(j)} \sum_{t=1}^T w_{it} (v_t^l - \mu_j^l)^2$$

Obviously, the EM algorithm doesn't change the link between a Gaussian function and the tissue described through it, but the voxels can change the tissue they belong to during this phase because the training takes place simultaneously on all tissues.

During this phase, each voxel is assigned to a Gaussian to build a feature space (estimated image representation). During the probabilistic segmentation phase, there is backward process to pass from the feature space to the original image. The procedure for assigning a voxel to one of the Gaussians of the model occurs by maximizing the posterior probability, and this brings directly to the final probabilistic segmentation. The CGMM (Constrain Gaussian Mixture Model) uses more than one Gaussian to describe a tissue, so the Bayes' rule for the posterior probability that a voxel belong to the j -th tissue is (Greenspan et al., 2006):

$$p(\text{tissue } j | v_t) = \frac{\sum_{i \in \pi^{-1}(j)} \alpha_i f_i(v_t | \mu_i, \Sigma_i)}{\sum_{i=1}^n \alpha_i f_i(v_t | \mu_i, \Sigma_i)}, \quad j = 1, \dots, k, \quad t = 1, \dots, T$$

The tissue that maximizes the posterior probability for each t -th voxel is:

$$tissue - label_t = \underset{j \in \{1, \dots, k\}}{\operatorname{argmax}} p(tissue\ j|v_t), \quad t = 1, \dots, T$$

tissue-label_t ∈ {1, ..., k} is one of the considered tissues. The final segmentation map is derived directly starting from the assignment of each voxel to a tissue through a label.

Tadayon et al. modify the GMM approach for the whole brain segmentation task to segment the ChP (Tadayon, Moret, et al., 2020). *Figure 2.2* shows the procedure pipeline. The starting point was the T1-w MRI image. First, a mask that covers CSF, ChP and the ventricular wall is created with FreeSurfer, then the ChP and the lateral ventricles are jointed (step 1). Second, it is applied a Bayesian GMM with two components to the masked voxels (step 2). In this way, two clusters are obtained: a major part of the first cluster is composed by CSF voxels (lower intensity), a major part of the second cluster is composed by the ChP and the ventricular wall voxels (higher intensity). Third, take the second cluster and apply 3D Susan Smoothing (sigma=1 mm, three dimensions) (step 3). In this way, the ventricular wall voxels are smoothed near to zero (as CSF in the mask), while ChP voxels near to one (as ChP in the mask). Finally, a Bayesian GMM with 3 components is applied and the voxels in the cluster with the highest average intensity values are classified as ChP voxels (step 4).

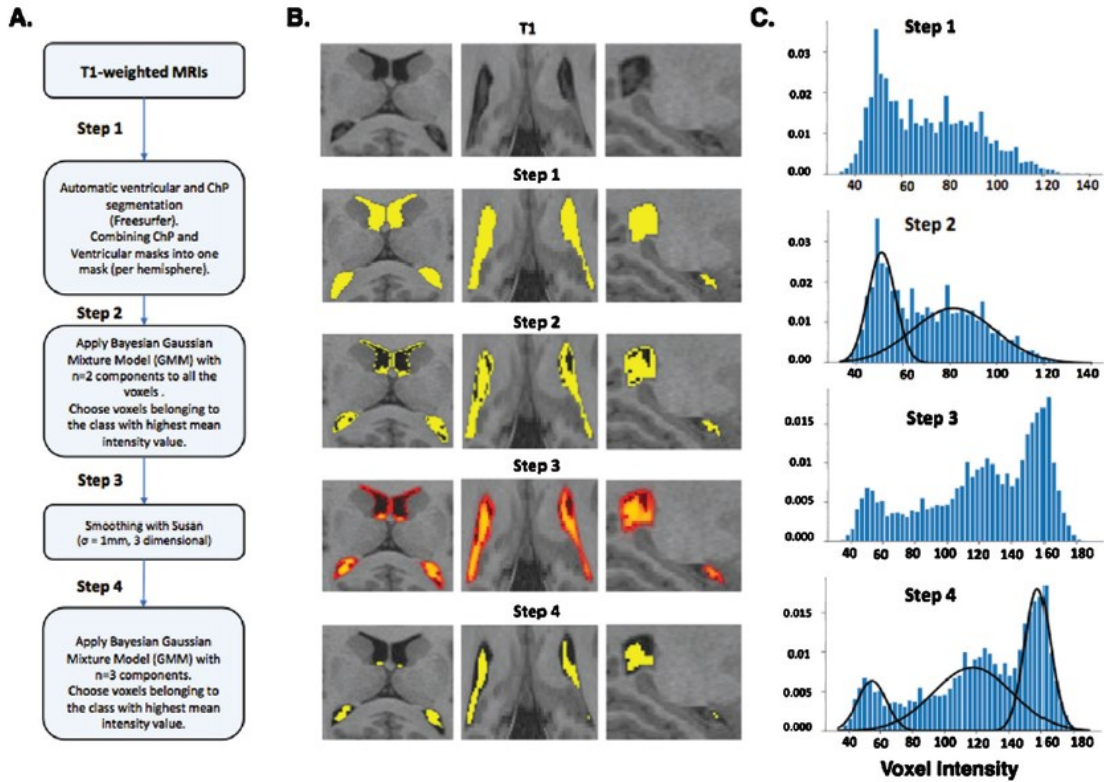


Figure 2.2: GMM segmentation pipeline (Tadayon, Moret, et al., 2020).

Figure 2.3 shows the comparison between the T1-w image, the FreeSurfer segmentation, the GMM segmentation and the validation with the PET study done by Tadayon et al.

GMM method is quite accurate if it is compared with FreeSurfer (Tadayon, Moret, et al., 2020).

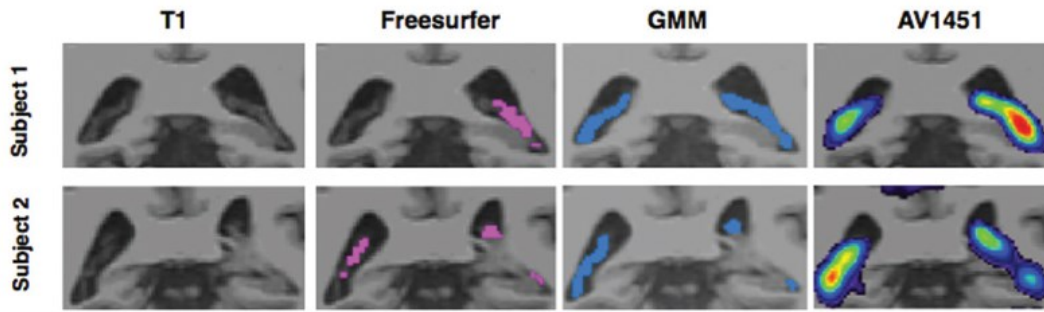


Figure 2.3: Comparison of segmentation methods in two subjects of ADNI3 dataset. Validation of the GMM results with AV1451 PET images (Tadayon, Moret, et al., 2020).

2.4 DEEP LEARNING NEURAL NETWORKS FOR IMAGE SEMANTIC SEGMENTATION

To perform an image based semantic segmentation means assigning each voxel of the image to a class to separate the image into label-uniform semantic regions. Due to this operative definition, the result depends not only on the image dataset, but also on the complexity of the task (Ghosh et al., 2019). The ChP segmentation is so complex that the automatic algorithms proposed in literature have poor results compared to the ground truth. For these reasons, the new goal has become to find an automatic alternative method to the proposed algorithms.

Recently, the focus has shifted to deep learning algorithms since they became the new state-of-the-art for Natural Language Processing (NLP), speech recognition, classification tasks, image recognition, and computed vision in general (Hatamizadeh et al., 2021; LeCun et al., 2015; Ruby Usha & Yendapalli Vamsidhar, 2020). The added value brought by the deep learning algorithms with respect to simple machine-learning algorithms is to consider many parameters and map the dataset into a higher dimensional space, making the algorithm capable of improving learning directly from the single input and not simply from the entire input training set.

A deep learning neural network architecture usually is a neural network architecture with more layers, so ‘deeper’. The structure is a combination of non-linear operations (layers) that work in parallel for which the output of the previous layer becomes the input of the next layer.

A general deep learning algorithm (LeCun et al., 2015) can be divided into two phases: training and validation. In the training phase, looking at the input, the DNN produces a vector of scores for each class as output. The final aim is to maximize the output values

for each class, and it is reached through an objective function that estimates the error between the real output and the reached one. The DNN modifies its internal parameters (also called weights) minimizing the estimated error. However, the procedure to modify the internal parameters consists in calculating the gradient vector of the error's trend as a function of weight's variation for each adjustable parameter. Consequently, each weight is adjusted in the direction that lowers the error. The considered objective function is the mean across all training samples and, since the goal is minimizing the output error, the target is the absolute minimum of this function, that can be reached looking at the negative gradient vector. This is the simplest procedure (and the first proposed) to update the weights, called Stochastic Gradient Descent (SGD), however, in the last few years it has been introduced the Adam optimizer (Kingma & Ba, 2015) and its weighted version AdamW (Loshchilov & Hutter, 2019). After the training, the performances of the DNN are tested over a new independent dataset (validation set). To update the internal parameters of the DNN the training considers also the backpropagation algorithm as showed in *Figure 2.4* (LeCun et al., 2015).

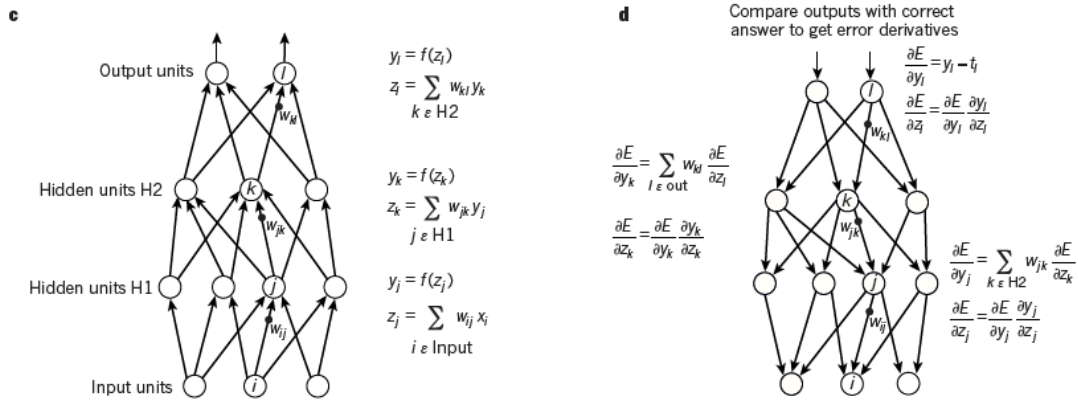


Figure 2.4: Multilayer neural networks and backpropagation, c) forward pass, d) backward pass (LeCun et al., 2015).

Usually, non-linear classifiers like ReLU, PReLU or sigmoid function are used to classify the input voxels. Also considering the high number of layers, a DNN becomes sensitive to the very detailed features without considering the irrelevant ones.

Lastly, Dosovitskiy et al. (Dosovitskiy et al., 2021) demonstrates the importance of adding an unsupervised pre-training step before the DNN, above all if the dataset is small. Moreover, this can help prevent overfitting (LeCun et al., 2015).

2.4.1 Optimizer: Adam

Adam is a novel version of the classic SGD method. It can be referred as a stochastic optimization algorithm used to update the neural network parameters in the training phase (Kingma & Ba, 2015). Adam means ‘adaptive movement estimation’; the main advantages

of this method are little memory requisites because it considers only first-order gradients, it labors also with sparse and noisy gradients, it is helpful for big data and big parameters problems, it changes the weights regardless of the diagonal rescaling of the gradients, there is not the needing for a stationary objective. The Adam convergence has an empirical demonstration. In recent years, it has become the most used optimizer in DNN. However, Adam regularization performances can be improved separating the gradient-base update from the weight decay, making weight decay and learning rate hyperparameters more independent and improving the parameters optimization, as proposed by Loshchilov et al. with Adamw (Loshchilov & Hutter, 2019). It is important to point out that the learning rate is an hyperparameter that controls how fast the model is to adapt itself to the proposed problem basing on the error estimate. The learning rate is one of the most important parameters for an optimizer. Instead, the weight decay is a penalty added to the loss function, based on model weights, and it is used to prevent overfitting. Both Adam and AdamW have been implemented in PyTorch. In our work, AdamW optimizer have been used.

2.4.2 DNN Architectures for Semantic Segmentation

DNN design is a very florid research field, therefore many novel architectures are produced every year, some of them are of general use, other are tailored for specific tasks or domain of application. However, consolidated DNN used for computed vision first, then for the semantic segmentation of images are mainly three: Connected Neural Networks (CNN), Fully Convolutional Neural Networks (FCN), and Encoder-Decoder model-based networks (Minaee et al., 2021). Recently, due to the excellent performances obtained for Natural Language Processing tasks, the attention is shifted also to Transformers combined with DNN (see paragraph 2.7).

The CNN have four fundamental traits: the use of many convolutional layers, common weights, pooling layers, and local connections (LeCun et al., 2015; Minaee et al., 2021). *Figure 2.5* shows the typical CNN structure. The main advantage of this type of networks is the lower number of parameters with respect to the FCN.

2.State of the art of image semantic segmentation

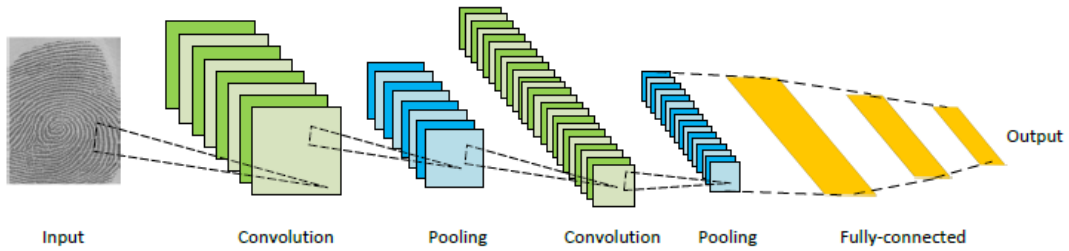


Figure 2.5: CNN structure. Sequence of operations: convolutions, subsampling (pooling), convolutions, subsampling, local connections (Minaee et al., 2021).

The FCN (Long et al., 2014; Minaee et al., 2021) is composed only by convolutional layers and the output is a spatial segmentation map instead of a vector of classification scores. It uses the skip-connections like the CNN, however, it is computationally expensive, even if the segmentations are accurate. Figure 2.6 shows the FCN structure.

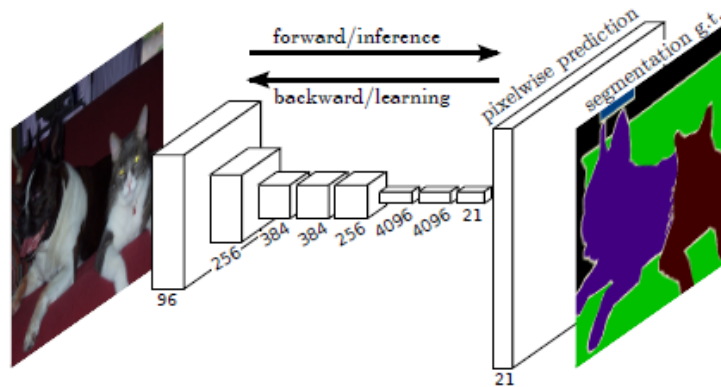


Figure 2.6: FCN structure. Operations: only convolutions, with skip connections to improve the segmentation (Long et al., 2014).

The Encoder-Decoder based models are based on a two-path structure: an encoding path that compresses the input into another space, while the decoding path starts from this new space and produces the output segmentation. The Autoencoders are architecture Encoder-Decoder based where the input is equal to the output. Figure 2.7 shows the Encoder-Decoder type structure.

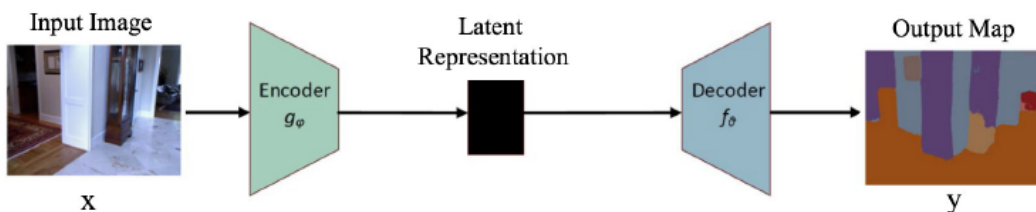


Figure 2.7: Encoder-Decoder structure. The input image is compressed in the encoder part to create a features map, represented in a higher dimensional space, then the decoder part scales back the feature map representation into the input image dimension to obtain the segmentation (Minaee et al., 2021).

In the next paragraphs, the DNN used in our work will be explained. It is important to point out that it is imprecise to place these networks in any of the above categories. The proposed membership class is only an indication and combination of the above categories in a single architecture can take place. In this thesis, all the DNN tested have been implemented in a patch-based fashion, considering as input the images as they have been acquired.

2.5 U-NET

A U-Net is an encoder-decoder based network architecture. It is a reimplement of the fully convolutional neural network (FCN) (Long et al., 2014) to allow the training with few images but with higher performances compared to the FCN. It was proposed by Ronneberger et al. (Ronneberger et al., 2015). As showed in the *Figure 2.8*, this network has a U-shape with a contracting part (encoder) followed by an expansive one (decoder).

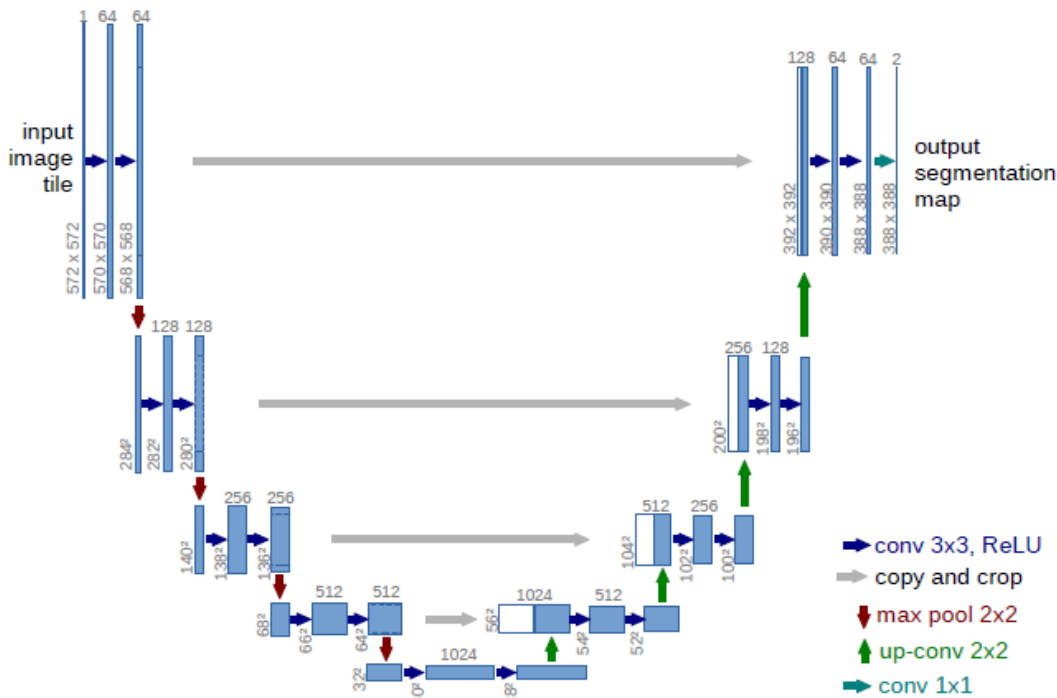


Figure 2.8: U-net architecture. Example for 32x32 pixels in the lowest resolution (Ronneberger et al., 2015).

With respect to the FCN, the U-Net has increased the number of feature channels of the up-sampling path, creating a symmetrical architecture. In this proposed architecture there are no fully connected layers and after the convolution operation only the target information is retained, and the context information is derived by the input image. This is useful to avoid interruptions using an overlap-tyle strategy. The voxels in the border regions are predicted looking at the mirror image.

There are 23 convolutional layers in total. The contracting path is a sequence of three operations: two 3x3 unpadded convolutions, a rectified linear-unit operation (ReLU), a 2x2 max-pooling operation to halve the resolution doing the down-sampling, doubling the number of feature channels at the same time.

In the second part of the architecture, the expansive path, the feature map is up-sampled; thereafter, they are applied in sequence: a 2x2 up-convolution, so the number of feature channels becomes the half, a concatenation with the same-level cropped feature map through shortcuts (horizontal lines), two 3x3 convolutions, a ReLU. The shortcuts are introduced to enter the contribution of non-granular information that would otherwise be lost. The last layer performs a 1x1 convolution to map the feature vector (64 features) into the classes.

Zhao et al. modify this standard model considering a parametric rectified linear unit (PReLU) activation function to have more parameters, so to improve performances (Zhao et al., 2020). Moreover, they extract more features at the first level of convolution (96 instead of 64). In addition, they consider a 3D U-Net, so 3x3x3 kernels for a convolutional layer and 2x2x2 max pooling layer. So, encoding and decoding layers have depth equal to three.

Zhao et al. find that performances are better with preprocessed images, data augmentation, non-uniform patch extraction, and weighted binary cross-entropy loss function.

In *Figure 2.9* we can see the comparison between a FreeSurfer segmentation, a 3D U-Net segmentation and a manual segmentation on a T1-weighted image.

The U-Net segmentation gives results very close to those obtained with the manual segmentation with respect to FreeSurfer.

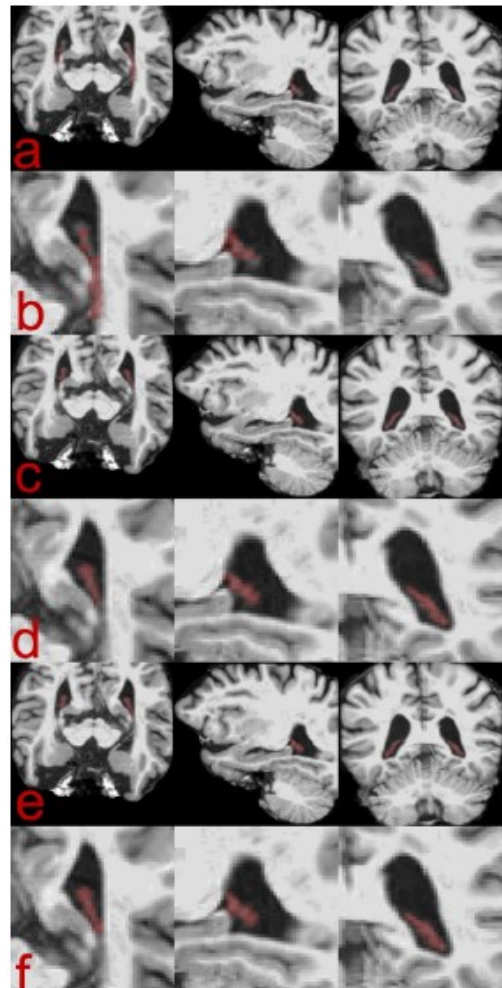


Figure 2.9: Comparisons of choroid plexus segmentation methods. (a)(b) the FreeSurfer results. (c)(d) the optimized 3D U-Net with preprocessed T1-weighted images. (e)(f) the manual segmentation (Zhao et al., 2020).

2.6 V-NET

Milletari et al. describe the V-Net architecture like a fully convolutional neural network that performs volumetric convolutions to segment MRI images (Milletari et al., 2016) (Figure 2.10). It has a V-shape.

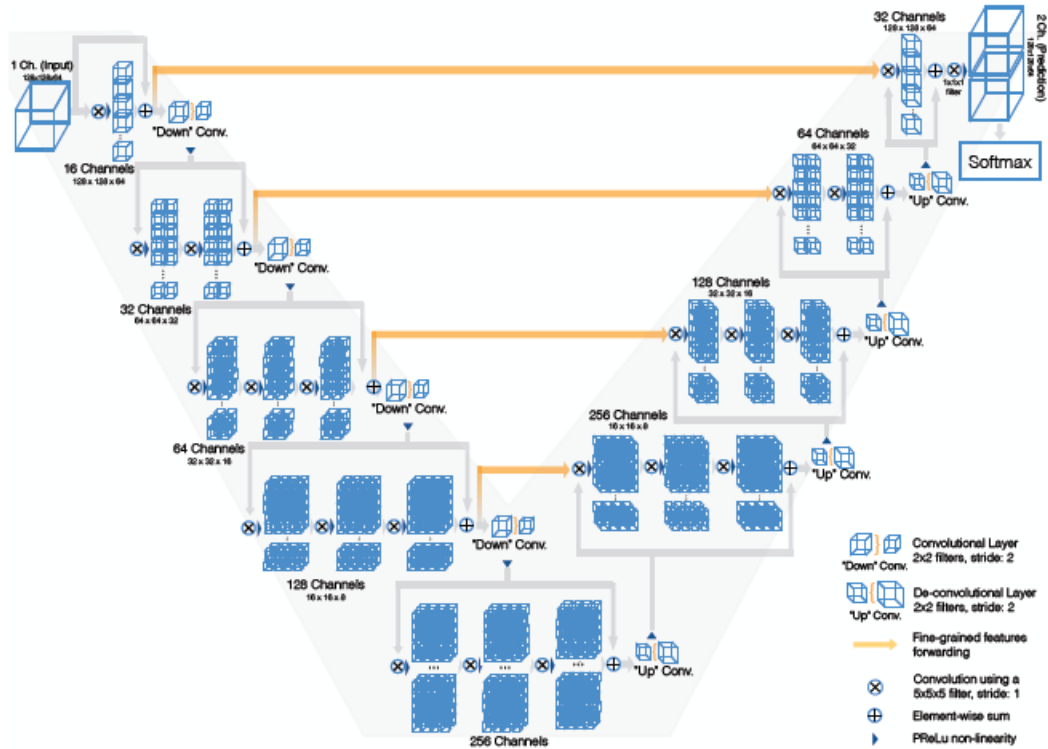


Figure 2.10: V-Net architecture (Milletari et al., 2016).

The V-Net architecture is characterized by a compressive part followed by an expansive one. Each part is composed by steps. In the left part, each step is composed from one to three convolutional layers, and it optimizes a residual function in a convergence time like that of other neural networks. This net uses the PReLU activation function. The convolutional layers use volumetric kernels with 5x5x5 voxel size, and they are followed by 2x2x2 voxel size wide kernels that halved the size of the stage feature map because of the use of 2x2x2 non overlapping patches. Differently from the U-Net architecture, in the V-Net the pooling layers are replaced by the convolutional ones because the number of feature channels is doubled at each stage so there is the needing of double the number of feature maps because their resolution is halved.

In this way, the right part of the architecture is not composed by up-pooling layers but only by de-convolutional ones. In fact, the expansive part of the V-Net extracts the features, starting from the two features maps extracted by the last left kernel, applying a soft-max layer to build probabilistic segmentations. Then, to increase the resolution, the

de-convolutional layers are applied and then the convolutional step starts (from 1 to three convolutional layers) but using the halved number of kernels (size 5x5x5) of the previous stage. This is repeated for the same number of stages used in the left part of the architecture.

Milletari et al. discover that the combined use of the V-Net and the Dice loss objective function brings to better experimental results when the segmentation task is a small anatomical region.

It could be interesting to see what happens if this network is used to segment the ChP region.

2.7 UNETR: TRANSFORMERS FOR IMAGE SEGMENTATION TASKS

The Transformers have been introduced in literature by Vaswani et al. (Vaswani et al., 2017). They became the new state-of-the-art for the Natural Language Processing (NLP) tasks and as encoder for the computed vision tasks due to their capability to extract the global information from the analysed context thanks to the attention mechanism. The attention mechanism uses as inputs and output only vectors. The output derives from the application of an attention function to the input values. This function maps the input through a set of keys and query, and the output is computed through a weighted sum of the mapped values, using a compatibility function that examines the degree of compatibility between the query and the key.

Starting from this idea, Hatamizadeh et al. have developed a new DNN model, called the UNet Transformers (UNETR) and they have obtained good results in applying this architecture to medical images. The main key to adapt this framework to images is that an image can be viewed as a text to be analysed but, instead of recognizing the words, the task is to find the voxels that contains the ChP.

2.7.1 UNETR structure

The UNet Transformers (UNETR) is a DNN that uses a transformer as an encoder and a U-Net in the decoder part. UNETR uses the patch-based approach. *Figure 2.11* shows the structure of a UNETR.

2.State of the art of image semantic segmentation

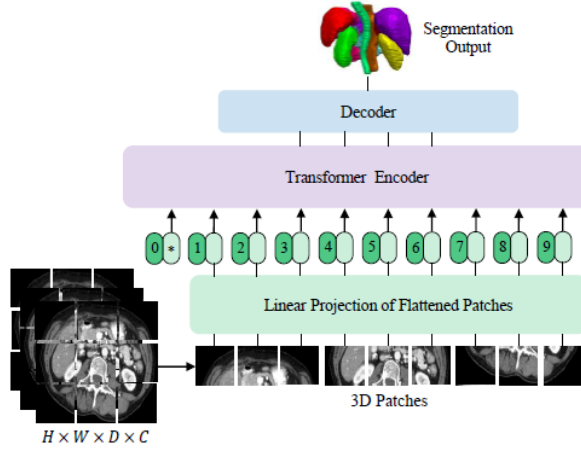


Figure 2.11: Overview of UNETR architecture (Hatamizadeh et al., 2021).

Hatamizadeh et al. describe the UNETR architecture for 3D segmentation images as follows (Hatamizadeh et al., 2021) (Figure 2.12).

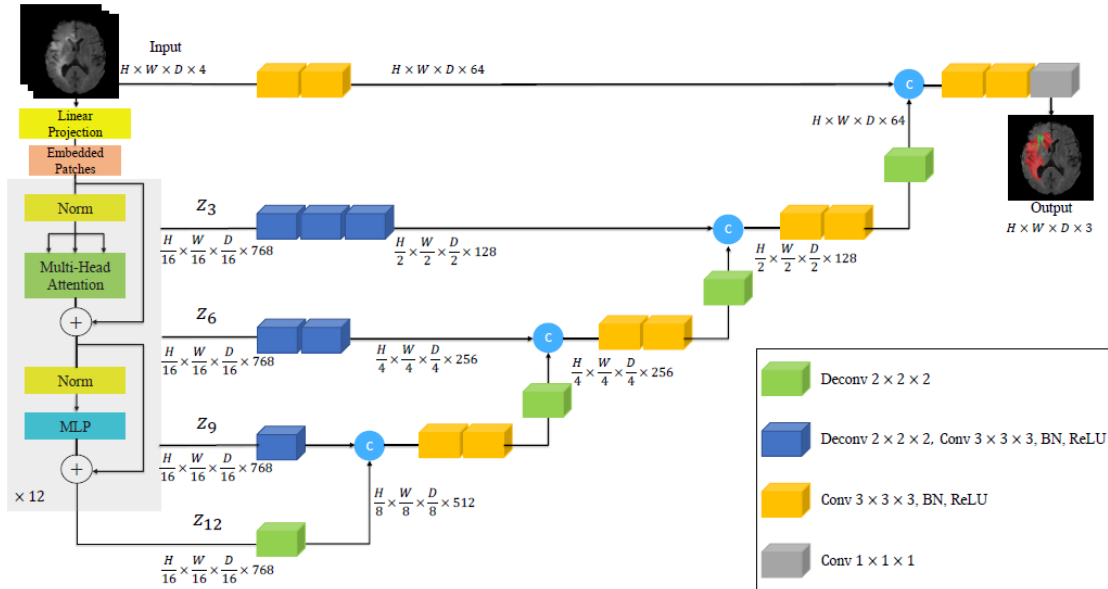


Figure 2.12: Overview of UNETR architecture. Patch resolution $P=16$, embedding size $K=768$ (Hatamizadeh et al., 2021).

The use of a transformer followed by a U-Net combines the ability of the transformer in learning the global information, and the feature of the U-Net in learning the localized one. Skip-connections between the transformer-encoder and the U-Net-decoder are always present to predict the segmentation using the contextual information.

Looking at the encoding part, the transformer needs a 1D sequence of inputs, so for a 3D image the 1D input sequence is $x \in \mathbb{R}^{HxWxDxC}$, where (H, W, D) is the resolution and C is the number of channels through which the input volume is divided into uniform non-overlapping patches $x_v \in \mathbb{R}^{N \times (P^3 \cdot C)}$ with resolution (P, P, P). N is the length of the overall sequence: $N = \frac{HxWxD}{P^3}$. After the input preparation, the first step is the linear layer that

brings the patches into a K-dimensional space. The spatial information about the patches is preserved using a 1D positional embedding $\mathbf{E}_{pos} \in \mathbb{R}^{N \times K}$ in addition to the projected patch one $\mathbf{E} \in \mathbb{R}^{(P^3 \cdot C) \times K}$, following the formulation: $z_0 = [x_v^1 \mathbf{E}; x_v^2 \mathbf{E}; \dots; x_v^N \mathbf{E}] + \mathbf{E}_{pos}$.

After that, there are some transformers blocks: multi-head self-attention (MSA) and multilayer perceptron (MLP) layers. The formulations are:

$$z'_i = MSA(Norm(z_{i-1})) + z_{i-1}, \quad i = 1, \dots, L$$

$$z_i = MLP(Norm(z'_i)) + z'_i, \quad i = 1, \dots, L$$

i : number of intermediate blocks

L : number of transformer layers

A MLP is composed by two linear layers and a GELU (Gaussian Error Linear Unit) activation functions (Hendrycks & Gimpel, 2016).

Instead, MSA has n parallel self-attention heads (SA), where the SA block is a function that learns the mapping between a query q , a key k and a value v in the input sequence $z \in \mathbb{R}^{N \times K}$, as explained before. The parameters of SA are called attention weights (A) and they are derived from: $A = Softmax(\frac{qk^T}{\sqrt{K_h}})$, where $K_h=K/n$ is the scaling factor

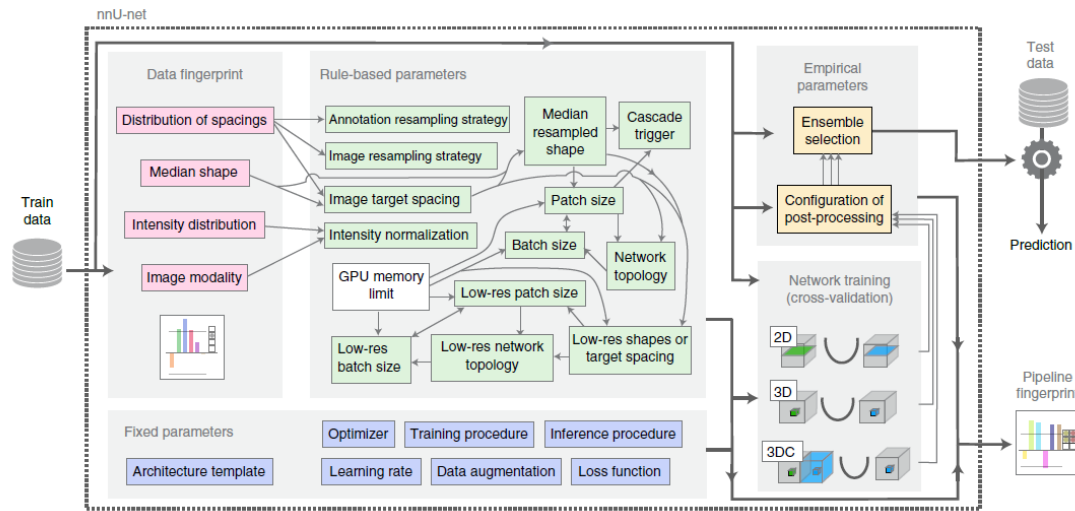
The yield of SA is: $SA(z) = Av$

The MSA elaboration is $MSA(z) = [SA_1(z); SA_2(z); \dots; SA_n(z)]W_{msa}$, where W_{msa} are the multi-headed trainable parameters weights.

After these blocks, z_i is obtained (size $(H \times W \times D/P^3) \times K$) and it is reshaped into a tensor to become an output of the transformers in the embedding space. Then, there is the passage from the embedding space into the original space using $3 \times 3 \times 3$ convolutional layers and normalization layers. At the end of the encoder, the architecture proceeds bottom up applying a deconvolutional layer to the output of the last layer of the transformers to double the resolution, then the concatenation of each resized feature map with that not resized of the previous layer (from z_{12} to the original input) starts, applying $3 \times 3 \times 3$ convolutional layers and then a deconvolutional one. When the concatenation reaches the original resolution, after the $3 \times 3 \times 3$ convolutional layers, a $1 \times 1 \times 1$ convolutional layer with Softmax activation function is applied to generate the predicted segmentation.

2.8 nnU-NET (DYN U-NET): A SELF-CONFIGURING DNN

The Dyn U-Net is a MONAI implementation of the nnU-net (no-new-Net) proposed by Isensee et al. (Isensee et al., 2021). This neural network has the peculiarity to build itself automatically in terms of pre-processing, architecture, training, post-processing, basing on the input-net dataset. In other terms, it is a holistic network because also the topology of the architecture does not have to be determined a priori. The need to build a self-configuring DNN arises since the DNN performances are strictly dependent on the parameters chosen by the operator. *Figure 2.13* presents the implementation of the nnU-Net for a segmentation task proposed by Isensee et al. (Isensee et al., 2021).



Design choice	Required input	Automated (fixed, rule-based or empirical) configuration derived by distilling expert knowledge (more details in online methods)			
Learning rate	-	Poly learning rate schedule (initial, 0.01)	Image target spacing	Distribution of spacings	If anisotropic, lowest resolution axis tenth percentile, other axes median. Otherwise, median spacing for each axis. (computed based on spacings found in training cases)
Loss function	-	Dice and cross-entropy	Network topology, patch size, batch size	Median resampled shape, target spacing, GPU memory limit	Initialize the patch size to median image shape and iteratively reduce it while adapting the network topology accordingly until the network can be trained with a batch size of at least 2 given GPU memory constraints. for details see online methods.
Architecture template	-	Encoder-decoder with skip-connection ('U-Net-like') and instance normalization, leaky Rel.U, deep supervision (topology-adapted in inferred parameters)	Trigger of 3D U-Net cascade	Median resampled image size, patch size	Yes, if patch size of the 3D full resolution U-Net covers less than 12.5% of the median resampled image shape
Optimizer	-	SGD with Nesterov momentum ($\mu = 0.99$)	Configuration of low-resolution 3D U-Net	Low-res target spacing or image shapes, GPU memory limit	Iteratively increase target spacing while reconfiguring patch size, network topology and batch size (as described above) until the configured patch size covers 25% of the median image shape. For details, see online methods.
Data augmentation	-	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring	Configuration of post-processing	Full set of training data and annotations	Treating all foreground classes as one; does all-but-largest-component-suppression increase cross-validation performance? Yes, apply; reiterate for individual classes No, do not apply; reiterate for individual foreground classes
Training procedure	-	1,000 epochs \times 250 minibatches, foreground oversampling	Ensemble selection	Full set of training data and annotations	From 2D U-Net, 3D U-Net or 3D cascade, choose the best model (or combination of two) according to cross-validation performance
Inference procedure	-	Sliding window with half-patch size overlap, Gaussian patch center weighting			
Intensity normalization	Modality, intensity distribution	If CT, global dataset percentile clipping & z score with global foreground mean and s.d. Otherwise, z score with per image mean and s.d.			
Image resampling strategy	Distribution of spacings	If anisotropic, in-plane with third-order spline, out-of-plane with nearest neighbor. Otherwise, third-order spline			
Annotation resampling strategy	Distribution of spacings	Convert to one-hot encoding \rightarrow If anisotropic, in-plane with linear interpolation, out-of-plane with nearest neighbor. Otherwise, linear interpolation			

Figure 2.13: Proposed automated method configuration for deep learning-based biomedical image segmentation (Isensee et al., 2021).

As shown in the *Figure 2.13*, the nnU-Net involves the use of three different types of parameters: the fixed parameters (data independent), the rule-based parameters, and the empirical parameters. The rule-based parameters are a compromise between the dataset properties (dataset fingerprint, like image size) and the design choices (pipeline fingerprint, like patch size or batch size) to enable an instant adaptation to the contest. The empirical parameters, that concern the model selection and the post processing, are set last according to the training phase.

The loss function proposed by the authors is a combination of both the Dice loss and the Cross-Entropy loss.

As regards the configuration of the model, the choice must fall on one of the following neural networks (or on a combination of two): a 2D U-Net, a 3D U-Net (full image resolution), a 3D U-Net cascade. The 3D U-Net cascade is a combination of a 3D U-net applied on down-sampled images followed by a 3D U-Net applied on the full images to refine the segmentation maps of the previous one. The model with the best cross-validation performances is then selected.

The nnU-Net is fast, there are few decisions to make a priori using it, and it is data efficient. Finally, this neural network is not focused only on the net architecture, but it finds the best architecture looking at the dataset, so it has great potential for applicability in various fields.

3 MATERIALS AND METHODS

This section is dedicated to the Materials and Methods of this research. In addition to the description of the procedural choices, the comparison with the literature’s works is introduced but it will be expanded in the chapter of the Discussion. This chapter describes the operational choices that were made, starting from the initial choice of the MRI sequences to be used.

3.1 DATASET: MRI SCANS AND GENERAL DESCRIPTION

Sixty Relapsing Remitting MS patient (Age 39.9 ± 9.5 years). The dataset was provided by the Multiple Sclerosis Center of the University Hospital of Verona, in collaboration with which this research was carried out.

Scanner: Philips Elition 3T, equipped with dedicated 32 channels head coil. Software version R5.7.2.1.

Protocol and MRI sequence parameters (*Table 3.1*):

MRI Sequence name	T1w (same for T1w and cT1w)	FLuid Attenuated Inversion Recovery (FLAIR)	Double Inversion Recovery (DIR)
MRI Sequence type	Gradient Echo Flash	Turbo Spin Echo with variable flip angle (BrainView)	Turbo Spin Echo with variable flip angle (BrainView)
Resolution	1x1x1 mm	1x1x1 mm	1x1x1 mm
Compressed SENSE acceleration factor	4	5	7
Echo Time	3.8 ms	376 ms	323 ms
Repetition Time	8.5 ms	8000 ms	5500 ms
Inversion Time	N.A.	2356 ms	525/2550 ms
Flip Angle	8	90	90
Acquisition time	3min 20s	4min 20s	7min

Table 3.1: Protocol and MRI sequence parameters.

The sequences were controlled by visual inspection. Furthermore, the correspondence of the header files between manual segmentations and the respective sequence was verified, as well as the size of the images themselves. The choice to keep the T1-w, cT1-w and FLAIR sequences co-registered with linear interpolation with respect to T1-w was made during the preliminary analysis, reported in Appendix A and in paragraph 3.2. The preliminary analysis was done first over nine subjects, ten over the whole dataset (sixty-one patients).

From the initial sixty-one patients, one was excluded by the training and validation sets because of the presence of ChP cysts which could invalidate the final prediction.

The dataset was divided into two parts: 45 subjects for the training set, on which the DNNs were trained, and 15 subjects for the validation set, on which the trained DNNs were tested.

3.2 PRELIMINARY ANALYSIS FOR THE DATASET COMPOSITION

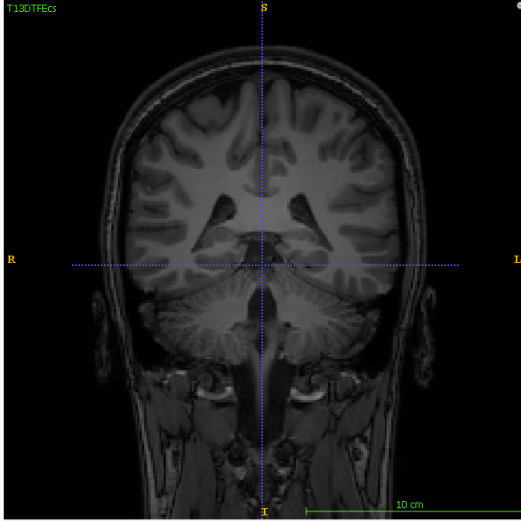
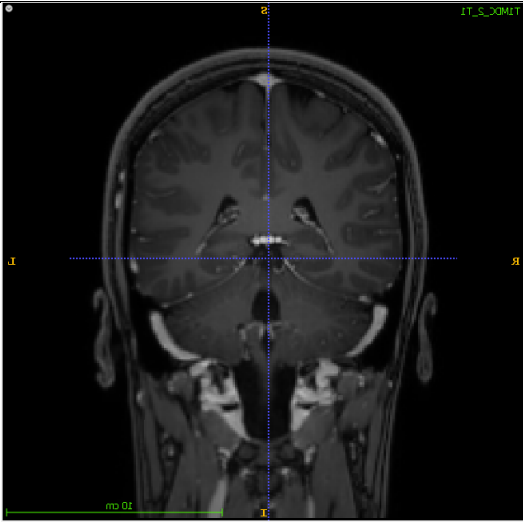
The first analysis that was carried out concerned the composition of the dataset in terms of image sequences to be considered. As a matter of fact, the Choroid Plexus segmentation studies in the literature (Schmidt-Mengin et al., 2021; Zhao et al., 2020) use T1-weighted (T1-w) segmentations (manual or FreeSurfer based) as ground truth and evaluate, with respect to these, the performances of the proposed method on T1-w images. Only one study (Tadayon, Moret, et al., 2020) uses as ground truth the manual segmentation depicted on cT1-w images in a subsample of subjects.

Our goal is to investigate whether other sequences, for example FLAIR or DIR, given as input to Deep Learning Neural Networks (DNNs), can give good results, even in combination with T1-w, and comparable to those obtainable with the gold-standard, the manual segmentation on cT1-w images.

However, no study has systematically investigated differences in ChP segmentation with images obtained with other sequences than T1-w ones. Therefore, a first explanatory analysis is performed on a sub-dataset consisting of the first 9 patients for each of which the following scans are performed: 3DT1TFEMDC, 3DT1TFE, 3DFLAIR, 3DDIR (Table 3.2). Consequently, the manual segmentations available are obtained from depicting ChP on DIR, T1-w and FLAIR sequences. Moreover, the T1-w sequence was used to perform the ChP segmentation with the automatic algorithm included in FS and the GMM method. The segmentations provided are then compared to the gold standard segmentation, so the manual segmentation based on the cT1-w sequence.

3. Materials and Methods

This analysis is carried out in MATLAB (version 2019b). Moreover, in the Results chapter this analysis was carried out over all sixty-one subjects, always on MATLAB. The nine patients were: 000091, 001889, 001922, 002043, 002045, 002050, 002056, 002059, 002060.

SEQUENCES	IMAGES OF PATIENT 2060 – ITK-SNAP Viewer
<p>T1- weighted image without contrast</p> <p>3DT1TFE is a sequence in which the image is T1 weighted (grey matter is gray, white matter white, CSF black). It is a TFE (Turbo Field Echo) sequence so a gradient echo pulse sequence. The Choroid Plexus can be detected, however the low contrast between the ChP and the neighboring structures requires a trained operator.</p>	
<p>T1- weighted image with contrast</p> <p>3DT1TFEMDC is the same of 3DT1TFE, acquired after the injection of contrast agent. The Choroid Plexus is enhanced since the contrast is accumulated in major vessels. On one hand, it's easier to segment it manually; on the other hand, as said in the introduction, the contrast is performed only on specific clinical demands.</p>	

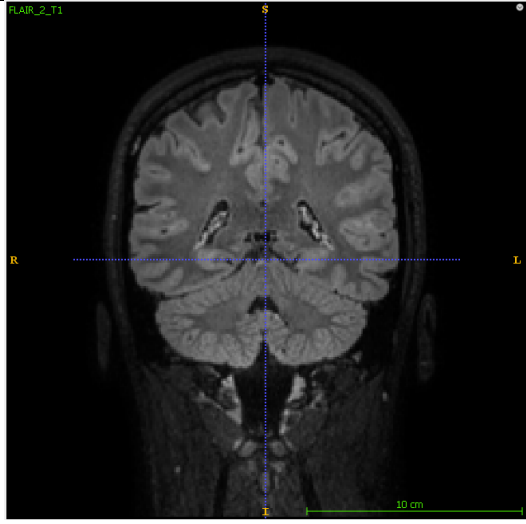
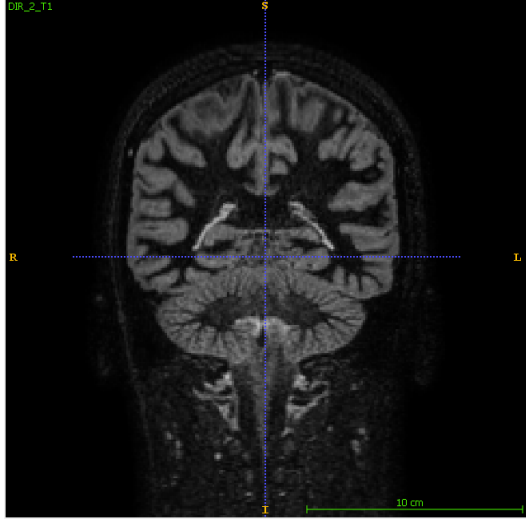
<p>FLAIR</p> <p>3D FLAIR is a sequence called Fluid Attenuated Inversion Recovery, the inversion time (TI) is longer to remove the fluid effect from the image (TI is set to erase fluid contribute). It is a very useful sequence to detect brain lesions that are normally covered by fluid signals. FLAIR give good contrast, almost comparable to the gold standard, however, there is a blurring phenomenon caused by TSE readout that does not allow a good definition of the edges.</p>	
<p>DIR</p> <p>3D DIR it is a Double Inversion Recovery pulse sequence; it is used to suppress signals derived from the CSF and the white matter thanks to two inversion pulses. It is useful to detect multiple sclerosis plaques. The contrast-to-noise ratio of this sequence is affected by the two inverse pulses that degrades the image quality.</p>	

Table 3.2: Description of the MRI sequences.

The first step of the preprocessing consists in the co-registration of all the sequences to the T1-w space. This is done because of the acquisition protocol. In fact, the first image to be acquired is the T1-w, the last is the cT1-w.

The metrics chosen to make the comparison between the segmentations of the ChP are the same that will be used in the training and validation procedure: Dice Coefficient, 95% Hausdorff Distance, Volume difference, RMSE (Root Mean Squared Error) and MSE (Mean Squared Error), Pearson's Correlation Coefficient (and relative p-values), linear regression between volumes of the investigated sequences (T1-w, FLAIR, DIR) with respect to the cT1-w one (gold standard) and OLS (linear regression without intercept), and Percentage Volume Difference.

In the following paragraph it will be discussed the analysis over the sequence's manual segmentations.

3.2.1 Manual segmentation

Manual segmentation of the available sequences was carried out by a trained radiologist that was blinded to subject identity. Subjects were shuffled and the radiologist performed separately the segmentation on one modality per time. Firstly, the FLAIR was segmented, secondly the T1-w, thirdly the DIR and lastly the cT1-w to minimize the bias in influencing the reader providing information on the gold-standard sequence.

3.2.2 Preliminary Analysis

A preliminary analysis was carried out on a subset of the whole dataset, with the aim of understanding the usefulness of each sequence tested. A comparison in term of ChP segmentation was made between each sequence available, considering as reference the segmentation obtained with the gold standard sequence (cT1-w).

Based on this analysis, a preliminary selection of the available sequences was made, to limit the time-consuming work done by the radiologist.

To provide a further insight on auto automated segmentation methods, the segmentation provided by both GMM and FS on the selected subset of subjects was included in the preliminary analysis.

Afterwards, a preliminary analysis over the whole dataset was carried out to verify if the preliminary results were consistent. It was taken the decision to discard the DIR sequence, as explained in the Results and Discussion chapters.

3.3 MONAI

This work has been implemented in MONAI version 0.8.1 (<https://monai.io>), using PyTorch version 1.10 (<https://pytorch.org>) and Python version 3.9.9, with NVIDIA-SMI 510.39.01, Driver Version: 510.39.01, CUDA Version: 11.6.

MONAI means Medical Open Network for Artificial Intelligence, it's a free PyTorch-based framework designed for deep learning in healthcare imaging (*Figure 3.1*).

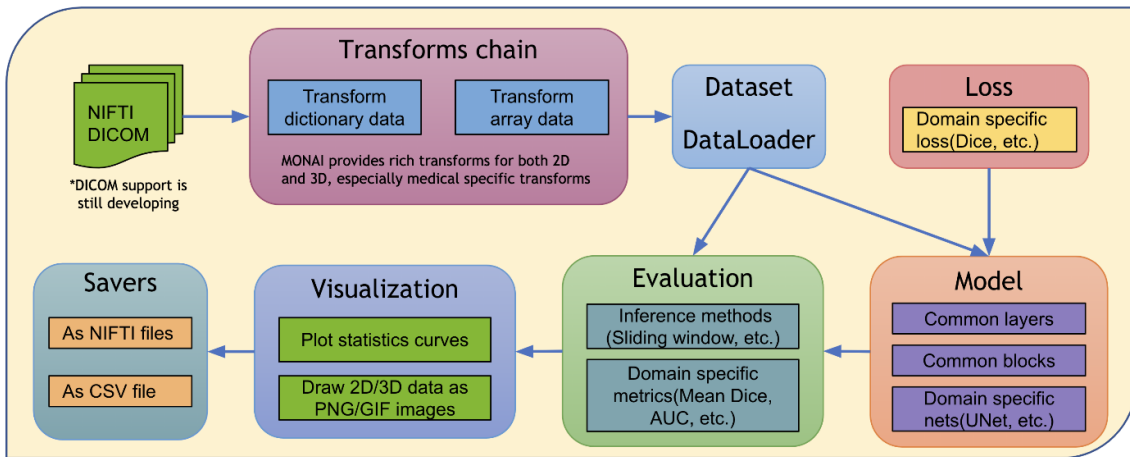


Figure 3.1: Illustration of a deep learning algorithm in MONAI (<https://docs.monai.io/en/stable/highlights.html>).

The main goal of MONAI Project is to create conditions that allow researchers to collaborate in building new Artificial Intelligence (AI) models using a standardized protocol. MONAI has a streaming loading modality to improve the training efficiency, many DNN already implemented, as domain specific metrics (i.e., Dice, Hausdorff Distance) and loss function (i.e., Dice, DiceCE). Moreover, there are specific transforms to preprocess the dataset with data augmentation to improve the nets performances. In addition, there is the possibility to plot the results obtained during the training phase and the possibility to save the results (i.e., predicted segmentations) with specific file extensions (i.e., NiFTi).

3.4 PREPROCESSING OF THE IMAGES

It is essential to remember that the 3D image over which the segmentation will be performed is the input for the model (or two images for the double 4D input), while at the output there will be the additional predicted segmentation. Before starting the training step, it is necessary to preprocess the images and, in some cases, also the labels, differently for training dataset and validation dataset. The applied transforms are the same for the one channel option (T1-w, FLAIR) or for the two-channel option (merge of T1-w and FLAIR), the only difference is the type of function to be applied.

Through the spatial operation the inputs are resampled into a voxel dimension of $1 \times 1 \times 1 \text{ mm}^3$ and oriented based on RAS specifications (Right, Anterior, Superior). The intensity operations allow to scale the intensity (only of the image) to the range $[0, 1]$. The crop transformations are used to make the training and validation step easier because they select the foreground with respect to the background. All these operations are done for both training and validation sets. However, the most important crop transform is the random cropping of patches in the image for the training set. As a matter of fact, the DNN

accepts as input the patches of the image over which the segmentation task is performed. The main issue of the ChP segmentation task is the label imbalance, since the ChP is tiny when compared to the whole brain. Due to this, two main problems emerge: the inter-patch imbalance and the intra-patch imbalance. Zhao et al. (Zhao et al., 2020) analyzed that large-patch size lower the intra-patch imbalance but the inter-patch imbalance became higher, and vice versa. For this reason, there is the need to find a balance between inter- and intra-patch imbalance. What is used in this work is to extract the patch with a 50% probability of having the central patch voxel containing the foreground or background. This is of help for the inter-patch imbalance. For the intra-patch imbalance, the solution is to use a loss function properly weighted.

3.4.1 Data Augmentation

Our dataset is small in term of subjects, so there is the need to apply some data augmentation strategies to increase the variability between the training data. The goal of these additional operations is to make the model more robust and more flexible to be applied to the validation set. This relation between the application of data augmentation transforms and the performance improvement was observed and tested by Dosovitskiy et al. (Dosovitskiy et al., n.d.). Indeed, elastic deformations are very common in human tissues and structures, in particular in presence of pathological conditions.

Like Schmidt-Mengin et al. and Hatamizadeh et al. (Hatamizadeh et al., 2021; Schmidt-Mengin et al., 2021), the data augmentation operations added are flips, rotations, intensity shifts, and small rotations and they are showed in the *Table 3.3* below accompanied with the respective application probability.

Operation	Direction/Angle/Offset	Probability
Flip	Y/N	10%
Rotation	90°/180°/270°	10%
Intensity Shift	0.1 offset	50%
Small Rotation	5° on x,y or z	10%

Table 3.3; Data Augmentation operations applied to the input training dataset with the indications of the respective probability and specifications (direction/angle/offset).

3.5 TRAINING PARAMETERS

During the training step, some parameters have been kept fixed, other have been varied to find the DNN model with the best combination of parameters to maximize the performance of the segmentation task. With contrast to literature studies (Hatamizadeh et al., 2021; Isensee et al., 2021; Milletari et al., 2016; Schmidt-Mengin et al., 2021; Zhao

et al., 2020), the training procedure consists of 40000 iterations and 400 epochs, over a training dataset of 45 subjects.

The optimizer used is Adam (Kingma & Ba, 2015) in its weighted version, like that used by Hatmizadeh et al. (Hatamizadeh et al., 2021), as described in the Chapter 2. The fixed parameters are the learning rate, the weight decay, and the batch size:

- The learning rate is an Adam hyperparameter that controls how much the model can be changed looking at the estimated error when the weights are updated. The learning rate is fixed at $1 \text{ e-}04$, with one exception for the nnU-Net (DynUnet) that has two different learning rates: $1 \text{ e-}02$ for iterations $[0, 10000]$ and $1 \text{ e-}03$ for iterations $[10000, 40000]$, taking inspiration from what was suggested for other tasks by the authors of the paper (Isensee et al., 2021).
- The weight-decay is a small penalty added to the loss function to stem overfitting and an enormous weight growth. The weight-decay is fixed at $1 \text{ e-}05$.
- The batch size is the number of examples done over the training set to estimate the error gradient. This hyperparameter controls how much the estimate of the error gradient is accurate. The batch size is fixed at 1.

As regards the variable parameters, the changes have involved: the type of input, the patch size, the loss function, and the preprocessing of the images with or without data augmentation, and, above all, the type of DNN architecture. The *Figures 3.2* below shows all the options for each of these categories. The total number of possible combinations used for the training step is 672.

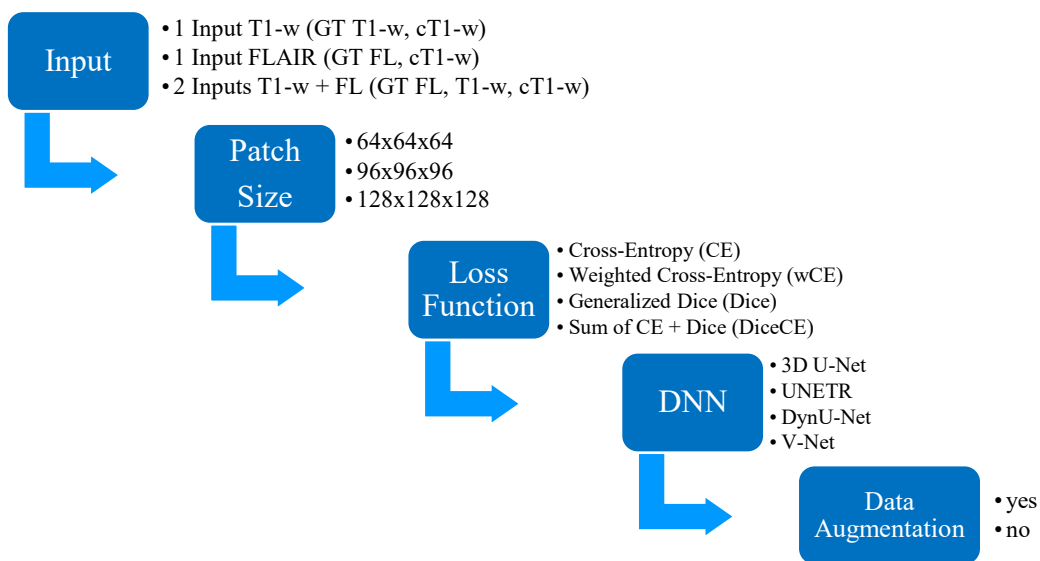


Figure 3.2: available options to train model varying inputs, patch size, loss function, DNN with or without data augmentation.

3.6 LOSS FUNCTION

The choice of the loss function is crucial to train a DNN. In fact, the training step of a deep learning algorithm performs a risk minimization, that means minimize a loss function that is used to evaluate the performances over the test set (Bertels et al., 2019). For this reason, the loss choice has an impact over the final predicted segmentations. Currently, there are two main lines of thought on the choice of the loss function: one uses the Cross-Entropy Loss (and its Weighted version) and use the Dice Coefficient (and other metrics) to evaluate the performances; the other uses loss functions that include the evaluation of the Dice Coefficient on their computation to improve the performances.

The empirical risk minimization of the learning phase aims to learn a segmentation algorithm f , the prediction, starting from an observed input x , the image, which is comparable to the reference segmentation y , the label, optimizing the expectation of a loss function over the training set:

$$\arg \min_{f \in F} \underbrace{\frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)}_{=: \hat{R}(f)}$$

l : loss function

F : set of functions represented by a CNN

$$\hat{R}(f) = E_{(x,y) \sim P_n} [l(x, y)]$$

E : expectation operation

P_n : bootstrap distribution from sample S of size n , where $S = \{(x_i, y_i)\}_{1 \leq i \leq n}$

For the neural network backpropagation training, it is calculated the gradient $\frac{\partial Loss}{\partial x}$, where x are the network parameters.

The tested DNNs were trained with four different Loss Functions, most of them found in literature's Choroid Plexus segmentation: the Cross-Entropy Loss function, the Weighted Cross-Entropy Loss function, the Generalized Dice Loss function, and the Dice-CE Loss function that is the union of Cross-Entropy Loss and Dice Loss.

3.6.1 Cross-Entropy Loss and Weighted Cross-Entropy Loss

The Weighted Cross-Entropy Loss is used by Ronneberger et al. and Zhao et al. (Ronneberger et al., 2015; Zhao et al., 2020) for optimizing the 3D U-Net. Zhao et al. study compares both Cross-Entropy Loss and Weighted Cross-Entropy Loss (Zhao et al.,

2020) to examine if the intra-patch label imbalance could be compensated assigning weights to the foreground region (ChP) and to the background.

The Weighted Cross Entropy in the two-class form can be summarized as (Sudre et al., 2017):

$$WCE = -\frac{1}{N} \sum_{n=1}^N w r_n \log(p_n) + (1 - r_n) \log(1 - p_n)$$

w: weights for the foreground class, $w = \frac{N - \sum_n p_n}{\sum_n p_n}$

r_n : voxel values of the R reference segmentation

p_n : predicted probabilistic map elements for the foreground label P (background class probability is 1-P)

The Cross-Entropy Loss was implemented extending from *torch.nn.CrossEntropyLoss*. Here the Loss function is implemented considering taking as input a tensor that contain scores for each C class (in this case, 2 class) and dimension. The weights could be added assigning a weight to each class. As in Zhao et al. it was set to 0.1 the weight for the 0 class (background) and weight 0.9 for the 1 class (foreground ChP) (Zhao et al., 2020). Therefore, the formulation of the loss function becomes (<https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html?highlight=crossentropyloss#torch.nn.CrossEntropyLoss>):

$$l(x, y) = \sum_{n=1}^N \frac{1}{\sum_{n=1}^N w_{y_n} \cdot 1\{y_n \neq ignore_index\}} l_n,$$

$$l_n = -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^C \exp(x_{n,c})} \cdot 1\{y_n \neq ignore_index\}$$

x: input (the model)

y: target (the manual segmentation)

w: weights

C: number of classes

N: number of voxels

ignore_index: report target values which are not to be included in the gradient input computation

3.6.2 Generalized Dice Loss

The Generalized Dice Loss proposed by Crum et al. (Crum et al., 2006) and by Sudre et al. (Sudre et al., 2017) for the training step of DNN was also tested. In fact, Milletari et al. (Milletari et al., 2016) proposed the use of the Dice Coefficient as the base to build an objective loss function to train a model in the learning step. This choice finds its bases on the fact that some anatomical regions of interest in medicine, like tumor or in this case the Choroid Plexus, occupy a very small region (of the brain) compared to the background, so the learning process could find with higher probability local minima and train a network with biased predictions. The consequence is that some parts of the region to be segmented are missing or finding only partially. However, if the loss function considers the optimization of the Dice Coefficient, in other words the degree of spatial overlap of the two examined segmentations, this problem could be resolved.

Starting from Milletari et al. formulation (Milletari et al., 2016), Sudre et al. describe the Generalized Dice Loss as (Sudre et al., 2017):

$$GDL = 1 - 2 \frac{\sum_{l=1}^2 w_l \sum_n r_{l_n} p_{l_n}}{\sum_{l=1}^2 w_l \sum_n r_{l_n} + p_{l_n}}$$

w_l : weight used to make the loss function invariant with respect to label set properties

r_{l_n} : voxel values of the R reference segmentation for the label set

p_{l_n} : predicted probabilistic map elements for the labels

When $w_l = 1 / (\sum_{n=1}^N r_{l_n})^2$ the loss function becomes GDL_v : in this case the correction for the inverse volume of each label is performed to reduce the correlation between Dice Coefficient and region dimension.

Considering the stochastic gradient descent in the two-class configuration, the gradient formulation is:

$$\frac{\partial GDL}{\partial p_i} = -2 \frac{(w_1^2 - w_2^2) [\sum_{n=1}^N p_n r_n + r_i \sum_{n=1}^N (p_n + r_n)] + N w_2 (w_1 + w_2) (1 - 2r_i)}{[(w_1 - w_2) \sum_{n=1}^N (p_n + r_n) + 2N w_2]^2}$$

This formulation allows the use for both balanced and unbalanced data and it uses a single score to evaluate the multiple class segmentation performances.

This loss function is implemented in the *class monai.losses.GeneralizedDiceLoss*. The background class was excluded from the loss calculation (*include_background = False*) and a sigmoid function was applied to provide a binary prediction (0: background voxels, 1: ChP voxels).

3.6.3 DiceCE Loss

The DiceCELoss is a weighted sum of both Cross-Entropy Loss and Dice Loss. It is used as loss function in some recent studies like that of Hatamizadeh et al. for the training step of the UNETR (Hatamizadeh et al., 2021). It is implemented in MONAI as *class monai.losses.DiceCELoss*. As in the case of Generalized Dice Loss, the background class was excluded from the loss calculation, and a sigmoid function was applied.

The Dice Loss is the version proposed by Milletari et al. (Milletari et al., 2016) and implemented in MONAI as *class monai.losses.DiceLoss*. The starting point is the calculation of the Dice Coefficient like:

$$Dice(P, G) = \frac{2 * \sum_{i=1}^N p_i g_i}{\sum_{i=1}^N p_i^2 + \sum_{i=1}^N g_i^2}$$

N : number of voxels

p_i : predicted binary segmentation volume (P)

g_i : ground truth binary segmentation volume (G)

The gradient formulation becomes:

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

p_j : j-th voxel prediction

The advantage is the absence of the need to assign weights to the various classes and the performances of the net are better than that trained with, for example, multinomial logistic loss.

Using this formulation, Hatamizadeh et al. build the DiceCELoss function as a combination of both *torch.nn.CrossEntropyLoss* and *monai.loss.DiceLoss*, that could be or not a weighted sum (Hatamizadeh et al., 2021). The loss formulation is:

$$L(G, P) = 1 - \frac{2}{C} \sum_{j=1}^C \frac{\sum_{i=1}^N G_{i,j} P_{i,j}}{\sum_{i=1}^N G_{i,j}^2 + \sum_{i=1}^N P_{i,j}^2} - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C G_{i,j} \log(P_{i,j})$$

N : number of voxels

C : number of classes

$P_{i,j}$: probability output for class j at i-th voxel

$G_{i,j}$: one hot encoded ground-truth for class j at i-th voxel

This formulation allows not to attribute weights to the classes and at the same time it allows to consider the trend of the Dice Coefficient to improve the performances of the DNN. It could prove to be the most promising loss function among those considered.

3.7 VALIDATION PROCEDURE: PERFORMANCE EVALUATION

To evaluate the performance of each DNN's models in terms of accuracy of the output segmentation, and then to make the comparison with FS and GMM, quantitative metric values were calculated during the training step over the validation set. The results reported in the Results chapter are referred to the model evaluation over the validation set to evaluate the DNN models' performances, as the comparison with FS and GMM.

3.7.1 Dice Coefficient

The Sørensen-Dice Coefficient or Dice Coefficient (DC) or Dice Similarity Coefficient is used as a measure of similarity between two binary segmentations (Dice, 1945; Hatamizadeh et al., 2021; Tadayon, Moret, et al., 2020; Zou et al., 2004) (Figure 3.3). In presence of high-class imbalance, that is our case, Dice Coefficient is a special case of kappa index, so chance-corrected, and the DC, reflecting the location and the size of the segmentation, appears to be more in line with the perceptual quality rather than with the precision of the voxel (Bertels et al., 2019).

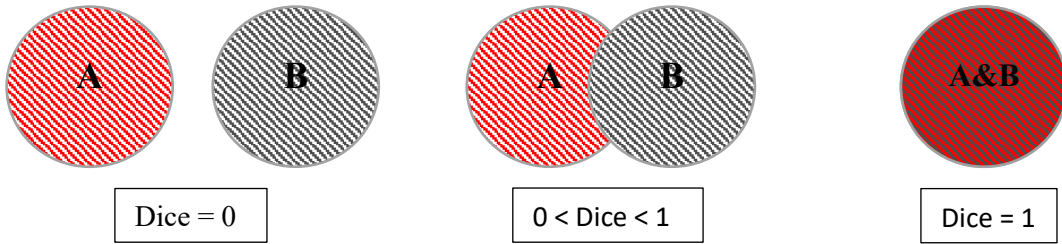


Figure 3.3: Visual demonstration of the Dice Score values.

DC is an indicator of spatial overlap of two segmentations under consideration. It ranges between 0 and 1, unitless: 0 indicates no overlap, 1 indicates complete overlap. There are two principal formulations:

$$Dice(G, P) = \frac{2 * |G \cap P|}{|G| + |P|} = \frac{2 * \sum_{i=1}^N G_i P_i}{\sum_{i=1}^N G_i + \sum_{i=1}^I P_i}$$

G : ground truth (manual segmentation), G_i : ground truth value for voxel i

P : predicted segmentation, P_i : predicted value for voxel i

N : number of voxels

$|G \cap P|$: number of ChP voxels in both segmentations, intersected region

$|G|, |P|$: number of ChP voxels in the segmentations

During the training and validation step, *class monai.metrics.DiceMetric* was used, setting *reduction = 'mean'* and *include_background = False* to exclude the computation of the Dice score over the first class (0: background). This is done because the Choroid Plexus is very small compared to the background, so it could be useful to exclude the Dice computation over the largest class to help convergence.

For the preprocessing in MATLAB, the built-in function for the Dice calculation in the case of a binary segmentation considers the table of segmentation classification probability (Zou et al., 2004).

$$Dice(G, P) = \frac{2 * TP}{2 * TP + FP + FN}$$

G : ground truth segmentation, P : predicted segmentation

TP : True Positive, number of voxels correctly classified as ChP voxels (with respect to the reference segmentation)

FP : False Positive, number of voxels wrongly classified as ChP voxels

FN : False Negative, number of voxels wrongly classified as background voxels

To evaluate the overall performances of a model, a mean value of this index over all the considered subjects was used.

3.7.2 Jaccard Coefficient

The Jaccard Similarity Index or Jaccard Similarity Coefficient is a measure of similarity of two segmentations. The index compares how many voxels are classified in the same way in both segmentations and how many are differently classified. There are two principal formulations of the Jaccard Coefficient (Bertels et al., 2019).

$$Jaccard(G, P) = \frac{|G \cap P|}{|G \cup P|}$$

$|G \cap P|$: Number of voxels with the same value in both sets

$|G \cup P|$: Number of voxels with different value in either set

$$Jaccard(G, P) = \frac{Dice(G, P)}{2 - Dice(G, P)}$$

The second formulation is significant because it explains the direct dependence of the Jaccard Coefficient on the Dice Coefficient.

To evaluate the overall performances of a model, it was considered a mean value of this index over all the considered subjects.

3.7.3 Hausdorff Distance

The Hausdorff Distance is a max-min distance used to evaluate the difference between the predicted segmentation and the ground truth segmentation (Hatamizadeh et al., 2021; Huttenlocher et al., 1993). In other words, it measures how far two subsets of a space are.

The Hausdorff Distance (HD) is defined as:

$$HD(G', P') = \max\{h(G, P), h(P, G)\} [mm]$$

Where $h(G, P)$ is the direct Hausdorff Distance from G to P:

$$h(G, P) = \max_{g' \in G'} \min_{p' \in P'} \|g' - p'\| [mm]$$

G', P' : ground truth and prediction surface point sets

g', p' : single points of the two sets

$\| \cdot \|$: Euclidean Norm

The Hausdorff Distance is the maximum value between the two direct Hausdorff Distances. In particular, $h(G, P)$ first finds the g' points ($\in G'$) that is farthest from all P' sets and then compute the distance as Euclidean norm between that point in G' and its nearest neighbor in P' . To find the most mismatched point of G' , it is computed the distance from each point of G' to its nearest point in P' , these distances are ranked from the smallest (low rank) to the largest (high rank) and the point with the higher rank is chosen. Using this method, the HD measures the degree of mismatch between two sets.

Following the literature (Hatamizadeh et al., 2021; Zhao et al., 2020), the used performance index is the 95% HD, so the 95th percentile of the Hausdorff Distance set points. This is done to exclude the 5% of outliers' points.

3.7.4 Percentage Volume Difference

Recent studies in literature (Fleischer et al., 2021; Lizano et al., 2019; Müller et al., 2022; Zhou et al., 2020) suggest the use of the ChP as a biomarker to study the evolution of a disease, that could be Multiple Sclerosis or a Psychotic disorder. However, what should be use as a biomarker is not the ChP as a structure, but the ChP volume. As a matter of

fact, it is observed, for example, in MS patients a higher ChP volume with respect to healthy control patients.

For these reasons, unlike Zhao et al. (Zhao et al., 2020), it was decided to calculate the volume percentage variation between the volume of the ground truth segmentation (manual segmentation) and the volume of the predicted segmentation to evaluate the error committed by the automatic algorithm (trained DNN model, FS, GMM) in estimating this parameter. The formulation is similar to that proposed by Schmidt-Mengin et al. (Schmidt-Mengin et al., 2021):

$$Volume_{percentage} = 100 * \frac{|Volume_{GT} - Volume_{Pred}|}{Volume_{GT}} [\%]$$

$Volume_{GT}$: volume of the ground truth segmentation [mm³]

$Volume_{Pred}$: volume of the predicted segmentation [mm³]

3.7.5 RMSE and MSE

RMSE (Root Mean Squared Error) and MSE (Mean Squared Error) are well described by Willmott (Willmott, 1982). The MSE is the mean quadratic-discrepancy between the predicted volume and the ground-truth volume, while the RMSE is its root version.

$$MSE = \frac{1}{N} \sum_{i=1}^N (Volume_{Pred} - Volume_{GT})^2 [\text{mm}^6]$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Volume_{Pred} - Volume_{GT})^2} [\text{mm}^3]$$

N : number of subjects

$Volume_{Pred}$: volume of the predicted segmentation [mm³]

$Volume_{GT}$: volume of the ground truth segmentation [mm³]

These two performance indices are calculated over: 9 subjects and all subjects in the preliminary analysis on MATLAB; over the 15 subjects of the validation set to compare the DNN models' performances; over the 15 subjects to make the comparison between the DNNs, FS, GMM and the gold-standard segmentation of the cT1-w.

3.7.6 Pearson's Correlation Analysis

The Pearson's Correlation Analysis between the predicted volume and the ground truth volume was performed only by Schmidt-Mengin et al. (Schmidt-Mengin et al., 2021). In this work, this parameter is evaluated in the primary analysis to identify the best MRI sequence whose manual segmentation is most related to the gold-standard manual

segmentation and in the comparison analysis (selected DNN, FS, GMM, cT1-w manual segmentation).

Both the Pearson's Linear Correlation Coefficient and the p-values are analyzed. In MATLAB Pearson's Correlation Analysis is the default correlation method. The input is a matrix that contains all the subject's volume values computed for each compared option. In the primary analysis, the compared options are the manual segmentations for each sequence (FLAIR; DIR; T1-w; cT1-w), FS and GMM. In the final analysis, the compared options are the trained DNN models, FS, GMM and the gold-standard manual segmentation of cT1-w.

The p-values matrix is a symmetric matrix used to test the hypothesis of a no correlation and that of nonzero correlation. The p-values are useful to choose the best option that is well correlated in term of significance with the reference method (the gold-standard), especially when two correlation coefficients are similar. If the p-value is lower than 0.05, the correlation is nonzero.

The Pearson's Linear Correlation Coefficient (PC) is the ratio between the covariance of two variables (methods to be compared) and the product of their standard deviation. In our case $PC(a, b)$ is a symmetric k-by-k matrix (k the number of columns, so the number of compared methods). The pairwise linear correlation coefficient is computed between column a and column b of the volume matrix X (<https://it.mathworks.com/help/stats/corr>):

$$PC(a, b) = \frac{Cov(X_a, X_b)}{\sigma(X_a)\sigma(X_b)} = \frac{\sum_{i=1}^N (X_{a,i} - \bar{X}_a)(X_{b,i} - \bar{X}_b)}{\sqrt{\{\sum_{i=1}^N (X_{a,i} - \bar{X}_a)^2 \sum_{j=1}^N (X_{b,j} - \bar{X}_b)^2\}}}$$

N : number of subjects (length of the column)

X : matrix of segmentations volume values (column: methods, rows: subjects)

$\bar{X}_a = \frac{\sum_{i=1}^N (X_{a,i})}{N}$: mean value of the column a (the same for the column b)

$Cov(X_a, X_b)$: covariance between column a and column b

$\sigma(X_a)$: sd of the column a (the same for column b)

The PC values are in range [-1, +1], where -1 is index of negative correlation, 0 no correlation, +1 positive correlation.

3.7.7 Linear Regression and OLS

These last two methods are used to compare the volume of the manual segmentations of various sequences with respect to that of the gold standard, FS and GMM in the primary analysis, but they are used also in the final comparative analysis.

The Linear Regression Analysis (Pandis, 2016) is used to predict the cT1-w volume using the volume values of the segmentations (manual or predicted) building straight-line:

$$Vol_{i-gold-standard} = \alpha + \beta * X_i$$

$Vol_{i-gold-standard}$: volume of the cT1-w segmentation (gold standard)

X_i : volume of the segmentation to be compared

α, β : parameters of the straight-line (intercept, angular coefficient)

Another annotation is:

$$Vol_{gold-standard} = X\beta + \epsilon$$

ϵ : error

β : parameters of the straight-line (intercept, angular coefficient)

The β coefficients are found minimizing an objective function S:

$$S(\beta) = \left\| Vol_{gold-standard} - X\beta \right\|^2,$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} S(\beta),$$

$$\hat{\beta} = (X^T X)^{-1} X^T Vol_{gold-standard}$$

This type of regression minimizes the sum of the squared residuals, where the residuals are the distances between each measured point $Vol_{i-gold-standard}$ and its prediction (calculated by the linear regression algorithm) among each axis.

From a clinical point of view, it is important to measure how much the volume prediction of each option is different from the gold-standard one. For this reason, the goal becomes to use a regression method where the intercept is zero, to make the comparison between the identity (perfect prediction of the gold-standard volume) with angular coefficient 1 and the linear regression built with a least squared regression method. The algorithm of the linear least squared regression has been modified to make this comparison. The name given is OLS, but it is a Linear Regression unless the intercept. Usually, OLS means Ordinary Least Squares that indicated the type of minimization done.

4 RESULTS

The first paragraph reports the preliminary analysis results performed on the sub-dataset, whereas the second paragraph reports an analysis on the whole dataset based on manual GT and existing software. The remaining paragraphs illustrates the performance of selected DNN in the trainings and validations phases. Moreover, a more extensive preliminary analysis is showed in Appendix A. The tables of the performance indices for each tested DNN are showed in Appendix B.

4.1 PRELIMINARY ANALYSIS ON THE SUB-DATASET

A preliminary analysis was carried out on a subset of the whole dataset, with the aim of understanding the usefulness of each sequence tested. A comparison in term of ChP segmentation was made between each sequence available, considering as reference the segmentation obtained with the gold-standard sequence (cT1-w).

To provide a further insight on automated segmentation methods, the segmentation provided by both GMM and FS on the selected subset of subjects was included in the preliminary analysis. The decisions taken after this preliminary analysis are illustrated in the Discussion chapter, paragraph 5.1.

Dice Coefficient: reference cT1-w

The *Table 4.1* shows the DC values for each subject and for each sequence or automated method when compared to the cT1-w sequence, considered as the Gold Standard ground truth. *Figure 4.1* reports boxplot of the same quantities reported in *Table 4.1* showing the variability for each compared method and sequence. T1-w and FLAIR sequences have the best mean scores with the lower variability, while the DIR sequence performs slightly worst. FS segmentation shows the lower Dice Coefficient, while GMM improves its results.

4. Results

SUBJECTS	T1-w	FLAIR	DIR	FS	GMM
2060	0.722	0.729	0.753	0.241	0.489
2059	0.705	0.740	0.597	0.259	0.525
2056	0.711	0.772	0.697	0.270	0.560
2051	0.595	0.584	0.556	0.324	0.499
2050	0.652	0.666	0.654	0.209	0.472
2045	0.694	0.710	0.694	0.449	0.535
1922	0.753	0.713	0.723	0.469	0.525
1889	0.662	0.683	0.688	0.312	0.514
91	0.741	0.698	0.750	0.382	0.541
MEAN	0.693	0.700	0.679	0.324	0.518
SD	0.049	0.053	0.066	0.091	0.027

Table 4.1: Dice Coefficient of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

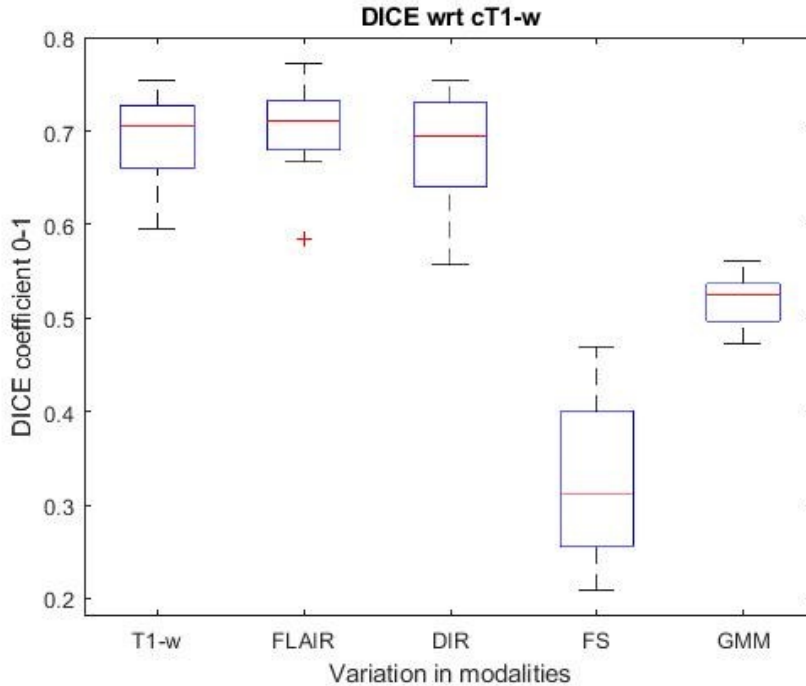


Figure 4.1: Boxplot of the Dice Coefficient median and variation intra-modality, calculated between each image modality (T1-w, FLAIR, DIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM).

Hausdorff Distance: reference cT1-w

For what concern the 95% Hausdorff Distance, DIR gives the best values in terms of mean value and variability between subjects, the SD obtained for DIR is probably contaminated by an outlier, while interquartile ranges provided by the boxplot shows a narrower interval (Table 4.2, Figure 4.2). FS and GMM segmentations give worse results with respect to both T1-w, FLAIR and DIR.

4. Results

SUBJECTS	T1-w	FLAIR	DIR	FS	GMM
2060	3.742	2.279	2.236	7.483	20.322
2059	17.148	6	22.045	10.464	24.353
2056	2.449	2.236	2.828	7	12.688
2051	11.180	8.161	6.403	8.544	10.244
2050	12.369	2.828	2.236	5.385	9.110
2045	9.222	8.367	2.236	5.661	6.481
1922	5.099	2.236	2	5.385	12.688
1889	5	3.162	3.317	6.0823	9.899
91	2	2.236	2.439	7.249	6.481
MEAN	7.579	4.167	5.082	7.028	12.474
SD	5.186	2.610	6.505	1.678	6.099

Table 4.2: 95% Hausdorff Distance of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

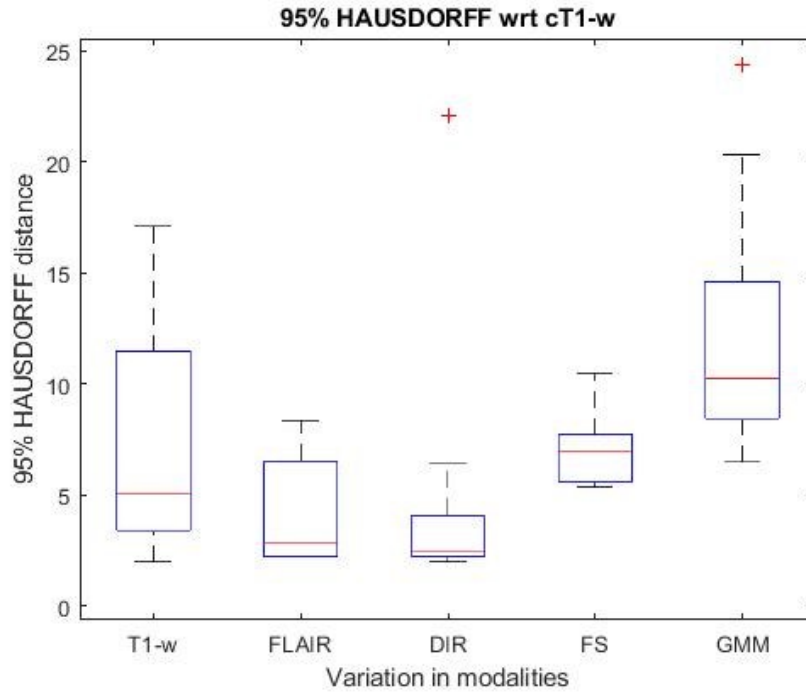


Figure 4.2: Boxplot of the 95% Hausdorff Distance median and variation intra-modality, calculated between each image modality (T1-w, FLAIR, DIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM).

Volume Analysis: reference cT1-w

The volume estimation, performed on T1-w manual segmentation has the mean value closest to the gold-standard, followed by the FLAIR sequence (Table 4.3). On the contrary, FS has the farthest mean value to the gold-standard.

4. Results

SUBJECTS	T1-w	FLAIR	DIR	FS	GMM	T1MDC
2060	4007	4941	5254	1112	1629	4381
2059	4123	4540	3825	1104	1942	3976
2056	3441	3484	5149	1085	2035	3044
2051	2960	3913	5693	1376	2946	3345
2050	2407	2557	4020	689	1401	2448
2045	3102	3026	3817	1686	2275	2621
1922	3436	3843	4367	2276	4145	3079
1889	2386	2650	3066	917	1545	2953
91	3056	3376	3473	2113	2346	2974
MEAN	3213.11	3592.22	4296	1373.11	2251.55	3202.33
SD	611.84	807.91	888.79	543.16	853.65	619.44

Table 4.3: Segmentation Volume ($1 \text{ voxel} = 1 \text{ mm}^3$) calculated for each patient for each available manual segmentation sequence, the automatic segmentations obtained with FS and GMM, and the gold-standard cT1-w sequence. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

The same trend is confirmed by the values of the calculated RMSE and MSE (Table 4.4).

SEQUENCE	MSE	RMSE
T1-w	128447	358
FLAIR	240104	490
DIR	1839042	1356
FS	4016569	2004
GMM	1957452	1399

Table 4.4: RMSE and MSE for each image modality (T1-w, FLAIR, DIR) and automatic segmentations (FS, GMM) with respect to the gold-standard cT1-w sequence.

For what concern Pearson's Correlation Analysis, T1-w and FLAIR are the sequences that provide volume estimation more correlated with cT1-w (Table 4.5). FS and GMM have lower correlation coefficient than DIR one, moreover, the p-values are very high, so there is no evidence of correlation with the gold standard segmentation.

P's CORR	T1-w	FLAIR	DIR	FS	GMM	cT1-w
T1-w	1	0.9088	0.370	0.173	0.161	0.809
FLAIR		1	0.541	0.128	0.199	0.936
DIR			1	-0.080	0.244	0.409
FS				1	0.809	-0.112
GMM					1	-0.068
cT1-w						1

Table 4.5: Pearson's Correlation Analysis coefficients between each image modality (T1-w, FLAIR, DIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM). The significative correlation coefficients ($\alpha=0,05$) are highlighted: the others are not significative.

4.2 PRELIMINARY ANALYSIS ON THE WHOLE DATASET

A further extensive preliminary analysis on the whole dataset was repeated in order to compare the gold-standard manual segmentation (cT1-w sequence) to other available approaches for the segmentation of the ChP: the manual segmentations considering T1-w or FLAIR sequences, and the automatic segmentations provided by GMM and FS. The

DIR sequences was discarded during the preliminary analysis on 9 subjects, to avoid unnecessary time-consuming procedures to the manual rater, as explained in the Discussion chapter (paragraph 5.2).

Dice Coefficient: reference cT1-w

FS segmentation shows the lower mean value of the DC, while FLAIR manual segmentation the highest one, followed by T1-w manual segmentation. The *Table 4.6* reports the mean DC value, performed with respect to the gold standard ground truth, for each compared method and sequence. The *Figure 4.3* reports the boxplot of the DC quantities, omitted in the *Table 4.6*, illustrating the variability between each term of comparison.

Segmentation	T1-w	FLAIR	FS	GMM
MEAN	0,668	0,677	0,320	0,514
SD	0,046	0,054	0,066	0,055

Table 4.6: Dice Coefficient of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Only mean and standard deviation of each image modality are reported. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

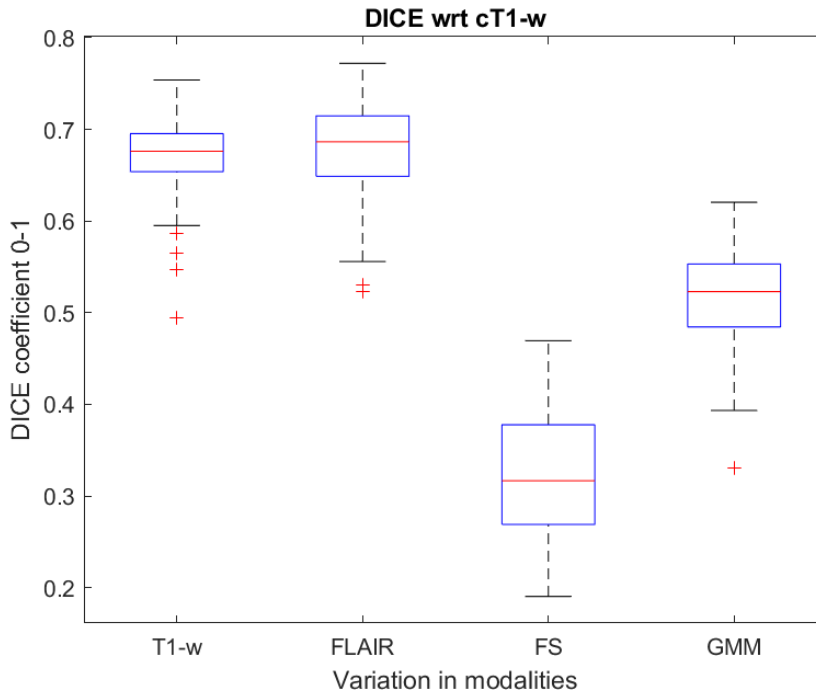


Figure 4.3: Boxplot of the Dice Coefficient median and variation intra-modality, calculated between each image modality (T1-w, FLAIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM).

The *Figure 4.4* below shows the outliers subjects for each term of comparison. T1-w and FLAIR manual segmentations have lower variability inside the dataset.

4. Results

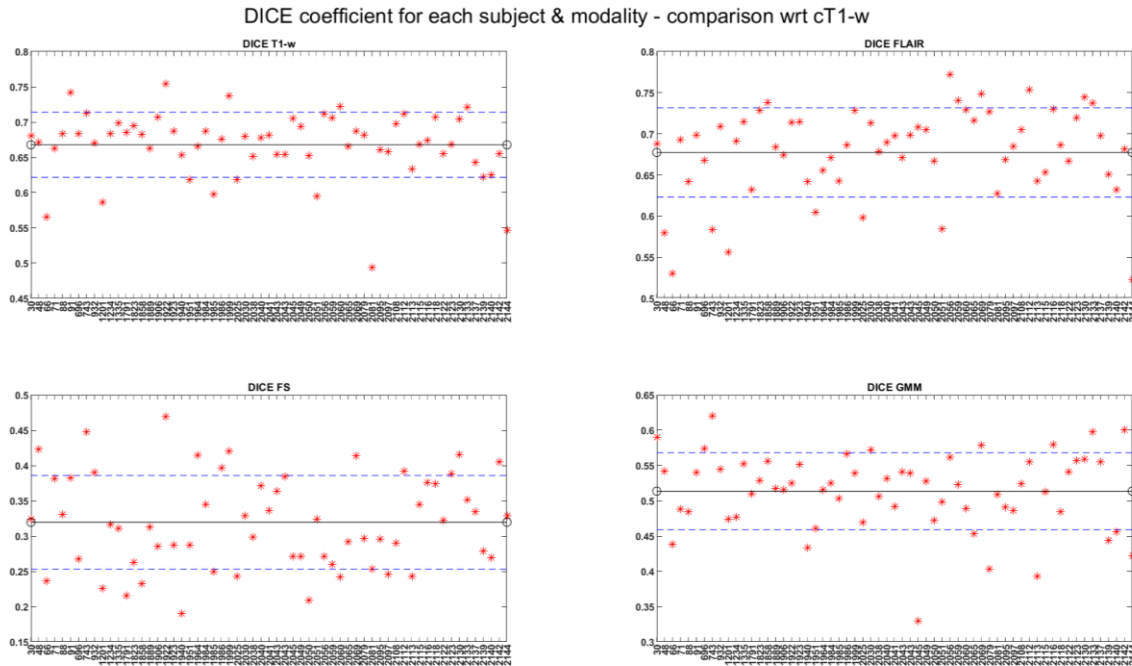


Figure 4.4: Plot of Dice Coefficient of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Mean and standard deviation of each modality are reported in the plots. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

Hausdorff Distance: reference cT1-w

Concerning the Hausdorff Distance, FLAIR gives the best values in terms of mean value and variability between subjects (Table 4.7). The trend showing FS and GMM segmentations giving less reliable results than T1-w and FLAIR ones (Figure 4.5).

Segmentation	T1-w	FLAIR	FS	GMM
MEAN	4,722	3,567	7,519	12,498
SD	3,054	2,271	1,843	4,526

Table 4.7: 95% Hausdorff Distance of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Only mean and standard deviation of each image modality are reported. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

4. Results

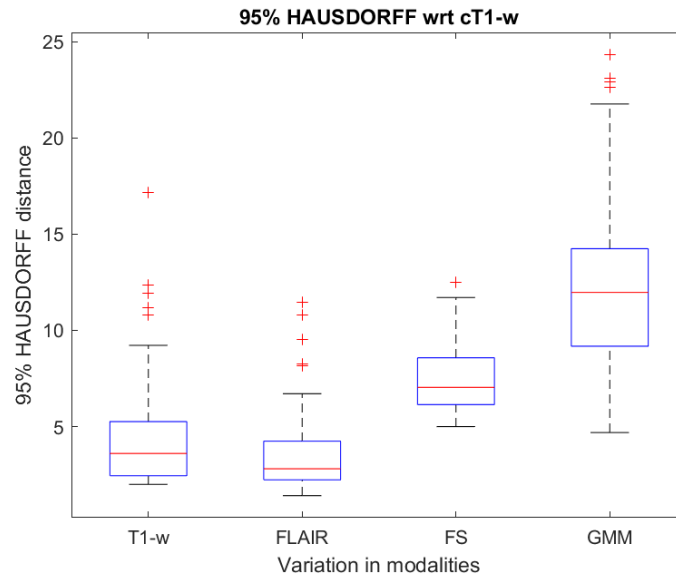


Figure 4.5: Boxplot of the 95% Hausdorff Distance median and variation intra-modality, calculated between each image modality (T1-w, FLAIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM).

The Figure 4.6 shows how the manual segmentations performed over T1-w and FLAIR images are more performant in terms of outliers than FS and GMM.

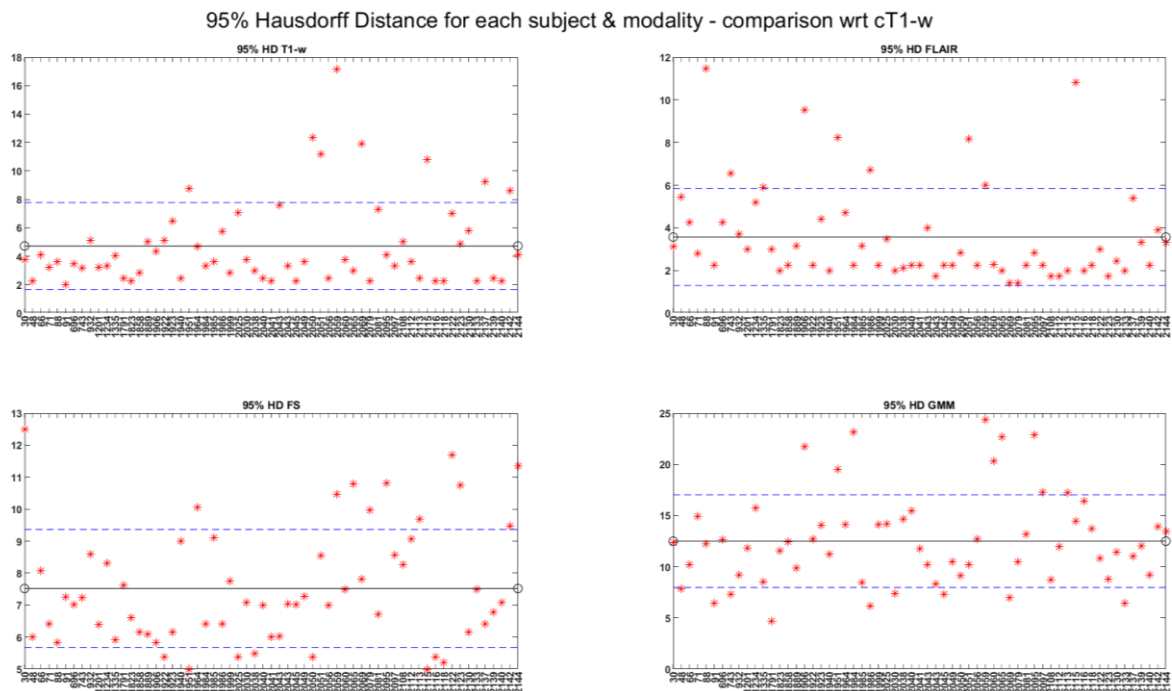


Figure 4.6: Plot of 95% Hausdorff distance of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Mean and standard deviation of each modality are reported in the plots. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

Volume Analysis: reference cT1-w

The volume analysis is the most significant because the possible biomarker is the volume estimation of the ChP that has to be as similar as possible to that extrapolated from the gold standard. T1-w manual segmentation has the mean value closest to the gold-

4. Results

standard, followed by the FLAIR one (Table 4.8). On the contrary, FS has the farthest mean value to the gold-standard one.

Segmentation	T1-w	FLAIR	FS	GMM	cT1-w
MEAN	3072,508	3786,852	1347,770	2055,573	2983,754
SD	563,077	679,237	445,649	773,060	505,905

Table 4.8: Segmentation Volume (1 voxel = 1 mm³) calculated for each patient for each available manual segmentation sequence, the automatic segmentations obtained with FS and GMM, and the gold-standard cT1-w sequence. Only mean and standard deviation of each image modality are reported.

RMSE and MSE confirms the results in term of volume estimation, showing T1-w minimum RMSE followed by FLAIR and the automated methods providing poorer reliability than manual approaches (Table 4.9).

SUBJECTS	MSE	RMSE
T1-w	138043	372
FLAIR	899576	949
FS	2981658	1727
GMM	1453816	1206

Table 4.9: RMSE and MSE for each image modality (T1-w, FLAIR) and automatic segmentations (FS, GMM) with respect to the gold-standard cT1-w sequence.

Observing the OLS (linear regression without intercept) and the linear regression analyses (Figure 4.7), GMM and FS are the segmentations that least accurately estimate the volume of the gold-standard segmentation, while T1-w is the more accurate one.

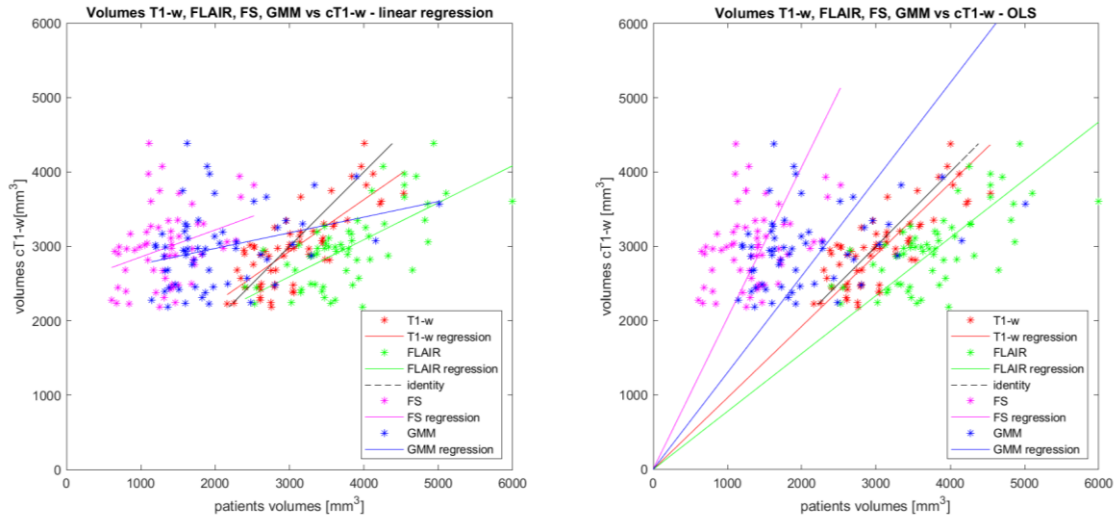


Figure 4.7: Panel left: Linear Regression of subject's volume for each modality; panel right: OLS volume representation for each modality.

For what concern Pearson's Correlation Analysis (Table 4.10, Figure 4.8), T1-w and FLAIR are the sequences more correlated with cT1-w. FS and GMM have far lower correlation coefficient than manual segmentation obtained with both T1-w and FLAIR.

4.Results

P's CORR	T1-w	FLAIR	FS	GMM	cT1-w
T1-w	1	0,752	0,389	0,427	0,773
FLAIR		1	0,316	0,414	0,667
FS			1	0,609	0,319
GMM				1	0,321
cT1-w					1

Table 4.10: Pearson's Correlation Analysis coefficients between each image modality (T1-w, FLAIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM). All Correlation coefficients are significant ($\alpha=0,05$).

Pearson Correlation Analysis

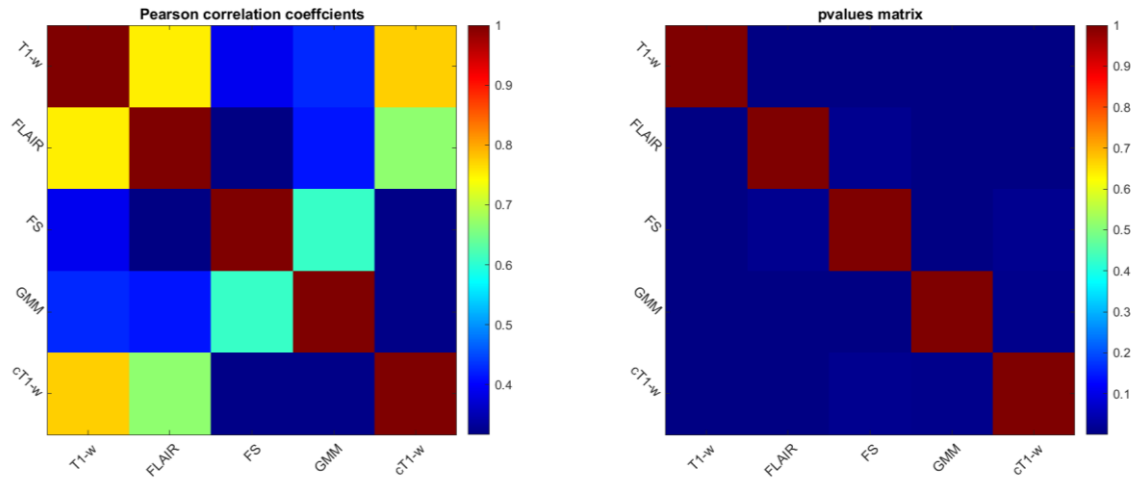


Figure 4.8: Left panel: Pearson's Correlation Coefficients; right panel: p-values matrix.

Percentage Volume Difference: reference cT1-w

In terms of Percentage Volume Difference (Table 4.11), that estimate the percentage discrepancy between the gold standard volume estimation and the proposed method or sequence one, T1-w is the sequence with the lower mean value and variability (Figure 4.9), followed by the FLAIR one. FS segmentation it is the method that gives the estimate of the volume furthest from the gold standard one.

SUBJECTS	T1-w	FLAIR	FS	GMM
MEAN	10,574	28,427	54,405	34,351
SD	7,600	18,395	14,144	17,291

Table 4.11: Percentage Volume Difference of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Only mean and standard deviation of each image modality are reported. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

4. Results

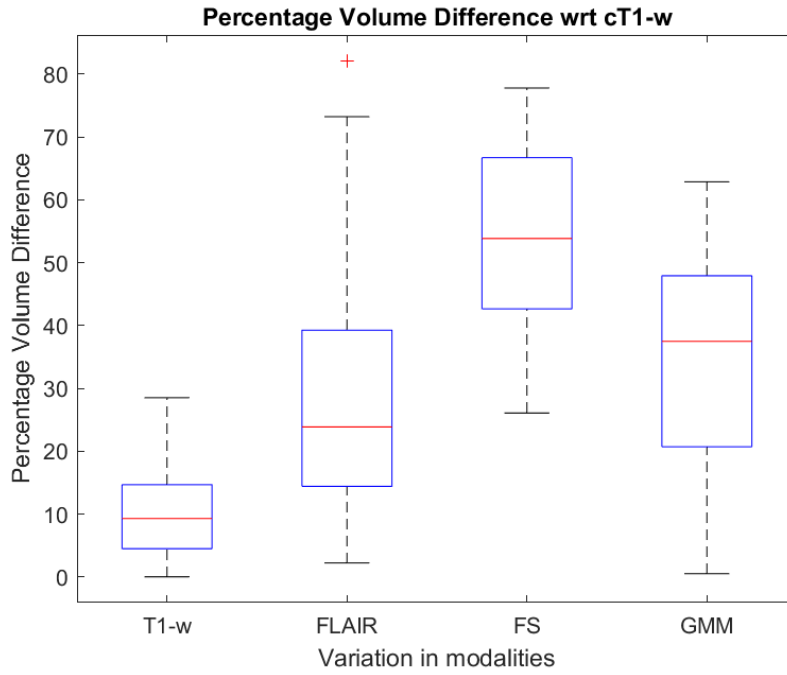


Figure 4.9: Boxplot of the Percentage Volume Difference median and variation intra-modality, calculated between each image modality (T1-w, FLAIR) and the gold-standard cT1-w sequence. The same was done for the automatic segmentations (FS, GMM).

Regarding the outliers, the FLAIR segmentation has the lower number, while the automatic methods the higher (Figure 4.10).

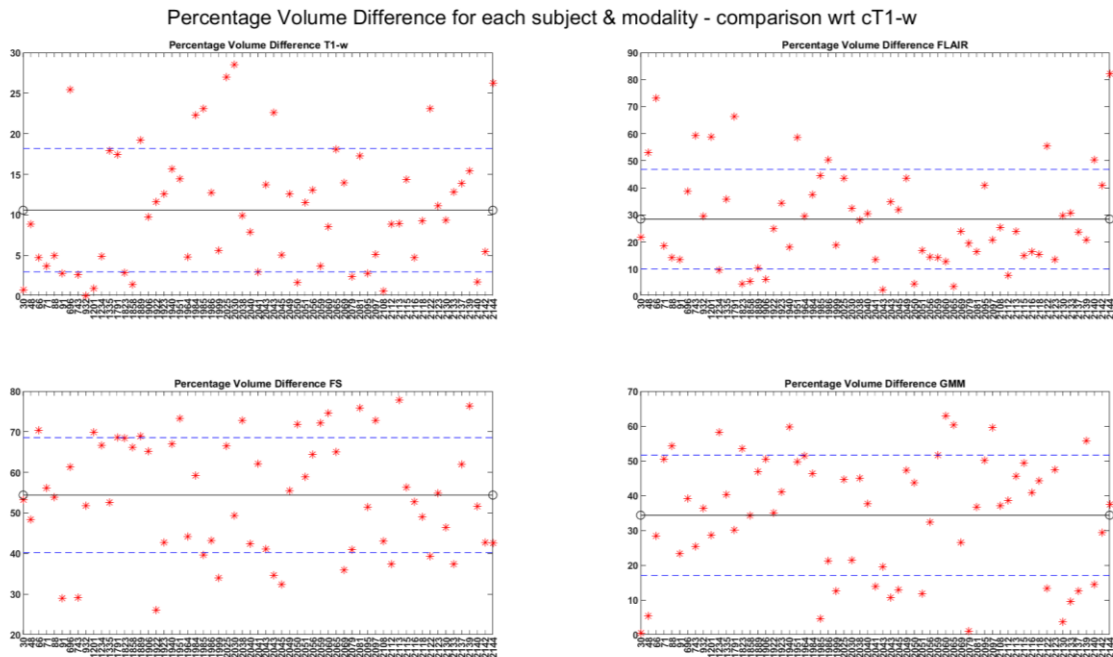


Figure 4.10: Plot of Percentage Volume Difference of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Mean and standard deviation of each modality are reported in the plots. Automated segmentation obtained with FS and GMM are also reported in the same fashion.

4.3 DNN VALIDATION SET RESULTS

The analysis of the performance indices for the validation set is reported in the following paragraphs. The paragraphs 4.4 and 4.5 show the above results for the trainings performed respectively with GT cT1-w (with all combination of inputs, explained in the paragraph) and GT without contrast (T1-w, FLAIR). The performance indices are calculated always considering as reference the gold standard ground truth (cT1-w manual segmentation), that is the target to be matched. The paragraph 4.6 reports the comparison between the trainings with and without contrast as GT to select the best input-GT combination, both with and without contrast, that minimizes the Percentage Volume Difference. The paragraph 4.7 reports the comparison analysis between the state-of-the-art automatic method (FS and GMM) and the two DNNs selected in the paragraph 4.6. In the following paragraphs there are references to Tables that are completely reported in Appendix B.

The Tables reported in Appendix B in paragraphs B.2 and B.3 reports the mean values (and standard deviation ones) of the performance indices or the single value over all subjects (RMSE). The MSE and the Volume Difference are not reported because of the presence of the RMSE and the Percentage Volume Difference. The discussion of the results for the paragraph 4.4. and 4.5 is reported in the Discussion chapter, paragraph 5.3.

4.3.1 Legend

The DNNs are named considering the training variable parameters used as follows:

DNN_PatchSize_LossFunction_DataAugmentation

where DNN, PatchSize, LossFunction and DataAugmentation represent:

- DNN (Deep Neural Network architecture): 3DUNET (3D U-Net), DynUNET (nnU-Net), VNET (V-Net), UNETR
- PatchSize (Patch Size): 64 (64x64x64), 96 (96x96x96), 128 (128x128x128)
- LossFunction (Loss Function): Dice (Generalized Dice Loss), DiceCE (Combination of Dice Loss and Cross-Entropy Loss), CE (Cross-Entropy Loss), wCE (Weighted-Cross-Entropy)
- DataAugmentation (Data Augmentation): DA (Data Augmentation Transforms applied), noDA (no application of Data Augmentation Transforms)

4.4 DNN VALIDATION SET RESULTS: TRAINING WITH GROUND TRUTH WITH CONTRAST

The performance indices are calculated making the comparison between the predicted segmentation obtained with each DNN and the manual segmentation obtained from the images with contrast, that is considered henceforth the gold standard ground truth (GT) (cT1-w). Three are the input – manual segmentation (MSeg) examined combinations: T1-w – cT1-w; FLAIR – cT1-w; T1-w+FLAIR – cT1-w.

The complete Tables of results (96 combinations of DNN, loss function, patch size and Data Augmentation) for each input-MSeg combination (paragraph 4.4.1, 4.4.2, 4.4.3) are showed in Appendix B- B.2, while the input-MSeg analysis for each DNN architecture (3D U-Net, UNETR, V-Net, nnU-Net) is reported in paragraph B.4.

The principal results for each combination are showed in the following sub-paragraphs.

Considering all 96 possible combinations of DNN, loss function, patch size and data augmentation transforms, the results were ordered according to the main performance indices: mean Dice Coefficient, mean 95% Hausdorff Distance, mean Percentage Volume Difference (RMSE follows the Percentage Volume Difference trend).

4.4.1 Input T1-w, GT cT1-w - reference cT1-w

Ordering by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN type as showed in *Table 4.12*, even if there is not a specific trend looking at the loss function, the patch size and the presence of Data Augmentation. The worst DNN in terms of DC is the V-Net. Generally, the Weighted Cross-Entropy loss function brings to lower DC and the higher the patch size, the better the performance except for the nnU-Net that performs equally well for independently from patch sizes (See table in Appendix B-B.2).

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_Dice_DA	0,77	0,04	1,70	1,35	9,33	7,68
nnU-Net_96_Dice_DA	0,77	0,04	1,59	0,72	9,36	7,34
nnU-Net_96_DiceCE_DA	0,77	0,03	1,47	0,33	7,86	6,60
nnU-Net_96_CE_DA	0,76	0,04	1,47	0,33	7,94	5,20
nnU-Net_128_DiceCE_DA	0,76	0,04	1,83	1,77	9,04	4,87
nnU-Net_64_CE_DA	0,76	0,04	1,59	0,77	8,58	5,26
nnU-Net_128_CE_DA	0,76	0,04	1,55	0,59	8,00	5,07
nnU-Net_96_Dice_noDA	0,76	0,04	1,50	0,53	9,34	6,34
nnU-Net_128_Dice_noDA	0,76	0,04	1,70	1,01	10,68	8,16
nnU-Net_64_Dice_noDA	0,76	0,04	9,76	30,07	10,12	8,06

Table 4.12: Best 10 DNNs sorted by mean Dice Coefficient.

4. Results

Ordering by 95% Hausdorff Distance (95% HD mean), nnU-Net is the most performant DNN architecture as showed in *Table 4.13*, mainly with higher patch size, as for the UNETR and 3D U-Net. However, there is no difference between the first ten DNNs in terms of 95% HD. Generally, V-Net is the worst DNN type, however the bigger the patch size, the better the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_wCE_noDA	0,75	0,04	1,47	0,34	19,55	11,09
nnU-Net_96_DiceCE_DA	0,77	0,03	1,47	0,33	7,86	6,60
nnU-Net_96_CE_DA	0,76	0,04	1,47	0,33	7,94	5,20
nnU-Net_128_CE_noDA	0,76	0,04	1,49	0,33	9,84	6,53
nnU-Net_96_Dice_noDA	0,76	0,04	1,50	0,53	9,34	6,34
UNETR_96_CE_DA	0,75	0,04	1,54	0,33	8,30	3,83
nnU-Net_128_CE_DA	0,76	0,04	1,55	0,59	8,00	5,07
nnU-Net_64_wCE_noDA	0,73	0,04	1,57	0,37	32,83	12,79
3DUNET_96_Dice_DA	0,74	0,04	1,58	0,32	13,81	9,59
3DUNET_128_Dice_DA	0,74	0,04	1,58	0,32	14,67	10,26

Table 4.13: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

Sorting by mean Percentage Volume Difference (% Vol Diff mean), UNETR and nnU-Net are the best DNNs type and the bigger the patch size, the better the performance as shown in *Table 4.14*. Once again, Weighted Cross-Entropy gives the worst performance as the V-Net (See table in Appendix B-B.2).

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
UNETR_128_Dice_DA	0,75	0,04	1,66	0,50	7,54	3,79
nnU-Net_96_DiceCE_DA	0,77	0,03	1,47	0,33	7,86	6,60
nnU-Net_96_CE_DA	0,76	0,04	1,47	0,33	7,94	5,20
nnU-Net_128_CE_DA	0,76	0,04	1,55	0,59	8,00	5,07
UNETR_96_CE_DA	0,75	0,04	1,54	0,33	8,30	3,83
nnU-Net_64_CE_DA	0,76	0,04	1,59	0,77	8,58	5,26
UNETR_128_DiceCE_noDA	0,73	0,05	1,92	0,89	8,72	6,69
3DUNET_64_DiceCE_DA	0,73	0,04	1,83	0,62	8,74	5,98
UNETR_128_CE_DA	0,74	0,04	1,90	1,12	8,98	4,78
nnU-Net_128_DiceCE_DA	0,76	0,04	1,83	1,77	9,04	4,87

Table 4.14: Best 10 DNNs sorted by mean Percentage Volume Difference.

Generally, no trends were found for Data Augmentation.

Considering all the combinations for each DNN architecture, the results highlight the following. The complete tables are showed in Appendix B – B.4.1. There are 24 possible combinations for each architecture.

4. Results

Across all architectures there are no differences in terms of standard deviation for the Dice Coefficient, except for the V-Net that has a higher variability. Generally, the Weighted Cross-Entropy loss function gives lower mean DC values and higher mean Percentage Volume Difference. Moreover, the six combinations with wCE are the outliers for the volume performance index (the reported range of values don't consider these six ones).

Concerning the 3D U-Net, the mean Dice Coefficient range of values is [0,68-0,74]. For what concern the mean 95% Hausdorff Distance, the range is [1,58-2,44] and the standard deviation correspondent range is [0,32-2,71]. There is only one outlier (3DUNET_64_Dice_DA) with a 95% HD of $(26,90 \pm 50,06)$. The results show that the lower is the patch size, the lower is the mean 95% HD. Sorting by mean Percentage Volume Difference, the range of values is [8,74-15,35], with standard deviation range [4,39-10,26].

For what concern the nnU-Net, the mean Dice Coefficient range is [0,72-0,77]. The range of the 95% Hausdorff Distance is [1,47-1,97] and the standard deviation correspondent range is [0,33-1,77]. There are two outliers (nnU-Net_64_Dice_noDA, nnU-Net_64_Dice_DA) with a 95% HD respectively of $(9,76 \pm 30,07)$ and $(25,54 \pm 43,70)$. Sorting by mean Percentage Volume Difference, the range of values is [7,86-11,12], with standard deviation range [4,87-10,32].

For the UNETR, the mean Dice Coefficient range of is [0,71-0,75]. The range of the 95% Hausdorff Distance is [1,54-2,33] and the standard deviation correspondent range is [0,32-2,00]. There are two outliers (UNETR_64_Dice_noDA, UNETR_96_Dice_noDA) with a 95% HD respectively of $(13,18 \pm 42,39)$ and $(19,87 \pm 46,47)$. The results show that the lower is the patch size, the lower is the mean 95% HD. Sorting by mean Percentage Volume Difference, the range of values is [7,54-18,51], with standard deviation range [3,79-11,55].

Concerning the V-Net, the mean Dice Coefficient range of values is [0,00-0,63] and there are differences in terms of standard deviation because of the higher number of outliers due to the Dice loss function that gives the lower mean DC values. The mean 95% Hausdorff Distance range is [2,47-7,63] and the standard deviation correspondent range is [0,5-4,10]. There are nine outliers, six of those are DNNs combinations with Dice loss function. Sorting by mean Percentage Volume Difference, the range of values is higher

than those of the other DNNs type and Weighted Cross-Entropy and Dice loss functions give the worst results in terms of this performance index.

4.4.2 Input FLAIR, GT cT1-w - reference cT1-w

Arranging by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN type as showed in *Table 4.15*. The Dice loss function gives better performance in terms of DC. Generally, the higher the patch size, the better the performance also for the nnU-Net.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_Dice_DA	0,76	0,05	1,81	1,06	12,10	8,91
nnU-Net_96_Dice_noDA	0,75	0,05	1,80	1,12	10,22	8,21
nnU-Net_96_Dice_DA	0,75	0,05	1,93	1,22	9,41	7,66
nnU-Net_64_Dice_noDA	0,75	0,05	9,97	29,97	10,13	7,05
nnU-Net_96_CE_DA	0,75	0,05	1,76	0,81	10,86	8,10
nnU-Net_128_CE_DA	0,75	0,06	2,19	2,15	11,48	9,14
nnU-Net_64_CE_DA	0,75	0,05	1,84	0,97	8,66	7,91
nnU-Net_128_Dice_noDA	0,75	0,05	1,72	0,69	9,92	7,58
nnU-Net_96_DiceCE_noDA	0,74	0,05	1,87	0,79	10,49	8,94
3DUNET_96_Dice_DA	0,74	0,05	1,75	0,62	11,04	9,74

Table 4.15: Best 10 DNNs sorted by mean Dice Coefficient.

Sorting by 95% Hausdorff Distance (95% HD mean), nnU-Net is the most performant DNN type as showed in *Table 4.16*, mainly with higher patch size, the second one is the 3D U-Net. However, there is no difference between the first ten DNNs in terms of 95% HD. Generally, the bigger the patch size, the better the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_Dice_noDA	0,75	0,05	1,72	0,69	9,92	7,58
nnU-Net_128_wCE_noDA	0,73	0,05	1,73	0,64	12,10	8,70
3DUNET_96_Dice_DA	0,74	0,05	1,75	0,62	11,04	9,74
nnU-Net_96_wCE_noDA	0,73	0,05	1,75	0,75	18,77	13,22
nnU-Net_96_CE_DA	0,75	0,05	1,76	0,81	10,86	8,10
nnU-Net_96_wCE_DA	0,71	0,04	1,76	0,54	55,23	13,62
3DUNET_64_DiceCE_DA	0,73	0,05	1,79	0,73	9,40	6,08
3DUNET_128_Dice_DA	0,73	0,05	1,80	0,73	11,76	9,73
nnU-Net_96_Dice_noDA	0,75	0,05	1,80	1,12	10,22	8,21
nnU-Net_128_Dice_DA	0,76	0,05	1,81	1,06	12,10	8,91

Table 4.16: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

4. Results

Ordering by mean Percentage Volume Difference (% Vol Diff mean), nnU-Net and UNETR (with high patch size) are the best DNNs type as shown in *Table 4.17*. Once again, Weighted Cross-Entropy gives the worst performance as the V-Net.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_64_CE_DA	0,75	0,05	1,84	0,97	8,66	7,91
3DUNET_64_DiceCE_DA	0,73	0,05	1,79	0,73	9,40	6,08
nnU-Net_96_Dice_DA	0,75	0,05	1,93	1,22	9,41	7,66
nnU-Net_64_Dice_DA	0,74	0,05	1,95	0,90	9,75	7,32
nnU-Net_128_Dice_noDA	0,75	0,05	1,72	0,69	9,92	7,58
UNETR_128_DiceCE_DA	0,72	0,06	2,00	1,06	9,92	5,45
UNETR_128_Dice_DA	0,73	0,05	1,86	0,69	9,95	7,50
nnU-Net_64_Dice_noDA	0,75	0,05	9,97	29,97	10,13	7,05
nnU-Net_96_Dice_noDA	0,75	0,05	1,80	1,12	10,22	8,21
3DUNET_64_Dice_DA	0,73	0,04	9,21	27,11	10,28	8,67

Table 4.17: Best 10 DNNs sorted by mean Percentage Volume Difference.

Generally, there is not a specific trend looking at the presence of Data Augmentation, while the V-Net and the Weighted Cross-Entropy lower the performances.

Considering all the combinations for each DNN architecture, the results confirm the above. The complete tables are showed in Appendix B – B.4.2. Indeed, generally, there are no trends for the application of Data Augmentation transforms and there are no differences in terms of standard deviation between the Dice Coefficient (except for the V-Net). The Weighted Cross-Entropy and the V-Net give lower performance above all in terms of DC and Percentage volume Difference. Moreover, the six combinations with wCE are the outliers for the volume performance index (the reported range of values don't consider these six ones).

For the 3D U-Net, the mean Dice Coefficient range of values is [0,69-0,74]. Generally, the Dice and DiceCE loss functions the higher ones. The mean 95% Hausdorff Distance range is [1,75-2,73] and the standard deviation correspondent range is [0,62-2,91]. There is only one outlier (3DUNET_64_Dice_DA) with a 95% HD of $(9,21 \pm 27,11)$. Sorting by mean Percentage Volume Difference, the range of values is [9,40-16,71], with standard deviation range [6,08-14,00].

Regarding the nnU-Net, the mean Dice Coefficient range of values is [0,71-0,76]. Generally, the Dice loss the higher one. For what concern the mean 95% Hausdorff Distance, the range is [1,72-2,51] and the standard deviation correspondent range is [0,69-2,15]. There is only one outlier (nnU-Net_64_Dice_noDA) with a 95% HD of $(9,97 \pm$

29,97). Sorting by mean Percentage Volume Difference, the range of values is [8,66-13,55], with standard deviation range [7,05-11,74]. Generally, Dice loss function gives better results.

Concerning the UNETR, the mean Dice Coefficient range of values is [0,67-0,74]. The mean 95% Hausdorff Distance range is [1,84-3,28] and the standard deviation correspondent range is [0,69-3,14]. There are no outliers. Sorting by mean Percentage Volume Difference, the range of values is [9,92-16,27], with standard deviation range [5,45-15,11]. The higher is the patch size and in presence of Dice loss function, the lower is the Percentage Volume Difference.

For the V-Net, the mean Dice Coefficient range of values is [0,00-0,69] and there are differences in terms of standard deviation because of the higher number of outliers. The mean 95% Hausdorff Distance range is [2,24-9,65] for a half of the combinations, while the others are all outliers, six of those are DNNs combinations with Dice loss function. Sorting by mean Percentage Volume Difference, the range of values is higher than those of the other DNNs type and Weighted Cross-Entropy and Dice loss functions give the worst results in terms of this performance index. Generally, the Dice loss function gives the worst results across all performance indices.

4.4.3 Input T1-w+FLAIR, GT cT1-w - reference cT1-w

Ordering by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN type as showed in *Table 4.18*. The Cross-Entropy and the Dice loss functions improve the performance. The worst DNN in terms of DC is the V-Net. The higher the patch size, the better the performance except for the nnU-Net.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_Dice_DA	0,77	0,04	1,56	0,72	8,62	6,33
nnU-Net_128_CE_DA	0,77	0,04	1,48	0,43	9,91	6,61
nnU-Net_128_Dice_DA	0,76	0,04	1,46	0,38	9,08	6,75
nnU-Net_128_Dice_noDA	0,76	0,04	1,58	0,56	8,34	5,92
nnU-Net_64_CE_noDA	0,76	0,04	1,79	1,33	8,75	5,22
nnU-Net_96_Dice_noDA	0,76	0,05	2,06	1,84	10,00	7,53
nnU-Net_64_CE_DA	0,76	0,05	1,52	0,65	9,37	5,87
nnU-Net_96_CE_DA	0,76	0,05	1,68	1,33	9,43	6,07
nnU-Net_128_CE_noDA	0,76	0,05	1,62	0,72	9,93	5,47
UNETR_96_DiceCE_DA	0,76	0,04	1,57	0,39	10,07	6,95

Table 4.18: Best 10 DNNs sorted by mean Dice Coefficient.

4. Results

Sorting by 95% Hausdorff Distance (95% HD mean), nnU-Net is the most performant DNN type as showed in *Table 4.19*, mainly with higher patch size, as for the UNETR and 3D U-Net. However, there is no difference between the first ten DNNs in terms of 95% HD. Generally, the bigger the patch size, the better the performance. No patterns were found in terms of loss function or Data Augmentation.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_Dice_DA	0,76	0,04	1,46	0,38	9,08	6,75
nnU-Net_128_CE_DA	0,77	0,04	1,48	0,43	9,91	6,61
nnU-Net_128_DiceCE_DA	0,76	0,04	1,50	0,38	8,23	5,81
3DUNET_96_Dice_DA	0,76	0,05	1,50	0,46	11,16	8,81
nnU-Net_64_CE_DA	0,76	0,05	1,52	0,65	9,37	5,87
nnU-Net_96_Dice_DA	0,77	0,04	1,56	0,72	8,62	6,33
UNETR_96_DiceCE_DA	0,76	0,04	1,57	0,39	10,07	6,95
nnU-Net_128_Dice_noDA	0,76	0,04	1,58	0,56	8,34	5,92
UNETR_64_Dice_noDA	0,76	0,04	1,59	0,53	11,53	7,97
nnU-Net_128_wCE_noDA	0,73	0,04	1,59	0,39	12,79	9,20

Table 4.19: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

Ordering by mean Percentage Volume Difference (% Vol Diff mean), UNETR and nnU-Net are the best DNNs type and the bigger the patch size, the better the performance as shown in *Table 4.20*. Generally, the DiceCE and the Dice loss function improve this performance index, as the high patch size.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_DiceCE_DA	0,76	0,04	1,50	0,38	8,23	5,81
nnU-Net_128_Dice_noDA	0,76	0,04	1,58	0,56	8,34	5,92
UNETR_128_DiceCE_DA	0,75	0,05	1,73	0,74	8,34	5,33
nnU-Net_96_DiceCE_noDA	0,75	0,05	1,66	0,55	8,47	6,23
nnU-Net_96_Dice_DA	0,77	0,04	1,56	0,72	8,62	6,33
nnU-Net_64_CE_noDA	0,76	0,04	1,79	1,33	8,75	5,22
nnU-Net_128_Dice_DA	0,76	0,04	1,46	0,38	9,08	6,75
UNETR_128_Dice_noDA	0,74	0,05	1,69	0,39	9,25	5,06
nnU-Net_64_CE_DA	0,76	0,05	1,52	0,65	9,37	5,87
UNETR_96_DiceCE_noDA	0,74	0,05	2,00	1,65	9,39	6,22

Table 4.20: Best 10 DNNs sorted by mean Percentage Volume Difference.

Generally, the presence of Data Augmentation does not modify the results. Moreover, the Weighted Cross-Entropy loss and the V-Net lower the performance.

Considering all the combinations for each DNN type, the results highlight the following (for the sake of clarity, the complete tables are showed in Appendix B – B.4.3). As in the previous input-MSeg combinations, there is no difference in terms of standard deviation

for the DC. As reported for the general performance indices analysis, the V-Net and the Weighted Cross-Entropy loss give lower performance, mostly for the Percentage Volume Difference.

For the 3D U-Net, the mean Dice Coefficient range of values is [0,71-0,76]. The Dice loss gives the higher DC values. The mean 95% Hausdorff Distance range is [1,50-2,24] and the standard deviation correspondent range is [0,35-2,00]. There are no outliers. The results show that the Dice loss gives lower mean 95% HD. Sorting by mean Percentage Volume Difference, the range of values is [9,67-13,91], with standard deviation range [4,52-10,03].

Concerning the nnU-Net, the mean Dice Coefficient range of values is [0,70-0,77]. Looking at the mean 95% Hausdorff Distance, the range is [1,46-3,07] and the standard deviation correspondent range is [0,38-1,84]. There are three outliers (nnU-Net_64_wCE_DA, nnU-Net_64_Dice_noDA, nnU-Net_64_Dice_DA). Sorting by mean Percentage Volume Difference, the range of values is [8,23-15,80], with standard deviation range [5,22-14,67].

Regarding the UNETR, the mean Dice Coefficient range of values is [0,71-0,76]. The range of the 95% Hausdorff Distance is [1,57-2,07] and the standard deviation correspondent range is [0,39-1,78]. There are two outliers (UNETR_96_Dice_DA, UNETR_64_Dice_DA) with a 95% HD respectively of $(11,31 \pm 35,52)$ and $(34,53 \pm 56,71)$. Sorting by mean Percentage Volume Difference, the range of values is [8,34-16,90], with standard deviation range [5,06-11,75]. Generally, the Dice and DiceCE loss functions and the higher patch size improve this performance index.

With reference to the V-Net, the mean Dice Coefficient range of values is [0,00-0,66] and there are differences in terms of standard deviation because of the higher number of outliers due to the Dice loss function that gives the lower mean DC values. Looking at the mean 95% Hausdorff Distance, only six combinations, four of which with Weighted Cross-Entropy loss function, have a mean value lower than 7,69. Sorting by mean Percentage Volume Difference, the range of values is higher than those of the other DNNs type and the Dice loss functions gives the worst results in terms of this performance index, while the DiceCE loss and the higher patch size improve the performance.

4.5 DNN VALIDATION SET RESULTS: TRAINING WITH GROUND TRUTH WITHOUT CONTRAST, PERFORMANCE INDICES CALCULATED WITH RESPECT TO THE CONTRAST GROUND TRUTH

The performance indices are calculated making the comparison between the predicted segmentation obtained with each DNN architecture, trained with the ground truth manual segmentation obtained from the images without contrast (T1-w, FLAIR), and the gold-standard manual segmentation (cT1-w). Four are the input – MSeg examined combinations: T1-w – T1-w; FLAIR – FLAIR; T1-w+FLAIR – T1-w; T1-w+FLAIR - FLAIR. Considering the bad performance obtained with the V-Net, it was excluded from this comparison.

The complete Tables of results (72 combinations of DNN, loss function, patch size and Data Augmentation) for each input-MSeg combination (paragraph 4.5.1, 4.5.2, 4.5.3, 4.5.4) are showed in Appendix B- B.3, while the input-MSeg analysis for each DNN architecture is reported in paragraph B.5.

The principal results for each combination are showed in the following sub-paragraphs.

Considering all 72 possible combinations of DNN, loss function, patch size and data augmentation transforms, the results were sorted according to the main performance indices: mean Dice Coefficient, mean 95% Hausdorff Distance, mean Percentage Volume Difference (RMSE follows the Percentage Volume Difference trend).

4.5.1 Input T1-w, GT T1-w – reference cT1-w

Ordering by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN type, followed by UNETR, as showed in *Table 4.21*. Generally, the higher the patch size, the better the performance also for the nnU-Net.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_CE_DA	0,71	0,04	3,48	1,87	7,40	5,91
nnU-Net_128_Dice_DA	0,71	0,04	2,99	0,97	12,92	10,40
nnU-Net_128_DiceCE_DA	0,71	0,04	3,05	0,98	8,99	7,45
nnU-Net_128_CE_DA	0,70	0,04	3,09	1,12	7,93	5,43
nnU-Net_96_Dice_noDA	0,70	0,04	2,86	0,97	8,68	6,60
UNETR_96_DiceCE_DA	0,70	0,04	3,57	2,47	8,61	6,67
UNETR_96_CE_DA	0,70	0,04	2,80	1,28	8,47	4,11
nnU-Net_96_Dice_DA	0,70	0,04	21,53	47,65	17,46	11,63
nnU-Net_96_DiceCE_noDA	0,70	0,04	3,73	1,59	8,29	6,24
nnU-Net_64_CE_DA	0,70	0,04	3,05	1,17	9,08	7,22

Table 4.21: Best 10 DNNs sorted by mean Dice Coefficient.

4. Results

Sorting by 95% Hausdorff Distance (95% HD mean), UNETR is the most performant DNN type as showed in *Table 4.22*, mainly with higher patch size, the second one is the nnU-Net. However, there is no difference between the first ten DNNs in terms of 95% HD. Generally, the Dice and DiceCE loss function give higher 95% HD, while Weighted Cross-Entropy lower.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
UNETR_128_Dice_DA	0,69	0,04	2,51	0,79	19,00	11,71
UNETR_128_DiceCE_DA	0,70	0,04	2,52	0,67	13,19	8,34
UNETR_64_wCE_noDA	0,67	0,04	2,61	0,65	43,77	18,20
UNETR_64_wCE_DA	0,68	0,04	2,63	1,04	27,90	16,77
nnU-Net_64_wCE_noDA	0,67	0,04	2,66	1,03	33,56	14,76
UNETR_64_CE_DA	0,70	0,05	2,70	0,41	8,61	6,85
3DUNET_96_wCE_noDA	0,64	0,05	2,77	0,62	50,15	18,66
nnU-Net_96_CE_noDA	0,69	0,05	2,79	0,89	9,28	5,37
UNETR_96_wCE_DA	0,67	0,04	2,80	0,58	43,31	14,15
UNETR_96_CE_DA	0,70	0,04	2,80	1,28	8,47	4,11

Table 4.22: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

Ordering by mean Percentage Volume Difference (% Vol Diff mean), nnU-Net and UNETR (with high patch size) are the best DNNs type as shown in *Table 4.23*. The CE and DiceCE loss functions improve the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_CE_DA	0,71	0,04	3,48	1,87	7,40	5,91
nnU-Net_128_CE_DA	0,70	0,04	3,09	1,12	7,93	5,43
3DUNET_64_DiceCE_DA	0,67	0,04	3,41	2,24	8,22	6,03
nnU-Net_96_DiceCE_noDA	0,70	0,04	3,73	1,59	8,29	6,24
nnU-Net_64_CE_noDA	0,69	0,04	3,33	1,37	8,44	5,63
UNETR_96_CE_DA	0,70	0,04	2,80	1,28	8,47	4,11
UNETR_96_DiceCE_noDA	0,69	0,05	4,03	2,86	8,47	5,85
UNETR_96_DiceCE_DA	0,70	0,04	3,57	2,47	8,61	6,67
UNETR_64_CE_DA	0,70	0,05	2,70	0,41	8,61	6,85
nnU-Net_96_Dice_noDA	0,70	0,04	2,86	0,97	8,68	6,60

Table 4.23: Best 10 DNNs sorted by mean Percentage Volume Difference.

Generally, there is not a specific trend looking at the presence of Data Augmentation, while the Weighted Cross-Entropy loss function lowers the performance.

Considering all the combinations for each DNN type, the results highlight the following. The complete tables are showed in Appendix B – B.5.1.

Generally, there is no difference in terms of standard deviation for the DC values, while the wCE lowers the performance, in particular for the Percentage Volume Difference.

For the 3D U-Net, the mean Dice Coefficient range of is [0,62-0,68]. Generally, the Dice and DiceCE loss functions the higher ones. Looking at the mean 95% Hausdorff Distance, the range is [2,77-4,53] and the standard deviation correspondent range is [0,62-2,31]. There are two outliers (3DUNET_64_Dice_DA, 3DUNET_96_Dice_DA) with a 95% HD respectively of $(11,23 \pm 31,19)$ and $(13,09 \pm 35,29)$. The Weighted Cross-Entropy improve the 95% HD values. Sorting by mean Percentage Volume Difference, the range of values is [8,22-15,03], with standard deviation range [6,03-10,96].

Concerning the nnU-Net, the mean Dice Coefficient range of values is [0,65-0,71]. Generally, the higher patch size, the higher the DC. The 95% Hausdorff Distance range is [2,66-4,32] and the standard deviation correspondent range is [0,68-1,90]. There are two outliers (nnU-Net_96_Dice_DA, nnU-Net_64_Dice_DA) with a 95% HD respectively of $(21,53 \pm 47,65)$ and $(63,57 \pm 62,99)$. Sorting by mean Percentage Volume Difference, the range of values is [7,40-17,46], with standard deviation range [5,37-11,63].

With respect to the UNETR, the mean Dice Coefficient range of values is [0,66-0,70]. The mean 95% Hausdorff Distance range is [2,51-4,03] and the standard deviation correspondent range is [0,41-2,86]. There are two outliers (UNETR_96_Dice_DA, UNETR_64_Dice_DA) with a 95% HD respectively of $(20,09 \pm 42,87)$ and $(100,48 \pm 68,38)$. Sorting by mean Percentage Volume Difference, the range of values is [8,47-14,65], with standard deviation range [4,11-10,83]. The higher is the patch size and in presence of DiceCE loss function, the lower is the Percentage Volume Difference.

4.5.2 Input FLAIR, GT FLAIR – reference cT1-w

Sorting by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN type, followed by 3D U-Net, as showed in *Table 4.24*. The Dice and DiceCE loss functions improve the results.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_Dice_DA	0,72	0,05	2,10	1,05	23,22	12,02
nnU-Net_64_DiceCE_DA	0,72	0,05	2,10	1,19	17,50	11,64
nnU-Net_64_CE_DA	0,72	0,05	2,16	1,04	18,16	11,76
nnU-Net_64_CE_noDA	0,72	0,05	1,91	0,97	22,63	12,47
nnU-Net_96_Dice_noDA	0,72	0,05	1,96	0,49	22,20	10,81
3DUNET_96_DiceCE_DA	0,72	0,05	2,19	0,94	19,79	11,84
nnU-Net_128_Dice_DA	0,72	0,05	2,27	0,84	28,60	12,48
nnU-Net_64_Dice_noDA	0,72	0,05	1,96	0,52	25,44	11,57
3DUNET_64_DiceCE_DA	0,71	0,05	2,14	0,98	16,75	10,98
nnU-Net_96_CE_noDA	0,71	0,05	1,89	0,52	19,18	9,84

Table 4.24: Best 10 DNNs sorted by mean Dice Coefficient.

4. Results

Ordering by 95% Hausdorff Distance (95% HD mean), nnU-Net is the most performant DNN type as showed in *Table 4.25*. However, there is no difference between the first ten DNNs in terms of 95% HD.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_CE_noDA	0,71	0,05	1,89	0,52	19,18	9,84
nnU-Net_128_Dice_noDA	0,71	0,05	1,89	0,59	24,76	11,18
3DUNET_64_CE_DA	0,71	0,05	1,91	0,60	20,89	10,77
nnU-Net_128_DiceCE_noDA	0,71	0,04	1,91	0,47	26,18	14,06
nnU-Net_64_CE_noDA	0,72	0,05	1,91	0,97	22,63	12,47
nnU-Net_64_Dice_noDA	0,72	0,05	1,96	0,52	25,44	11,57
nnU-Net_96_Dice_noDA	0,72	0,05	1,96	0,49	22,20	10,81
3DUNET_64_Dice_noDA	0,71	0,05	1,97	0,59	31,85	12,34
nnU-Net_96_wCE_noDA	0,69	0,04	1,97	0,50	45,38	12,34
3DUNET_128_DiceCE_DA	0,71	0,05	1,97	0,78	23,99	10,87

Table 4.25: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

Sorting by mean Percentage Volume Difference (% Vol Diff mean), nnU-Net and UNETR are the best DNNs type as shown in *Table 4.26*. The CE and DiceCE loss functions improve the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_DiceCE_DA	0,70	0,06	3,54	2,63	11,78	9,73
UNETR_64_CE_noDA	0,69	0,06	2,91	2,10	14,65	10,78
UNETR_96_CE_DA	0,70	0,06	3,17	2,56	14,88	12,17
UNETR_64_DiceCE_DA	0,69	0,06	2,33	0,83	15,44	11,74
UNETR_64_Dice_DA	0,69	0,06	17,32	29,77	15,80	11,03
3DUNET_64_DiceCE_DA	0,71	0,05	2,14	0,98	16,75	10,98
3DUNET_64_Dice_DA	0,70	0,05	21,13	37,91	17,10	14,25
nnU-Net_64_DiceCE_DA	0,72	0,05	2,10	1,19	17,50	11,64
UNETR_128_CE_noDA	0,68	0,06	2,50	0,98	18,00	7,98
nnU-Net_64_CE_DA	0,72	0,05	2,16	1,04	18,16	11,76

Table 4.26: Best 10 DNNs sorted by mean Percentage Volume Difference.

Generally, there is not a specific trend looking at the presence of Data Augmentation, while the Weighted Cross-Entropy loss function lowers the performance.

Considering all the combinations for each DNN type, the results highlight the following. The complete tables are showed in Appendix B – B.5.2.

Generally, there is no difference in terms of standard deviation for the DC values, while the wCE lower the performance, in particular for the Percentage Volume Difference.

For the 3D U-Net, the mean Dice Coefficient range of values is [0,64-0,72]. Generally, the CE and DiceCE loss functions the higher ones. The mean 95% Hausdorff Distance range is [1,91-2,77] and the standard deviation correspondent range is [0,59-3,01]. There

are two outliers (3DUNET_96_Dice_DA, 3DUNET_64_Dice_DA) with a 95% HD respectively of $(20,72 \pm 47,59)$ and $(21,13 \pm 37,91)$. Sorting by mean Percentage Volume Difference, only two combinations have mean values under the 18% (3DUNET_64_DiceCE_DA, 3DUNET_64_Dice_DA).

Concerning the nnU-Net, the mean Dice Coefficient range of values is [0,65-0,72]. The mean 95% Hausdorff Distance range is [1,89-3,54] and the standard deviation correspondent range is [0,49-2,63]. There are two outliers (nnU-Net_64_Dice_DA, nnU-Net_128_CE_DA) with a 95% HD respectively of $(15,76 \pm 34,50)$ and $(25,40 \pm 46,24)$. Sorting by mean Percentage Volume Difference, only two combinations have a mean value under the 18% (nnU-Net_96_DiceCE_DA, nnU-Net_64_DiceCE_DA).

With reference to the UNETR, the mean Dice Coefficient range of values across all the 24 possible combination is [0,65-0,71]. Looking at the mean 95% Hausdorff Distance, the range is [2,10-3,17] and the standard deviation correspondent range is [0,63-2,56]. There is only one outlier (UNETR_64_Dice_DA) with a 95% HD of $(17,32 \pm 29,77)$. Sorting by mean Percentage Volume Difference, only five combinations have a mean value under the 18%.

4.5.3 Input T1-w+FLAIR, GT T1-w – reference cT1-w

Sorting by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN type, followed by UNETR, as showed in *Table 4.27*. Generally, the higher the patch size, the higher the mean DC values.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_CE_DA	0,71	0,05	2,50	0,55	8,98	6,74
nnU-Net_64_CE_DA	0,71	0,04	13,65	41,58	9,38	6,85
nnU-Net_96_Dice_DA	0,70	0,05	3,46	2,41	10,95	8,73
nnU-Net_128_DiceCE_DA	0,70	0,04	3,02	1,73	9,34	6,97
UNETR_96_DiceCE_DA	0,70	0,04	2,17	0,50	9,67	6,56
nnU-Net_128_Dice_noDA	0,70	0,04	3,22	1,31	10,30	6,46
UNETR_96_Dice_DA	0,70	0,05	2,66	0,76	9,98	7,18
nnU-Net_96_CE_DA	0,70	0,05	2,49	0,65	8,55	6,36
3DUNET_64_Dice_DA	0,70	0,04	2,82	0,94	12,31	9,21
nnU-Net_128_Dice_DA	0,70	0,04	2,83	0,72	12,32	9,84

Table 4.27: Best 10 DNNs sorted by mean Dice Coefficient.

Ordering by 95% Hausdorff Distance (95% HD mean), UNETR is the most performant DNN type, followed by nnU-Net, as showed in *Table 4.28*. However, there is no difference between the first ten DNNs in terms of 95% HD.

4. Results

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
UNETR_96_DiceCE_DA	0,70	0,04	2,17	0,50	9,67	6,56
nnU-Net_96_wCE_noDA	0,69	0,04	2,37	0,85	20,12	13,30
UNETR_96_CE_DA	0,69	0,05	2,48	0,37	8,68	7,74
nnU-Net_96_CE_DA	0,70	0,05	2,49	0,65	8,55	6,36
nnU-Net_64_DiceCE_noDA	0,70	0,04	2,49	0,77	9,95	9,23
nnU-Net_128_CE_DA	0,71	0,05	2,50	0,55	8,98	6,74
3DUNET_64_wCE_DA	0,66	0,04	2,55	0,70	51,62	19,82
nnU-Net_64_CE_noDA	0,69	0,04	2,60	0,73	11,73	8,88
UNETR_64_DiceCE_DA	0,69	0,05	2,64	0,62	8,75	7,03
UNETR_96_wCE_DA	0,68	0,04	2,65	0,74	42,48	16,92

Table 4.28: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

Sorting by mean Percentage Volume Difference (% Vol Diff mean), nnU-Net and UNETR are the best DNNs type as shown in Table 4.29. The higher patch size improves the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_Dice_noDA	0,70	0,05	2,91	1,40	8,34	7,78
nnU-Net_128_CE_noDA	0,69	0,04	2,80	0,97	8,47	7,47
nnU-Net_96_CE_DA	0,70	0,05	2,49	0,65	8,55	6,36
UNETR_96_CE_DA	0,69	0,05	2,48	0,37	8,68	7,74
UNETR_64_DiceCE_DA	0,69	0,05	2,64	0,62	8,75	7,03
UNETR_64_DiceCE_noDA	0,69	0,05	3,85	2,61	8,94	7,29
3DUNET_96_DiceCE_DA	0,69	0,04	3,97	2,42	8,96	8,18
nnU-Net_128_CE_DA	0,71	0,05	2,50	0,55	8,98	6,74
3DUNET_96_Dice_DA	0,69	0,05	3,07	0,93	9,21	6,72
3DUNET_128_Dice_DA	0,69	0,04	3,45	1,69	9,32	7,48

Table 4.29: Best 10 DNNs sorted by mean Percentage Volume Difference.

Generally, there is not a specific trend looking at the presence of Data Augmentation, while the Weighted Cross-Entropy loss function lowers the performance

Considering all the combinations for each DNN type, the results highlight the following. The complete tables are showed in Appendix B – B.5.3. The Weighted Cross-Entropy loss function gives the worst results for all performance indices.

For the 3D U-Net, the mean Dice Coefficient range of values is [0,66-0,70]. Generally, the Dice loss function the higher ones. Looking at the mean 95% Hausdorff Distance, the range is [2,55-4,08] and the standard deviation correspondent range is [0,70-2,56]. There is one outlier (3DUNET_64_DiceCE_DA) with a 95% HD of $(31,09 \pm 54,28)$. The Weighted Cross-Entropy loss function gives the better results in term of 95% HD. Sorting by mean Percentage Volume Difference, the range of values is [8,96-12,71], with

standard deviation range [6,72-10,24]. Generally, the higher the patch size, the lower the Percentage Volume Difference is.

Concerning the nnU-Net, the mean Dice Coefficient range of values is [0,66-0,71]. Looking at the mean 95% Hausdorff Distance, the range is [2,37-8,35] and the standard deviation correspondent range is [0,55-2,41]. There are four outliers (nnU-Net_64_CE_DA, nnU-Net_64_wCE_DA, nnU-Net_64_Dice_DA, nnU-Net_64_Dice_noDA). Sorting by mean Percentage Volume Difference, the range of values is [8,34-15,83], with standard deviation range [6,36-12,77]. Generally, the higher the patch size, the lower the Percentage Volume Difference is.

Regarding the UNETR, the mean Dice Coefficient range of values is [0,66-0,70]. The Dice loss function gives a higher mean DC value. Looking at the mean 95% Hausdorff Distance the range is [2,17-4,24] and the standard deviation correspondent range is [0,37-2,61]. There is only one outlier (UNETR_64_Dice_DA) with a 95% HD of (35,53 ± 58,29). Sorting by mean Percentage Volume Difference, the range of values is [8,68-13,51], with standard deviation range [7,03-12,30]. The DiceCE and CE loss functions gives lower Percentage Volume Difference values.

4.5.4 Input T1-w+FLAIR, GT FLAIR – reference cT1-w

Sorting by Dice Coefficient (Dice mean), the nnU-Net is the most performant DNN architecture, followed by 3D U-Net, as showed in *Table 4.30*. Generally, the higher the patch size, the higher the mean DC values.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_96_Dice_noDA	0,72	0,05	1,82	0,69	22,58	11,53
3DUNET_96_Dice_DA	0,72	0,05	10,04	30,75	25,25	13,17
nnU-Net_128_DiceCE_DA	0,72	0,05	1,80	0,54	19,97	13,38
3DUNET_96_Dice_noDA	0,72	0,05	1,82	0,52	18,29	9,93
nnU-Net_128_CE_DA	0,72	0,05	1,73	0,58	20,12	11,59
nnU-Net_96_DiceCE_DA	0,72	0,05	1,84	0,57	17,17	11,80
nnU-Net_96_Dice_DA	0,72	0,06	15,56	36,66	24,86	14,40
3DUNET_64_Dice_noDA	0,72	0,05	1,85	0,55	22,99	11,62
nnU-Net_128_Dice_DA	0,72	0,05	1,75	0,56	24,32	15,73
3DUNET_128_CE_DA	0,72	0,06	1,85	0,55	18,71	13,06

Table 4.30: Best 10 DNNs sorted by mean Dice Coefficient.

Ordering by 95% Hausdorff Distance (95% HD mean), nnU-Net is the most performant DNN type, followed by 3D U-net, as showed in *Table 4.31*. However, there is no

4. Results

difference between the first ten DNNs in terms of 95% HD. The dice and DiceCE loss functions improve the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
nnU-Net_128_CE_DA	0,72	0,05	1,73	0,58	20,12	11,59
nnU-Net_128_Dice_DA	0,72	0,05	1,75	0,56	24,32	15,73
nnU-Net_128_DiceCE_DA	0,72	0,05	1,80	0,54	19,97	13,38
nnU-Net_128_Dice_noDA	0,72	0,05	1,80	0,57	20,12	15,10
nnU-Net_96_Dice_noDA	0,72	0,05	1,82	0,69	22,58	11,53
3DUNET_96_Dice_noDA	0,72	0,05	1,82	0,52	18,29	9,93
3DUNET_96_DiceCE_DA	0,72	0,05	1,83	0,57	28,68	12,71
nnU-Net_96_DiceCE_DA	0,72	0,05	1,84	0,57	17,17	11,80
3DUNET_128_CE_DA	0,72	0,06	1,85	0,55	18,71	13,06
3DUNET_64_Dice_noDA	0,72	0,05	1,85	0,55	22,99	11,62

Table 4.31: Best 10 DNNs sorted by mean 95% Hausdorff Distance.

Sorting by mean Percentage Volume Difference (% Vol Diff mean), nnU-Net and 3D U-Net are the best DNNs type as shown in Table 4.32. The Dice and DiceCE improve them. The higher patch size improves the performance.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd
3DUNET_128_Dice_noDA	0,71	0,05	1,94	0,50	15,51	10,61
nnU-Net_96_DiceCE_DA	0,72	0,05	1,84	0,57	17,17	11,80
nnU-Net_64_DiceCE_DA	0,70	0,06	12,62	35,99	17,28	13,58
3DUNET_96_Dice_noDA	0,72	0,05	1,82	0,52	18,29	9,93
3DUNET_128_CE_DA	0,72	0,06	1,85	0,55	18,71	13,06
nnU-Net_96_DiceCE_noDA	0,71	0,06	2,29	0,76	18,81	10,37
3DUNET_128_DiceCE_noDA	0,71	0,05	1,94	0,47	19,04	9,53
UNETR_128_Dice_noDA	0,70	0,06	1,99	0,60	19,09	10,27
3DUNET_96_CE_noDA	0,71	0,06	2,42	1,88	19,32	10,52
UNETR_128_CE_DA	0,71	0,05	2,07	0,63	19,65	13,15

Table 4.32: Best 10 DNNs sorted by mean Percentage Volume Difference.

Considering all the combinations for each DNN type, the results highlight the following. The complete tables are showed in Appendix B – B.5.4. As for the other combinations, the wCE lowers the performances and there are no trends for the Data Augmentation application.

For the 3D U-Net, the mean Dice Coefficient range of values is [0,64-0,72]. Generally, the Dice loss function the higher ones. Looking at the mean 95% Hausdorff Distance the range is [1,82-2,70] and the standard deviation correspondent range is [0,43-1,93]. There are two outliers (3DUNET_96_Dice_DA, 3DUNET_64_DiceCE_DA) with a 95% HD respectively of $(10,04 \pm 30,75)$ and $(12,69 \pm 37,99)$. Generally, the Dice loss function

gives the better results in term of 95% HD. Sorting by mean Percentage Volume Difference, only 3DUNET_128_Dice_noDA has a value under the 18%. Generally, the higher the patch size, the lower the Percentage Volume Difference is.

Concerning the nnU-Net, the mean Dice Coefficient range of values is [0,64-0,72]. The higher the patch size, the higher the mean DC values. Looking at the mean 95% Hausdorff Distance the range is [1,73-2,49] and the standard deviation correspondent range is [0,54-1,09]. There are seven outliers, mostly with Dice loss function. Sorting by mean Percentage Volume Difference, only two combinations (nnU-Net_96_DiceCE_DA, nnU-Net_64_DiceCE_DA) have values under the 18%. Generally, the DiceCE loss function lowers the Percentage Volume Difference.

Regarding the UNETR, the mean Dice Coefficient range of values is [0,66-0,72]. The Dice and DiceCE loss functions give a higher mean DC value. Looking at the mean 95% Hausdorff Distance, the range is [1,86-2,56] and the standard deviation correspondent range is [0,55-1,67]. There are two outliers (UNETR_96_DiceCE_DA, UNETR_64_Dice_DA) with a 95% HD respectively of $(10,02 \pm 30,70)$ and $(119,52 \pm 50,75)$. No combination has Percentage Volume Difference lower than 18%.

4.6 PERFORMANCE ANALYSIS: COMPARISON BETWEEN MODELS TRAINED WITH GT cT1-w AND THOSE WITH GT WITHOUT CONTRAST (T1-w, FLAIR)

The aim of this paragraph is to make a comparison between the performance of the DNNs trained with GT cT1-w manual segmentation and those trained with GT without contrast (T1-w, FLAIR) manual segmentation. As showed in the previous paragraphs, the performance indices for the DNNs trained with GT without contrast were calculated with respect to the cT1-w manual segmentation (gold-standard).

The comparison was made considering the correspondent performance combinations individually and with respect to the others:

- Input T1-w, GT T1-w compared with Input T1-w, GT cT1-w
- Input FLAIR, GT FLAIR compared with Input FLAIR, GT cT1-w
- Input T1-w+FLAIR, GT T1-w and Input T1-w+FLAIR, GT FLAIR compared with Input T1-w+FLAIR, GT cT1-w

The results in the Tables are showed considering the ten best combination of training parameters sorted by mean Percentage Volume Difference, that is the possible biomarker.

The ten best nets were then sorted by the $m+2*sd$ and $m+3*sd$ values, that represent respectively the 95% and the 99% probabilities of correctly segment the ChP in all subjects of the validation set, keeping below the discriminating threshold (21,4 %) selected by the literature (Müller et al., 2022) between healthy control patients and patients with MS, as commented in the Discussion chapter (paragraph 5.4). The reported performance indices are the Dice Coefficient and the Percentage Volume Difference.

4.6.1 Input T1-w, GT T1-w compared with Input T1-w, GT cT1-w

The *Table 4.33* below shows that the best DNN to be used both with and without contrast is the UNETR. Moreover, there is no statistical difference between UNETR and nnU-Net. The best configuration to segment the ChP in presence of GT cT1-w is the UNETR with patch size 128, trained with Dice loss function and Data Augmentation transforms applied to the training dataset. The second-best configuration is the UNETR with patch size 96, trained with Cross-Entropy loss function and Data Augmentation transforms applied to the training dataset. This configuration is also the best configuration to segment the ChP in absence of contrast information. Looking at the training with contrast, nine configurations can be selected to correctly segmenting the ChP across all subjects with a probability of 95%, while only the first two UNETR are able to do this with a probability of 99%. Looking at the training without contrast, the results have the same trend. The Dice Coefficient is performant for both terms of comparison.

Training input T1-w, GT cT1-w, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	$m+2*sd$	$m+3*sd$
UNETR			x	x			x		0,75	0,04	7,54	3,79	15,11	18,90
UNETR		x			x		x		0,75	0,04	8,30	3,83	15,96	19,79
nnU-Net			x		x		x		0,76	0,04	8,00	5,07	18,15	23,22
nnU-Net		x			x		x		0,76	0,04	7,94	5,20	18,34	23,54
UNETR			x		x		x		0,74	0,04	8,98	4,78	18,54	23,32
nnU-Net			x			x	x		0,76	0,04	9,04	4,87	18,79	23,66
nnU-Net	x				x		x		0,76	0,04	8,58	5,26	19,09	24,35
3D U-Net	x					x	x		0,73	0,04	8,74	5,98	20,70	26,68
nnU-Net		x				x	x		0,77	0,03	7,86	6,60	21,06	27,66
UNETR			x			x		x	0,73	0,05	8,72	6,69	22,11	28,80
MEAN									0,75	0,04	8,37	5,21		

4. Results

Training input T1-w, GT T1-w, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	m+2*sd	m+3*sd
UNETR		x			x		x		0,70	0,04	8,47	4,11	16,68	20,78
nnU-Net			x		x		x		0,70	0,04	7,93	5,43	18,79	24,22
nnU-Net		x			x		x		0,71	0,04	7,40	5,91	19,23	25,14
nnU-Net	x				x			x	0,69	0,04	8,44	5,63	19,70	25,33
UNETR		x				x		x	0,69	0,05	8,47	5,85	20,16	26,01
3D U-Net	x					x	x		0,67	0,04	8,22	6,03	20,29	26,32
nnU-Net		x				x		x	0,70	0,04	8,29	6,24	20,78	27,02
nnU-Net		x		x				x	0,70	0,04	8,68	6,60	21,87	28,47
UNETR		x				x	x		0,70	0,04	8,61	6,67	21,95	28,62
UNETR	x				x		x		0,70	0,05	8,61	6,85	22,31	29,16
MEAN									0,70	0,04	8,31	5,93		

Table 4.33: Input T1-w, GT T1-w compared with Input T1-w, GT cT1-w. The nets are sorted first by Percentage Volume Difference, then for the m+2*sd (and m+3*sd as consequence).

4.6.2 Input FLAIR, GT FLAIR compared with Input FLAIR, GT cT1-w

As for the T1-w input images, the Table 4.34 below shows how the most performant nets are again UNETR and nnU-Net. However, although for the DNNs trained with contrast it is possible to select a net with a 95% probability of correctly segmenting the ChP, that is again a UNETR, this is not possible for the DNNs trained without contrast. Even if the mean Dice Coefficient is not different from that obtained with the T1-w input, the Percentage Volume Difference is higher for the best-case sequence.

Training input FLAIR, GT cT1-w, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	m+2*sd	m+3*sd
UNETR			x			x	x		0,72	0,06	9,92	5,45	20,81	26,26
3D U-Net	x					x	x		0,73	0,05	9,40	6,08	21,55	27,63
nnU-Net	x			x				x	0,75	0,05	10,13	7,05	24,23	31,28
nnU-Net	x			x			x		0,74	0,05	9,75	7,32	24,39	31,72
nnU-Net	x				x		x		0,75	0,05	8,66	7,91	24,49	32,40
nnU-Net		x		x			x		0,75	0,05	9,41	7,66	24,73	32,40
UNETR			x	x			x		0,73	0,05	9,95	7,50	24,95	32,45
nnU-Net			x	x				x	0,75	0,05	9,92	7,58	25,08	32,66
nnU-Net		x		x				x	0,75	0,05	10,22	8,21	26,64	34,84
3D U-Net	x			x			x		0,73	0,04	10,28	8,67	27,61	36,28
MEAN									0,74	0,05	9,76	7,34		

4. Results

Training input FLAIR, GT FLAIR, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	m+2*sd	m+3*sd
nnU-Net		x				x	x		0,70	0,06	11,78	9,73	31,25	40,98
UNETR			x		x			x	0,68	0,06	18,00	7,98	33,96	41,94
UNETR	x				x			x	0,69	0,06	14,65	10,78	36,21	47,00
UNETR	x			x			x		0,69	0,06	15,80	11,03	37,87	48,90
3D U-Net	x					x	x		0,71	0,05	16,75	10,98	38,71	49,68
UNETR	x					x	x		0,69	0,06	15,44	11,74	38,92	50,66
UNETR		x			x		x		0,70	0,06	14,88	12,17	39,21	51,38
nnU-Net	x					x	x		0,72	0,05	17,50	11,64	40,79	52,44
nnU-Net	x				x		x		0,72	0,05	18,16	11,76	41,69	53,45
3D U-Net	x			x			x		0,70	0,05	17,10	14,25	45,60	59,85
MEAN									0,70	0,06	16,01	11,21		

Table 4.34: Input FLAIR, GT FLAIR compared with Input FLAIR, GT cT1-w. The nets are sorted first by Percentage Volume Difference, then for the m+2*sd (and m+3*sd as consequence).

4.6.3 Input T1-w+FLAIR, GT T1-w and Input T1-w+FLAIR, GT FLAIR compared with Input T1-w+FLAIR, GT cT1-w

For what concern the two inputs analysis, the Table 4.35 below shows how once again the UNETR and nnU-Net are the best DNNs to lower the Percentage Volume Difference.

Training input T1-w+FLAIR, GT cT1-w, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	m+2*sd	m+3*sd
UNETR			x			x	x		0,75	0,05	8,34	5,33	19,01	24,34
nnU-Net	x				x			x	0,76	0,04	8,75	5,22	19,20	24,42
UNETR			x	x			x		0,74	0,05	9,25	5,06	19,37	24,43
nnU-Net			x			x	x		0,76	0,04	8,23	5,81	19,85	25,65
nnU-Net			x	x				x	0,76	0,04	8,34	5,92	20,17	26,09
nnU-Net		x				x		x	0,75	0,05	8,47	6,23	20,93	27,16
nnU-Net	x				x		x		0,76	0,05	9,37	5,87	21,11	26,97
nnU-Net		x		x			x		0,77	0,04	8,62	6,33	21,28	27,61
UNETR		x				x		x	0,74	0,05	9,39	6,22	21,84	28,06
nnU-Net			x	x			x		0,76	0,04	9,08	6,75	22,59	29,34
MEAN									0,76	0,05	8,78	5,87		

Training input T1-w+FLAIR, GT T1-w, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	m+2*sd	m+3*sd
nnU-Net		x			x		x		0,70	0,05	8,55	6,36	21,27	27,63
nnU-Net			x		x		x		0,71	0,05	8,98	6,74	22,46	29,20
3D U-Net		x		x			x		0,69	0,05	9,21	6,72	22,65	29,36
UNETR	x					x	x		0,69	0,05	8,75	7,03	22,82	29,85
nnU-Net			x		x			x	0,69	0,04	8,47	7,47	23,40	30,87
UNETR	x					x		x	0,69	0,05	8,94	7,29	23,53	30,82
nnU-Net		x		x				x	0,70	0,05	8,34	7,78	23,90	31,68

4. Results

UNETR		x			x		x		0,69	0,05	8,68	7,74	24,16	31,91
3D U-Net			x	x			x		0,69	0,04	9,32	7,48	24,28	31,76
3D U-Net		x				x	x		0,69	0,04	8,96	8,18	25,33	33,51
MEAN									0,69	0,05	8,82	7,28		

Training input T1-w+FLAIR, GT FLAIR, reference cT1-w

Deep Neural Network	Patch size			Loss Function			DA		Performance Indices				< 21,4 %	
	64	96	128	Dice	CE	DiceCE	Yes	No	Dice mean	Dice sd	% Vol Diff mean	% Vol Diff sd	m+2*sd	m+3*sd
3D U-Net			x	x				x	0,71	0,05	15,51	10,61	36,73	47,35
3D U-Net			x			x		x	0,71	0,05	19,04	9,53	38,11	47,64
3D U-Net		x		x				x	0,72	0,05	18,29	9,93	38,15	48,08
nnU-Net		x				x		x	0,71	0,06	18,81	10,37	39,56	49,93
UNETR			x	x				x	0,70	0,06	19,09	10,27	39,62	49,89
3D U-Net		x			x			x	0,71	0,06	19,32	10,52	40,36	50,87
nnU-Net		x				x	x		0,72	0,05	17,17	11,80	40,78	52,58
nnU-Net	x					x	x		0,70	0,06	17,28	13,58	44,44	58,03
3D U-Net			x		x		x		0,72	0,06	18,71	13,06	44,83	57,89
UNETR			x		x		x		0,71	0,05	19,65	13,15	45,95	59,10
MEAN									0,71	0,05	18,29	11,28		

Table 4.35: Input T1-w+FLAIR, GT T1-w and Input T1-w+FLAIR, GT FLAIR compared with Input T1-w+FLAIR, GT cT1-w. The nets are sorted first by Percentage Volume Difference, then for the m+2*sd (and m+3*sd as consequence).

However, no combination can be selected for the training without contrast, both with GT T1-w (borderline) and GT FLAIR. Moreover, using GT FLAIR improves the Dice Coefficient but gives worst results for what concern the Percentage Volume Difference, that is the biomarker.

The performance analysis led to consider the UNETR and the nnU-Net as the most performant DNNs with T1-w input to improve the performance of the Percentage Volume Difference index.

4.7 COMPARISON WITH THE STATE-OF-THE-ART AUTOMATIC METHODS FS AND GMM

As explained in the discussion (par 5.4), the previous analysis led to consider the following DNNs and the following input – GT MSeg combinations for comparison with state-of-the-art automatic methods: UNETR with patch size 128x128x128, Dice loss function and Data Augmentation for the input T1-w with GT cT1-w; UNETR with patch size 96x96x96, Cross-Entropy loss function and Data Augmentation for the input T1-w with GT T1-w.

The comparison between the automatic segmentation obtained with FS, GMM, and the two UNETR trained with GT with and without contrast T1-w MSeg is performed in

MATLAB. The comparison considers also the MSeg obtained on T1-w images to analyze how much better or worse than the GT the DNNs predict the segmentation of the ChP.

Dice Coefficient: reference cT1-w

The compared method with the higher mean Dice Coefficient is the DNN trained on T1-w images, using GT cT1-w MSeg. However, the DNN trained over T1-w images, using GT T1-w MSeg, gives a DC higher than those obtained with state-of-the-art automatic methods (FS, GMM) or the T1-w MSeg too. The *Table 4.36* below shows the DC values for each subject and for each method when compared to the cT1-w sequence, considered as the Gold Standard ground truth. *Figure 4.11* reports boxplot of the same quantities reported in *Table 4.36* showing the variability for each compared method.

SUBJECTS	T1-w	FS	GMM	Contrast UNETR 128, Dice, DA	No contrast UNETR 96, CE, DA
30	0,681	0,324	0,590	0,774	0,733
48	0,671	0,423	0,542	0,769	0,711
66	0,566	0,237	0,438	0,645	0,606
1923	0,687	0,287	0,552	0,735	0,722
1964	0,666	0,415	0,515	0,778	0,732
1984	0,687	0,345	0,525	0,782	0,746
1985	0,598	0,250	0,503	0,697	0,635
2043	0,654	0,385	0,539	0,754	0,697
2045	0,705	0,271	0,330	0,756	0,712
2050	0,653	0,209	0,472	0,746	0,710
2056	0,712	0,271	0,561	0,812	0,734
2060	0,722	0,242	0,489	0,816	0,771
2113	0,634	0,243	0,393	0,716	0,653
2118	0,708	0,374	0,485	0,754	0,703
2139	0,622	0,279	0,444	0,701	0,665
MEAN	0,664	0,304	0,492	0,749	0,702
SD	0,044	0,069	0,069	0,045	0,045

Table 4.36: Dice Coefficient between respectively the T1-w manual segmentation, the automatic method proposed by the literature (FS, GMM) and the proposed DNNs trained with contrast (Contrast UNETR, 128, Dice, DA) and without contrast (No contrast UNETR, 96, CE, DA), and the reference gold-standard cT1-w sequence. The performance index was estimated for each subject of the validation set. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

4. Results

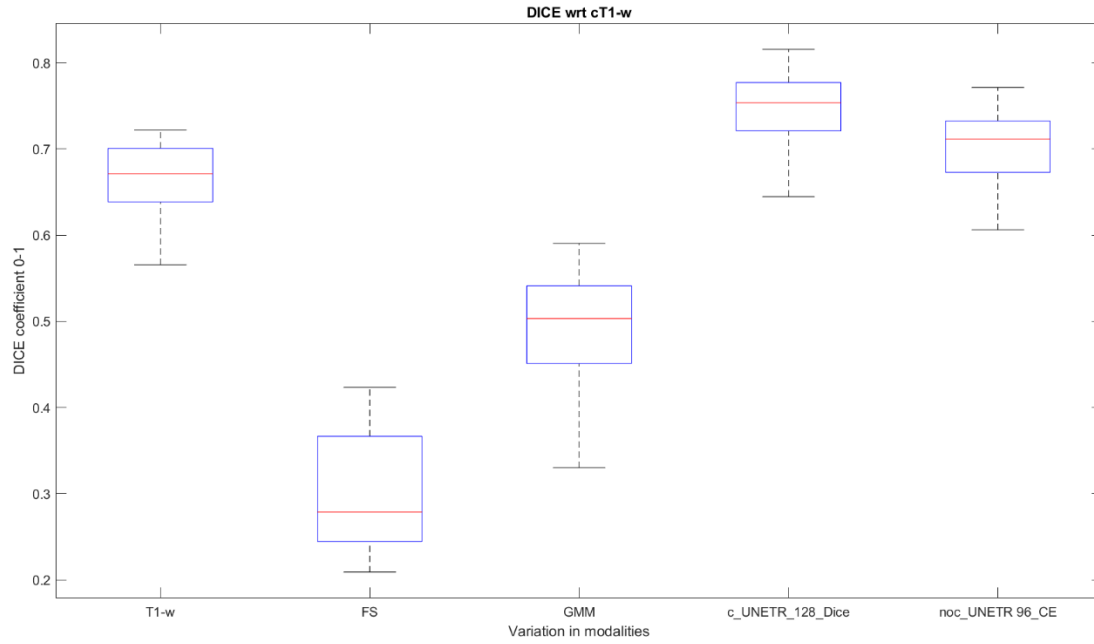


Figure 4.11: Boxplot of the Dice Coefficient median and variation intra-modality, calculated between each compared method (T1-w MSeg, FS, GMM, contrast UNETR_128_Dice, no contrast UNETR_96_CE) and the gold-standard cT1-w sequence.

The Figure 4.12 below shows the outlier subjects for each term of comparison. FS and GMM have the higher variability inside the dataset.

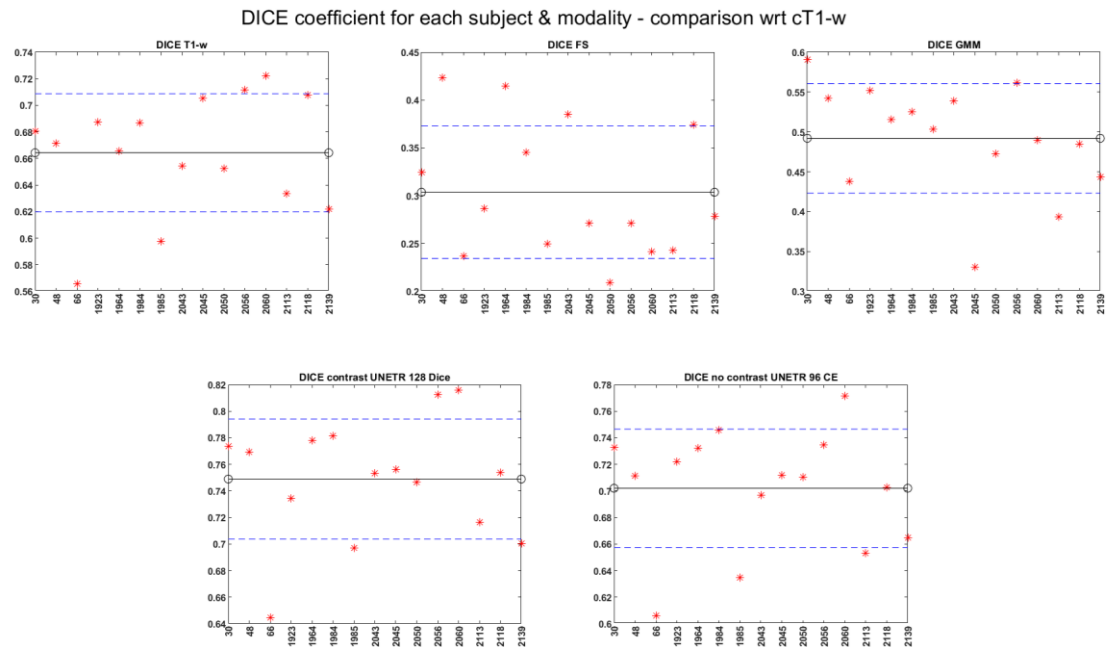


Figure 4.12: Plot of Dice Coefficient calculated for each patient between each compared method (T1-w MSeg, FS, GMM, the proposed DNNs) and the gold-standard cT1-w sequence. Mean and standard deviation of each modality are reported in the plots.

Hausdorff Distance: reference cT1-w

For what concern the 95% Hausdorff Distance, the contrast UNETR_128_Dice gives the best values in terms of mean value and variability between subjects, the SD obtained for this DNN is contaminated by only one outlier, while interquartile ranges provided by the boxplot shows a narrower interval (*Table 4.37, Figure 4.13*). FS and GMM segmentations give worse results with respect to both T1-w MSeg and the two proposed DNNs.

SUBJECTS	T1-w	FS	GMM	Contrast UNETR 128, Dice, DA	No contrast UNETR 96, CE, DA
30	3,74	12,49	12,34	2,24	6,06
48	2,24	6,01	7,87	2,00	2,24
66	4,12	8,06	10,20	2,83	3,57
1923	6,48	6,16	14,03	2,83	3,26
1964	4,69	10,05	14,12	2,24	3,16
1984	3,32	6,40	23,11	2,24	2,45
1985	3,61	9,11	8,49	3,00	5,32
2043	3,32	7,04	8,35	1,41	4,12
2045	2,24	7,01	7,28	3,74	9,70
2050	12,37	5,39	9,11	2,00	3,00
2056	2,45	7,00	12,71	2,00	3,16
2060	3,74	7,48	20,32	1,73	2,24
2113	2,45	9,68	17,24	2,24	3,00
2118	2,24	5,20	13,74	2,00	5,16
2139	2,45	6,78	12,04	2,45	2,45
MEAN	3,96	7,59	12,73	2,33	3,93
SD	2,60	1,98	4,64	0,57	1,98

Table 4.37: 95% Hausdorff Distance between respectively the T1-w manual segmentation, the automatic method proposed by the literature (FS, GMM) and the proposed DNNs trained with contrast (Contrast UNETR, 128, Dice, DA) and without contrast (No contrast UNETR, 96, CE, DA), and the reference gold-standard cT1-w sequence. The performance index was estimated for each subject of the validation set. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

4. Results

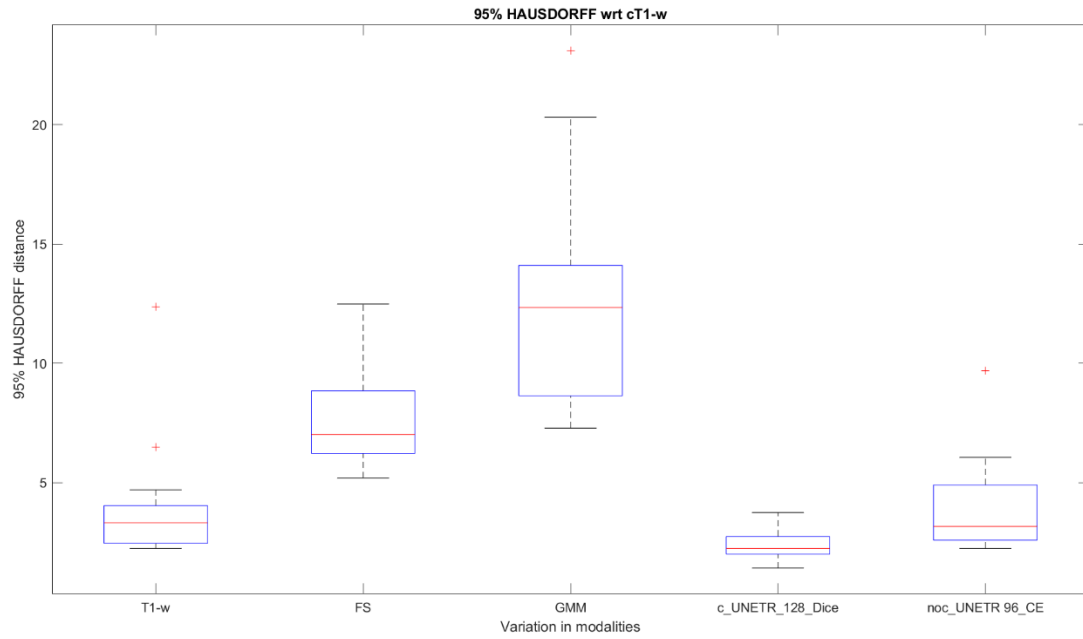


Figure 4.13: Boxplot of the 95% Hausdorff Distance median and variation intra-modality, calculated between each compared method (T1-w MSeg, FS, GMM, contrast UNETR_128_Dice, no contrast UNETR_96_CE) and the gold-standard cT1-w sequence.

The Figure 4.14 below shows the outlier subjects for each term of comparison. FS and GMM have the higher variability inside the dataset with respect to T1-w MSeg and the two proposed DNNs.

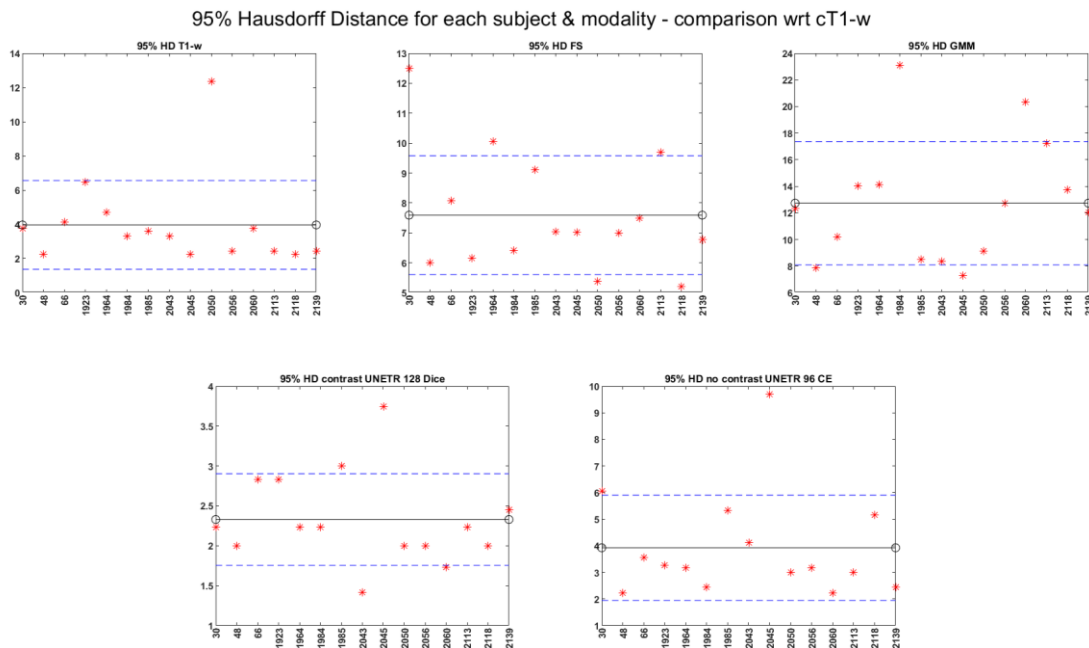


Figure 4.14: Plot of 95% Hausdorff Distance calculated for each patient between each compared method (T1-w MSeg, FS, GMM, the proposed DNNs) and the gold-standard cT1-w sequence. Mean and standard deviation of each modality are reported in the plots.

Volume Analysis: reference cT1-w

The volume analysis is the most significant because the possible biomarker is the volume estimation of the ChP that has to be as similar as possible to that extrapolated from the gold standard. The proposed DNN trained with contrast and the proposed DNN trained without contrast have the mean value closest to the gold-standard, followed by the T1-w MSeg (Table 4.38). On the contrary, FS has the farthest mean value to the gold-standard.

SUBJECTS	T1-w	FS	GMM	Contrast UNETR 128, Dice	No contrast UNETR 96, CE	cT1-w reference
30	3003	1414	3043	2835	2867	3027
48	2343	1327	2431	2466	2358	2570
66	2812	797	1920	2754	2483	2684
1923	2799	1425	1466	2701	2825	2486
1964	3458	1843	1602	3145	3088	3299
1984	4538	1514	1993	3923	3739	3711
1985	3478	1707	2698	3149	3133	2826
2043	2748	1464	2482	2396	2577	2241
2045	2620	1686	2171	2719	2710	2493
2050	2407	689	1380	2691	2582	2448
2056	3441	1085	2057	2801	2852	3044
2060	4007	1112	1627	3701	3779	4381
2113	2672	652	1597	2790	2692	2933
2118	2616	1467	1605	3273	2966	2882
2139	2534	708	1326	2943	2576	2995
MEAN	3031,73	1259,33	1959,87	2952,47	2881,80	2934,67
SD	628,85	395,77	517,39	423,90	416,79	546,17

Table 4.38: Segmentation Volume ($1 \text{ voxel} = 1 \text{ mm}^3$) calculated for each patient of the validation set for each compared method: the T1-w MSeg, the automatic method proposed by the literature (FS, GMM) and the proposed DNNs trained with contrast (Contrast UNETR, 128, Dice, DA) and without contrast (No contrast UNETR, 96, CE, DA), and the reference cT1-w MSeg. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

RMSE and MSE confirms the results in term of volume estimation, showing the contrast DNN minimum RMSE followed by no contrast proposed DNN and the T1-w MSeg. The state-of-the-art automated methods providing poorer reliability than the proposed DNNs approaches (Table 4.39).

4. Results

MODALITY	MSE	RMSE
T1-w MSeg	148301	385
FS	3222725	1795
GMM	1570243	1253
Contrast UNETR, 128, Dice	73562	271
No Contrast UNETR, 96, CE	78888	281

Table 4.39: RMSE and MSE for the T1-w MSeg, the state-of-the-art automatic segmentations (FS, GMM), and the segmentations of the proposed DNNs, with respect to the gold-standard cT1-w sequence.

Observing the OLS (linear regression without intercept) and the linear regression analyses (Figure 4.15), GMM and FS are the segmentations that least accurately estimate the volume of the gold-standard segmentation, while the proposed DNNs are the more accurate.

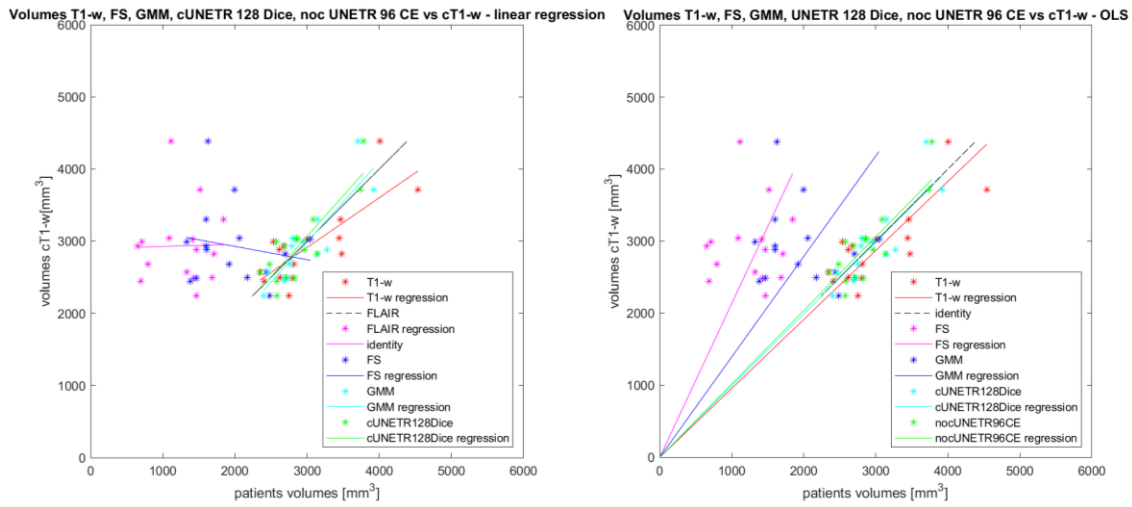


Figure 4.15: Panel left: Linear Regression of subject's volume for each modality; panel right: OLS volume representation for each modality.

For what concern Pearson's Correlation Analysis (Table 4.40, Figure 4.16), the proposed DNNs are the more correlated with cT1-w MSeg, followed by T1-w MSeg. FS and GMM have far lower correlation coefficients than the proposed DNNs or the T1-w MSeg.

P's CORR	T1-w MSeg	FS	GMM	cUNETR, 128, Dice	nocUNETR, 96, CE	cT1-w MSeg
T1-w MSeg	1	0,332	0,095	0,827	0,912	0,793
FS		1	0,461	0,200	0,346	0,021
GMM			1	-0,201	-0,056	-0,174
cUNETR, 128, Dice				1	0,928	0,863
nocUNETR, 96, CE					1	0,858
cT1-w MSeg						1

Table 4.40: Pearson's Correlation Analysis coefficients between each method (T1-w MSeg, FS, GMM, contrast UNETR_128_Dice, no contrast UNETR_96_CE) and the gold-standard cT1-w sequence. The significant correlation coefficients ($\alpha=0,05$) are highlighted: the others are not significant.

4. Results

Pearson Correlation Analysis

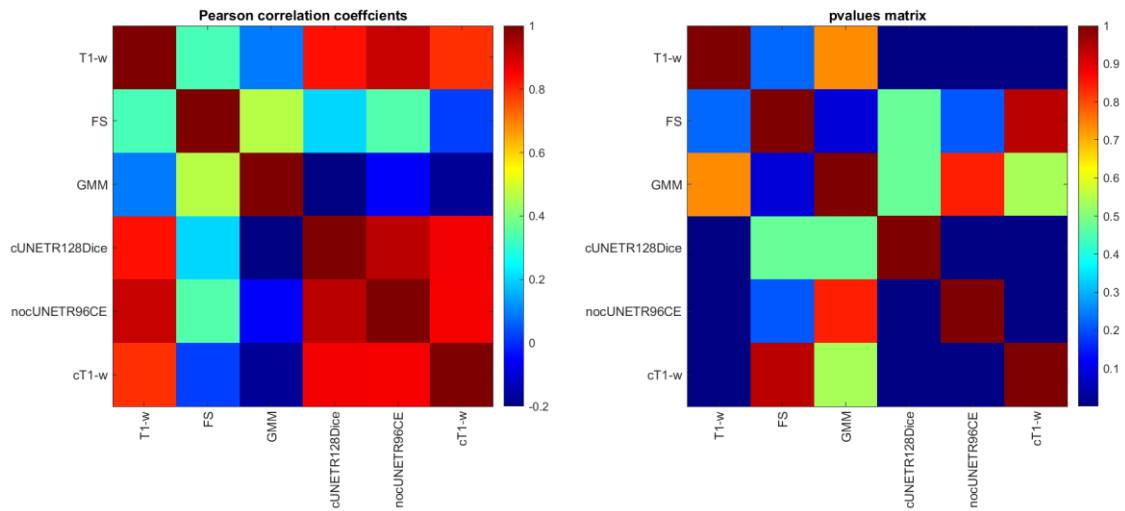


Figure 4.16: Left panel: Pearson's Correlation Coefficients; right panel: p-values matrix.

Percentage Volume Difference: reference cT1-w

In terms of Percentage Volume Difference (*Table 4.41*), that estimate the percentage discrepancy between the gold standard volume estimation and the compared method, the DNN trained with contrast is the method with the lower mean value and variability (*Figure 4.17*), followed by the no contrast DNN one and the T1-w MSeg. FS segmentation it is the method that gives the estimate of the volume furthest from the gold standard one.

SUBJECTS	T1-w	FS	GMM	Contrast UNETR 128, Dice, DA	No contrast UNETR 96, CE, DA
30	0,79	53,29	0,53	6,34	5,29
48	8,83	48,37	5,41	4,05	8,25
66	4,77	70,31	28,46	2,61	7,49
1923	12,59	42,68	41,03	8,65	13,64
1964	4,82	44,13	51,44	4,67	6,40
1984	22,29	59,20	46,29	5,71	0,75
1985	23,07	39,60	4,53	11,43	10,86
2043	22,62	34,67	10,75	6,92	14,99
2045	5,09	32,37	12,92	9,07	8,70
2050	1,67	71,85	43,63	9,93	5,47
2056	13,04	64,36	32,42	7,98	6,31
2060	8,54	74,62	62,86	15,52	13,74
2113	8,90	77,77	45,55	4,88	8,22
2118	9,23	49,10	44,31	13,57	2,91

4. Results

2139	15,39	76,36	55,73	1,74	13,99
MEAN	10,78	55,91	32,39	7,54	8,47
SD	7,34	15,78	20,58	3,92	4,25

Table 4.41: Percentage Volume Difference estimated between, respectively, the T1-w manual segmentation, the automatic method proposed by the literature (FS, GMM) and the proposed DNNs trained with contrast (Contrast UNETR, 128, Dice, DA) and without contrast (No contrast UNETR, 96, CE, DA), and the reference gold-standard cT1-w sequence. The performance index was estimated for each subject of the validation set. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

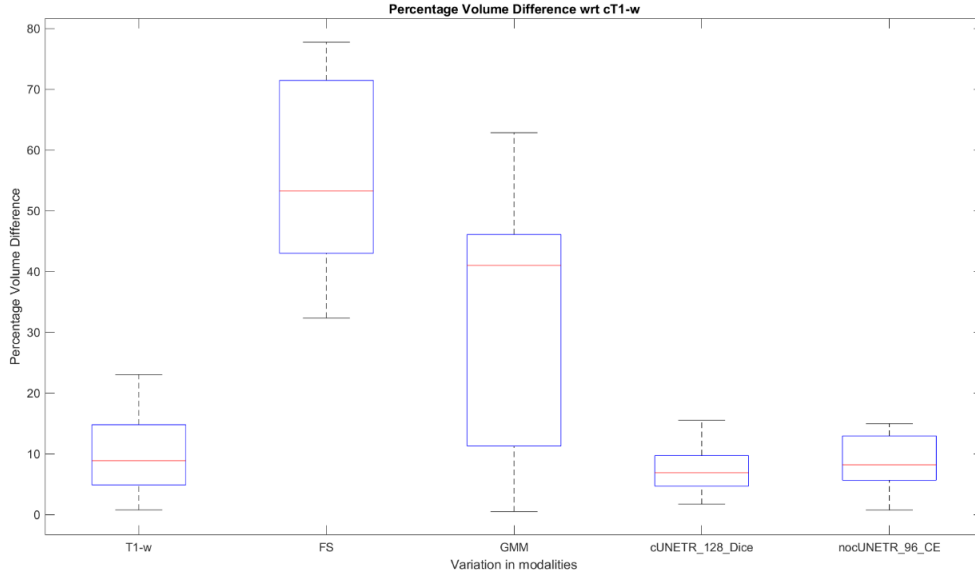


Figure 4.17: Boxplot of the Percentage Volume Difference median and variation intra-modality, calculated between each compared method (T1-w MSeg, FS, GMM, contrast UNETR_128_Dice, no contrast UNETR_96_CE) and the gold-standard cT1-w sequence.

Concerning the outliers, the segmentations obtained through the proposed DNNs have the lower number, while the other automatic methods the higher (Figure 4.18).

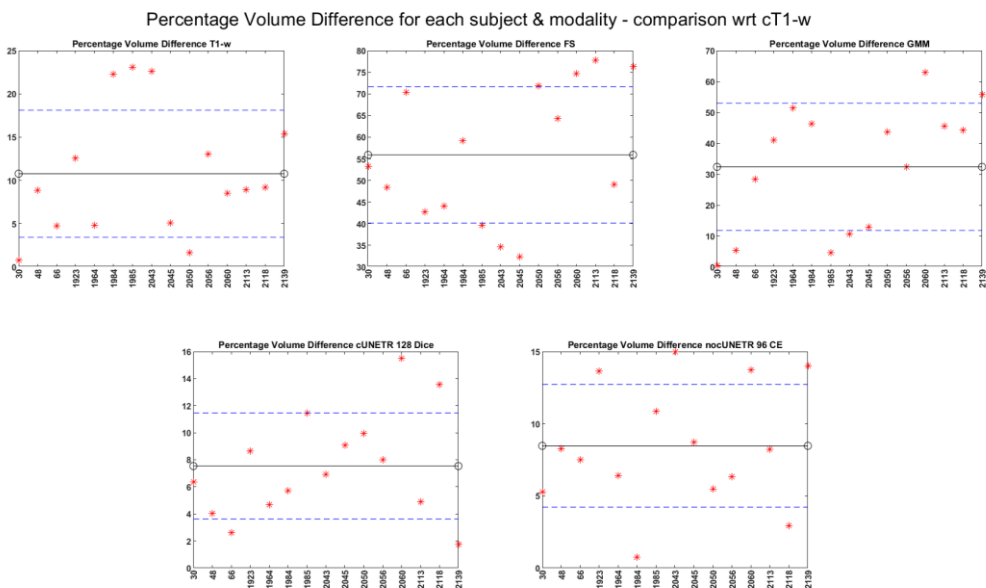


Figure 4.1: Plot of Percentage Volume Difference calculated for each patient between each compared method (T1-w MSeg, FS, GMM, the proposed DNNs) and the gold-standard cT1-w sequence. Mean and standard deviation of each modality are reported in the plots.

The comparison between the automatic segmentation obtained with FS, GMM, and the two UNETR trained with GT with and without contrast T1-w images has highlighted that the performance of the proposed DNNs, trained both with and without contrast, are superior to that of the automatic methods proposed in the literature and, moreover, to that of the manual segmentation performed on the T1-w images without contrast. The detailed analysis is reported in the Discussion chapter, paragraph 5.5.

4.8 VISUAL INSPECTION OF THE PREDICTED SEGMENTATIONS USING THE PROPOSED DNN

The above paragraph has highlighted the better performance of the two proposed DNNs with respect to the other compared methods, both automatic and manual (without contrast). This paragraph reports the images of the segmentations obtained with the UNETR using patch size 128 and Dice loss function with data augmentation trained with contrast and compared to the cT1-w MSeg. Those obtained with UNETR patch size 96 and CE loss function without contrast are not reported. The images were made with FSLeaves. Two patients were selected: the best patient and the worst one. The selection was made considering the best and the worst Percentage Volume Difference. The best-case patient has a Percentage Volume Difference of 1,74%, while the worst has a value of 15,52%.

The *Figure 4.19* shows the T1-w image given as input to the DNN for the best-case patient, while *Figure 4.20* the cT1-w target image. *Figure 4.21* shows the GT cT1-w MSeg overlapped to the cT1-w image, that is the target to be reached. *Figure 4.22* shows the predicted segmentation obtained with UNETR_128_Dice_DA trained with contrast GT overlapped to the cT1-w image.

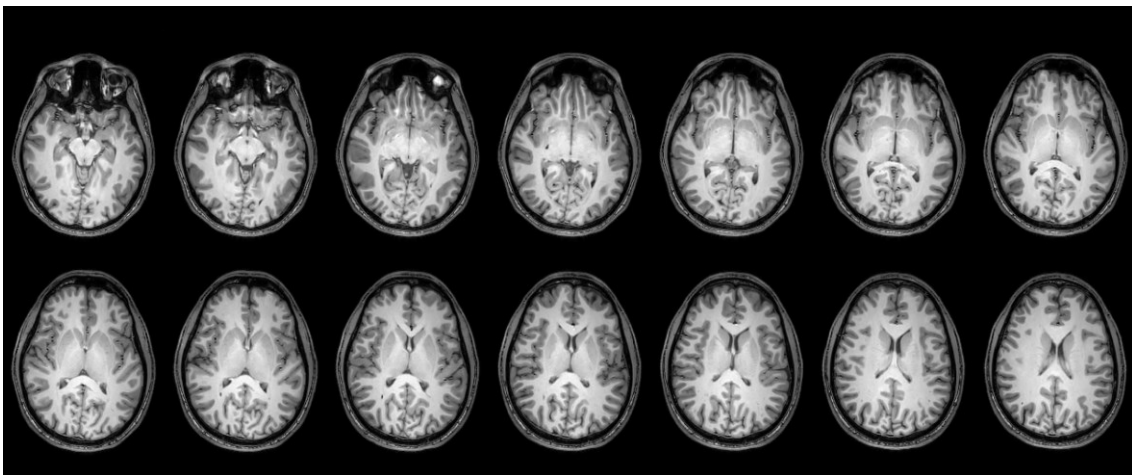


Figure 4.19: Best-case patient T1-w image. Axial view, FSLeaves.

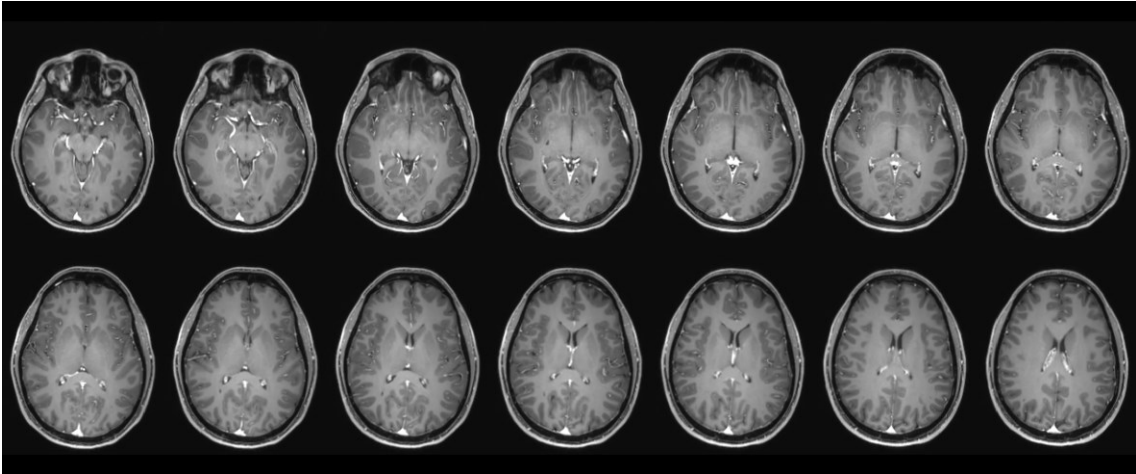


Figure 4.20: Best-case patient cT1-w target image. Axial view, FSLeves.

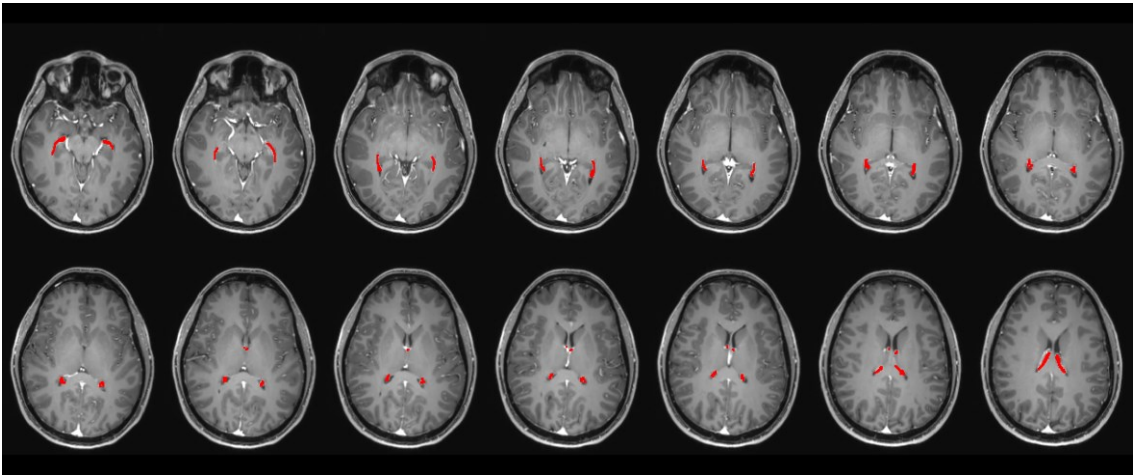


Figure 4.21: Best-case patient cT1-w target image and GT cT1-w MSeg in red. Axial view, FSLeves.

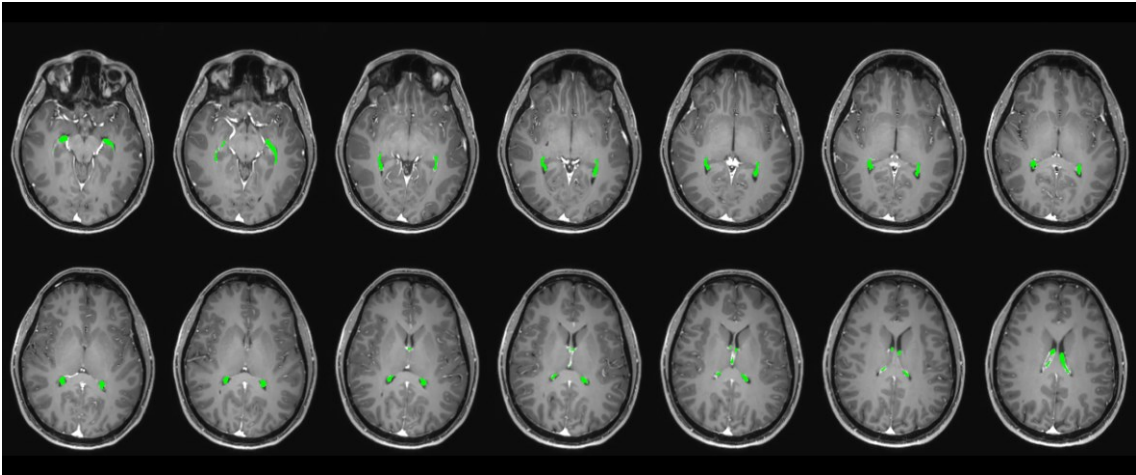


Figure 4.22: Best-case patient cT1-w target image and predicted segmentation (UNETR_128_Dice_DA) in green. Axial view, FSLeves.

Figure 4.23 shows the two segmentations, the gold standard and the predicted one, overlapped.

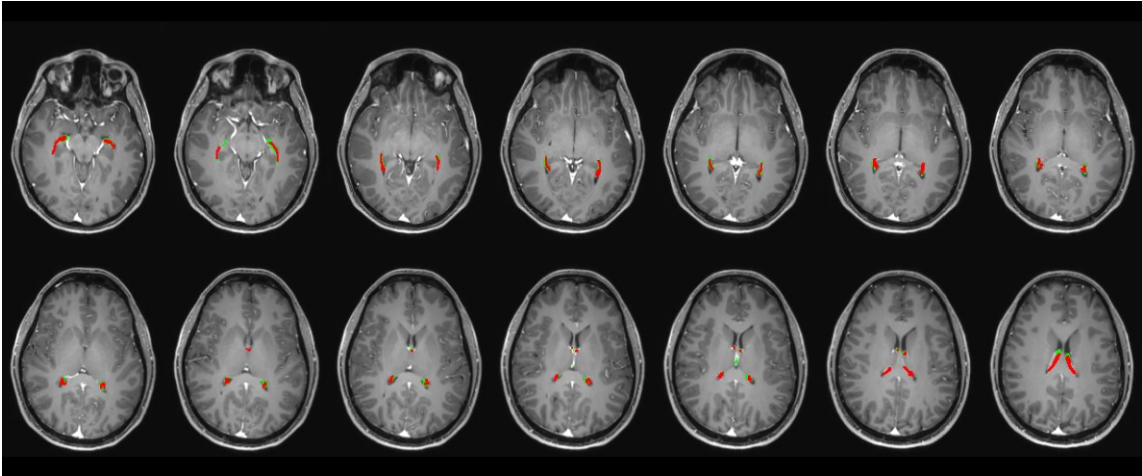


Figure 4.23: Best-case patient cT1-w target image, predicted segmentation (UNETR_128_Dice_DA) in green, GT cT1-w MSeg in red. Axial view, FSLeys.

Figure 4.24 shows the two segmentations overlapped over the worst-case patient cT1-w image.

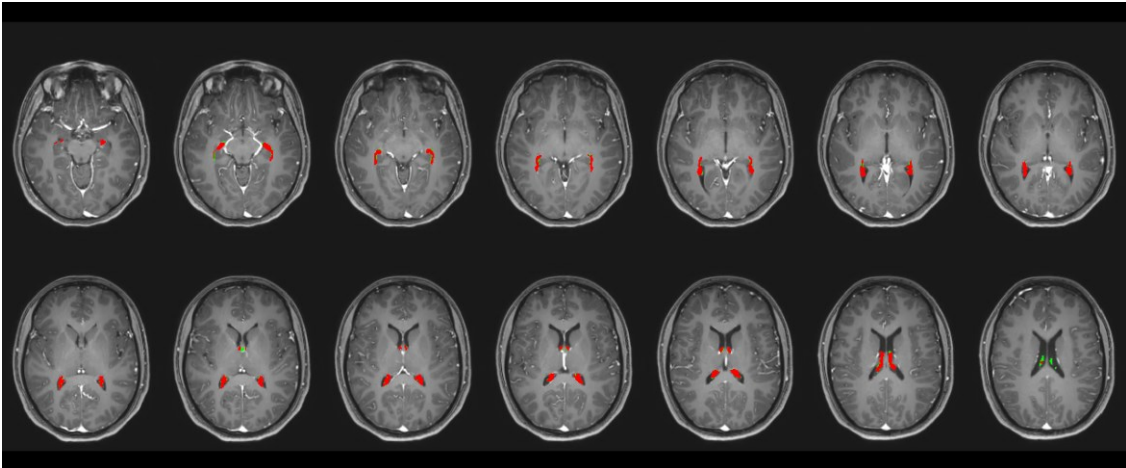


Figure 4.24: Worst-case patient cT1-w target image, predicted segmentation (UNETR_128_Dice_DA) in green, GT cT1-w MSeg in red. Axial view, FSLeys.

By visual inspection, there are no difference between the predicted segmentation and the gold standard one both for the best-case patient and for the worst-case one.

5 DISCUSSION

The first paragraph reports the discussion on the preliminary analysis performed on the sub-dataset, whereas the second paragraph reports an analysis on the whole dataset based on manual GT and existing software. The remaining paragraphs illustrates the performance of the selected DNNs in the trainings and validations phases.

5.1 PRELIMINARY ANALYSIS ON THE SUB-DATASET

The results obtained lead to draw the following conclusions. In terms of Dice, T1-w and FLAIR sequences have the best mean scores with the lower variability as compared to DIR sequence. FS segmentation shows the lower Dice Coefficient: this is due to the high error rate that the algorithm does by segmenting the ChP, losing accuracy especially in the inferior portion of the anatomical structure, confirmed by the visual inspection of the segmentation with ITK-SNAP. The GMM algorithm reaches a good Dice score for an automatic segmentation algorithm but always worse than that obtained with the ground truth (manual segmentation). Concerning the Hausdorff Distance, DIR gives the best values in terms of mean value and variability between subjects. Again, FS and GMM segmentations give worse results with respect to T1-w and FLAIR. In any case, there is not so much difference between FLAIR and DIR looking at the Hausdorff Distance. For what concern Pearson's Correlation Analysis, T1-w and FLAIR are the sequences more correlated with cT1-w and the more significative ones in terms of p-values. FS and GMM have lower correlation coefficient than DIR one, moreover, the p-values are very high, so there is no correlation with the gold standard segmentation.

To conclude, the segmentations obtained with FS and GMM have lower quality compared to the gold standard with respect to the manual segmentations depicted starting by FLAIR, DIR, and T1-w sequences. Therefore, it was decided to consider for further analysis as ground truth, the manual segmentations obtained by the FLAIR and T1-w sequences because they have metrics values comparable to that of the gold-standard and to exclude the DIR. Moreover, the possible biomarker to be used is the volume of the ChP, so the need is to find the sequence that minimizes the discrepancy with respect to the gold-standard segmentation volume. FLAIR and DIR are sequences more correlated with the cT1-w one in terms of volume, however both FLAIR and DIR tend to overestimate the volume of the ChP, with the DIR providing values far from the cT1-w even if mildly correlated to that.

Therefore, DIR was excluded from further analysis and the final goal of this work is now to train DNN over the remaining FLAIR, T1-w images and the combination of both, to predict the ChP segmentation, in order to obtain a result as comparable as possible with that derived from manual segmentation of the cT1-w sequence (the gold-standard).

5.2 PRELIMINARY ANALYSIS ON THE WHOLE DATASET

In terms of Dice, FS segmentation shows the lower mean value: this is due to the high error rate that the algorithm does by segmenting the ChP, losing accuracy especially in the inferior part of the anatomical structure, confirmed by the visual inspection of the segmentation with ITK-SNAP. The GMM algorithm reaches a good Dice score for an automatic segmentation algorithm but always worse than that obtained with the ground truth T1-w and FLAIR (manual segmentation). Concerning the Hausdorff Distance, FLAIR gives the best values in terms of mean value and variability between subjects. FS and GMM segmentations give worse results with respect to T1-w and FLAIR. For what concern Pearson's Correlation Analysis, T1-w and FLAIR are the sequences more correlated with cT1-w. FS and GMM have lower correlation coefficient, even if all p-values are significant. The previous consideration is confirmed by the Linear Regression Analysis and by the OLS Analysis. Observing the OLS (linear regression without intercept) GMM and FS are the segmentations that least accurately estimate the volume of the gold-standard segmentation. Taking into consideration the RMSE (and the MSE), T1-w and FLAIR segmentations are the best. In terms of Percentage Volume Difference, T1-w is the sequence with the lower mean value percentage and variability, followed by the FLAIR one. FS segmentation is again the worst.

To conclude, the segmentations obtained with FS and GMM have lower quality compared to the gold standard with respect to the manual segmentations depicted on FLAIR and T1-w sequences. Moreover, the possible biomarker to be used is the volume of the ChP, so the need is to find the sequence that minimizes the discrepancy with respect to the gold-standard segmentation volume. This preliminary analysis shows that best sequence to be used is the T1-w one in terms of Percentage Volume Difference. On the other hands, FLAIR sequence gives a higher Dice Coefficient, so a more delineated segmentation with lower variability. Even if the volume estimated by FLAIR sequence is correlated to that of cT1-w, the volume predicted by FLAIR is overestimated. However, these results show how difficult the ChP segmentation task is, above all if it is not done with the gold standard.

5.3 DNN VALIDATION SET RESULTS

The results showed in the Results chapter, paragraphs 4.4 and 4.5 led to the following considerations. All the performance indices are estimated with respect to the reference segmentation (the gold standard cT1-w MSeg), regardless of whether the training of the DNNs combination has been performed considering as GT the MSeg on sequences used without contrast agent (T1-w or FLAIR) or the cT1-w MSeg. This was done because the cT1-w MSeg is considered as the gold-standard reference.

The analyzed performance indices are the Dice Coefficient, the 95% Hausdorff Distance and the Percentage Volume Difference evaluated over the whole validation set of subjects. Even if the Jaccard Score, the Volume Difference and the RMSE (MSE) are calculated too, they are not reported in the Results chapter because they follow respectively the Dice Coefficient and the Percentage Volume Difference. The method used to analyze the results was sorting all combinations of DNNs (for each input-GT MSeg combination) for each of the three indices mentioned above.

The results of both paragraphs 4.4 and 4.5 show that sorting for each metric leads to different results in terms of better combinations of DNNs considering all the architectures (3D U-Net, nnU-Net, UNETR, V-Net). This consideration remains consistent even considering the results for each single architecture, always ordering by the three metrics. It is worth noting that considering single indices of performance tend to provide misleading conclusion, therefore it is considered a better practice to perform a multimodal evaluation, based on multiple indices to provide a more complete analysis of the performance.

Generally, the conclusions that can be drawn are the following. First, adding the data augmentation to the original data does not markedly modify the results, consequently, looking at the sorted DNNs combinations, there is no trend that justifies the best performance of the application of data augmentation transforms. This observation was not expected since data-augmentation has proved to help avoiding overfitting in the training phase. However, a hypothesis that can explain this evidence is the ChP size, that is only a small portion of the whole. As consequence, it might be enough to train a DNN using original images without adding transforms. Secondly, the Weighted Cross-Entropy loss function brings to lower performances. This is due probably to the weight associated to the foreground (the ChP) and the background. Even if the patches are extracted with an equal (50 %) probability of having a foreground or a background central voxel, and even if the patches with a foreground central voxel are weighted at 0,9 (the other at 0.1),

this is not enough to balance the discrepancy between the size of the ChP and that of the background. As consequence, for the ChP segmentation task it is preferable not to weigh the two classes in the loss function. Regarding this two initial consideration, Zhao et al. have concluded that the data augmentation transforms combined with the binary Weighted Cross-Entropy loss function bring to higher performances (Zhao et al., 2020). However, the differences with the U-Net trained over original images is minimal, moreover the sensitivity is higher.

Another consideration to do is that the patch size seems not to alter the results in general, particularly for the nnU-Net architecture. This is due to the self-configuring characteristics of the net, that can auto tune the parameters, following the features of the input images. Discarding the outlier combinations, the nnU-Net achieve higher performance regardless of patch size and loss function, except for the wCE. As the nnU-Net, the UNETR is not influenced by the patch size or the loss function. These two architectures are always in the top ten configurations, regardless for the metric you are ordered for. Consequently, on one hand, the nnU-Net is probably the most promising network for the segmentation of the ChP due to the excellent results obtained for the segmentation of various organs and especially for the segmentation of brain tumors and MS lesions (Isensee et al., 2021). On the other hand, the UNETR architecture was tested over an abdominal CT to segment different organs inside the body and the great results obtained by Hatamizadeh et al. are an excellent starting point to test the effectiveness of this method also for the ChP segmentation (Hatamizadeh et al., 2021).

On the contrary, the V-Net is the architecture that leads to the worst performance, above all for what concern the Percentage Volume Difference, that is the possible biomarker to be reached. Concerning this architecture, it is not surprisingly that it does not perform well for this difficult task: indeed, the great results obtained with the V-Net are referred to the prostate segmentation task, that is easier due to the configuration of the anatomical region to be segmented (Milletari et al., 2016).

For the 3D U-Net, this architecture does not follow any notable trend in any case.

The above considerations highlight the difficulty of the segmentation task and the consequent difficulty in choosing the best configuration that can maintain high performance for all the performance indices analyzed. However, this is due in part to the small dataset that may not allow a precise choice to be made, even if generally it is still possible to exclude the V-Net between the architectures and the wCE between the loss functions and to select both nnU-net and UNETR as equivalent suitable solution for this task.

5.4 PERFORMANCE ANALYSIS: COMPARISON BETWEEN MODELS TRAINED WITH GT CT1-W AND THOSE WITH GT WITHOUT CONTRAST (T1-w, FLAIR)

The aim of this thesis is to select the DNN configuration that allows to obtain a segmentation of the ChP for which the difference between the volume of the predicted segmentation and the volume of the GT obtained with the gold standard is minimal: indeed, the possible biomarker is the ChP volume. Moreover, the second goal is to avoid the patient injecting the contrast medium to make the procedure even less invasive. As consequence, in the paragraph 4.6 the comparison analysis was done between the input-GT MSeg without contrast and input-GT MSeg with contrast to investigate the presence of trends regarding the architectures, the patch size or the loss function consistent with respect to the same input regardless of the GT. In addition, it was done the comparison between macro-categories, so comparing the T1-w, the FLAIR or the combined inputs in general with the aim of investigating the best input sequence to be used to improve the performances of the DNNs.

As the main goal is to have higher performance not in terms of Dice Coefficient, in other words the perfect delineation of the segmentation, but in terms of Percentage Volume Difference, so a better volume estimate, the configurations are sorted only for the Percentage Volume Difference. The other two performance indices are reported. This analysis is performed only for all the configurations and not also for each architecture. As anticipated in the Results chapter, the best ten configurations, selected according to the lower Percentage Volume Difference sorting, are then ordered by the $\text{mean}+2*\text{sd}$ (and $\text{m}+3*\text{sd}$), where m is the mean Percentage Volume Difference value and sd is the standard deviation of the Percentage Volume Difference value for each configuration. These two values represent respectively the 95 % and 99 % probabilities to have a Percentage Volume Difference value for each patient in the validation set under a selected threshold. In the literature, a study has analyzed the volume difference between the healthy controls and the multiple sclerosis patients, and this value is around 21,4 % in the used independent dataset (Müller et al., 2022). This value is used as the threshold in our analysis. Only the configurations that are below this threshold (preferably 99 %) have been selected.

The results show similar trend across all the input configurations both with and without contrast. The best ten configurations are always UNETR or nnU-Net, even if it is not possible to select specific patch size or loss function. As said before, the data augmentation does not improve the results, and the wCE and the V-Net are excluded by

this analysis. Concerning the trend of the Dice Coefficient compared to the Percentage Volume Difference, it is generally high in the ten selected configurations, however it is difficult to find a network that maximizes the DC in this subset. Comparing the DNN with and without contrast as described in the paragraphs 4.6.1, 4.6.2 and 4.6.3, the performances are lower for the configurations trained without contrast (i.e., Input T1-w: DC 0.75 with contrast; DC 0.70 without contrast).

Concerning the single input T1-w comparison, two configurations are selected with a probability of 99% for GT cT1-w MSeg (UNETR, patch size 128, Dice loss, DA; UNETR, patch size 96, CE loss, DA) and one for GT T1-w MSeg (UNETR, patch size 96, CE loss, DA). In addition, a UNETR configuration is the same for both. The T1-w single input is able to predict a ChP segmentation extremely accurate in terms of volume, so it is preferable to be used as input for the DNNs.

Regarding the single input FLAIR comparison, none of the two options (GT FLAIR MSeg, GT cT1-w MSeg) give a configuration under the threshold with a 99% probability, even if the mean performance indices are near to that of the T1-w inputs. This is probably due to the type of sequence. The FLAIR sequence shows a contrast between the ChP and the other brain regions higher than the T1-w sequence. However, there is a blurred effect. This is probably the main cause of the lower accuracy in estimating the ChP volume, even if the Dice Coefficient is high. It could be said that the FLAIR is not the best input sequence to segment the ChP if the aim is to have the better volume estimate, while if the goal is to have a good delineation of the segmentation it could be used. Moreover, using the combined inputs (T1-w + FLAIR) the Dice Coefficient is higher than that of the single input, not only using the GT cT1-w, but also using the GT FLAIR. However, none of the configuration can be selected with a threshold of 21,4% with a probability of 99%.

In light of these results, the nnU-Net and UNETR are the architecture that predict the ChP segmentation with higher performances, but UNETR is more accurate considering the threshold value reported in literature. Moreover, the T1-w sequence is the only one that, given as single input to the DNNs, provides an accurate segmentation in terms of volume estimation. On the contrary, in the analyzed dataset, the FLAIR sequence alone does not obtain performance comparable to that of the T1-w single input. Moreover, the FLAIR sequence added to the T1-w one does not improve the performance.

It was decided to consider for the following analyses only the single T1-w input with the best configuration DNN for both GT T1-w and GT cT1-w. The best DNNs configurations to be compared with the state-of-the-art automatic method are two UNETR: UNETR,

patch size 128, Dice loss, DA for the DNN trained with contrast MSeg; UNETR, patch size 96, CE loss, DA for the DNN trained with no contrast MSeg.

5.5 COMPARISON WITH THE STATE-OF-THE-ART AUTOMATIC METHODS FS AND GMM

The aim of this comparison is to investigate if the proposed DNN have better performance, in particular in terms of volume estimation, with respect to the state-of-the-art automatic methods of FS and GMM and with respect to the T1-w MSeg. A second goal is to investigate if the DNN trained without the contrast has performance comparable to that trained with contrast. All the metrics are estimated with respect to the cT1-w MSeg, that is the gold standard segmentation.

The results obtained at the paragraph 4.7 lead to draw the following considerations. Concerning all the performance indices, that are the Dice Coefficient, the 95% Hausdorff Distance, the Percentage Volume Difference, the OLS linear regression, the Pearson's Correlation Analysis, the two proposed DNNs have higher performances compared to cT1-w MSeg than FS or GMM. Moreover, they are better than the T1-w MSeg too. In particular, concerning the Percentage Volume Difference boxplot, the interquartile ranges of both DNNs are under the 21,4 %, while the T1-w MSeg one not.

These results suggest that the main problem of the FS method is that it is quite inaccurate with respect to the MSeg. The GMM improve the FS performance, but the results are far from that obtained with the cT1-w MSeg. On the contrary, the DNNs are able to estimate the ChP volume similarly to the cT1-w MSeg even without the contrast injection.

Comparing to the literature results concerning the DNN architecture to segment the ChP, in this study the results obtained with the UNETR trained with contrast are DC ($0,749 \pm 0,045$) and Percentage Volume Difference ($7,54 \pm 3,92$) %, and with the UNETR without contrast are DC ($0,702 \pm 0,045$) and Percentage Volume Difference ($8,47 \pm 4,25$) %. Schmidt-Mengin et al. found a mean average volume error rate of 20% for the nnU-Net, moreover they don't compare the performance with the cT1-w MSeg but with the T1-w MSeg that was used during the training phase (Schmidt-Mengin et al., 2021). Zhao et al. don't compare to the cT1-w MSeg too, in addition, they don't estimate the discrepancy between the volume of the predicted segmentation and that of the MSeg one.

5.6 VISUAL INSPECTION OF THE PREDICTED SEGMENTATIONS USING THE PROPOSED DNN

The paragraph 4.7 has highlighted the higher performance of the two proposed DNNs with respect to the state-of-the-art automatic methods. The paragraph 4.8 has reported the segmentation, predicted and gold standard, overlapped to the target cT1-w image for both the best-case and worst-case patients for the UNETR_128_Dice_DA.

By visual inspection, the selected DNN performs in the same way for all patients of the validation set.

Generally, it can be concluded that the proposed DNNs outperform the literature and better approximate the gold standard MSeg, even without contrast.

6 CONCLUSION

The Choroid Plexus is a vascular tissue located in the brain ventricles that arouses particular interest in the last years because of its involvement in Alzheimer’s disease (Tadayon, Pascual-Leone, et al., 2020), psychiatric disorders (i.e., depression, schizophrenia) (Althubaity et al., 2022; Lizano et al., 2019) or Multiple Sclerosis inflammatory state (Ricigliano et al., 2021; Vercellino et al., 2008). In particular, the ChP appears to be enlarged in those diseased patients comparing to the healthy controls. Concerning the MS disease, a recent study (Fleischer et al., 2021) suggests the use of the ChP volume, estimated from the ChP segmentation, as a biomarker to evaluate the progression of the disease and the effectiveness of the therapy, becoming an image correlate of the inflammation state. As consequence, the main goal has become develop and validate an automatic method to segment the ChP due to the time-consuming feature of the manual segmentation approach (ground truth) performed preferably on cT1-w images to enhance the contrast between the ChP and the background (gold-standard). However, the contrast injection is not always an option due to its invasiveness.

The aim of this work was using DNN techniques, which have become the new state-of-the-art for image processing, to perform the ChP segmentation task over a cohort of RRMS patients as reliable and accurate as possible. The target to be reached was the cT1-w manual segmentation. The goal was to train different architectures (3D U-Net, UNETR, nnU-Net, V-Net), changing the training parameters (loss function, patch size, data augmentation transforms) and the input sequence without contrast (T1-w, FLAIR, T1-w+FLAIR) to find the combination with the higher performance, calculated with respect to the gold-standard manual segmentation, particularly in terms of ChP volume estimation, that could be the possible biomarker. The selected configuration was compared with the literature proposed automatic methods, FS (Fischl, 2012) and GMM (Tadayon, Moret, et al., 2020).

The main consideration that should be taken into account when comparing segmentation algorithm of the ChP, is that the ChP segmentation task is complex. The complexity is caused mainly by both the size of the ChP compared to the MRI resolution available and the variability of the vasculature inside the ventricles. As consequence, it is not possible to delineate a single-network parameter configuration that provides the best performance

across all the analyzed metrics, especially not using sequences that exploits a contrast injection.

Despite the previous consideration, the following general indications can be outlined. Firstly, the architecture that must be discarded from the initial list is the V-Net, because this network, independently from the parameter configuration, always provide the worst outcome. Secondly, the Weighted Cross-Entropy loss function worsens the performance across all the examined networks. Consequently, for this particular task it is suggested to avoid its use. No trends are found regarding the patch size and the other loss functions. Concerning the data augmentation transforms, it does not improve the results as hoped, nevertheless, the two best DNNs selected for T1-w input, with GT cT1-w MSeg (UNETR with patch size 128, Dice loss function and data augmentation) and GT T1-w MSeg (UNETR with patch size 96, CE loss function and data augmentation), are both trained with data augmentation transforms applied to the training dataset. Thirdly, the T1-w sequence used alone brings to higher performances, even if the sequence without contrast is used. Moreover, the results demonstrate that there is no statistical difference between the use of the cT1-w MSeg or the T1-w MSeg as label during the training phase. This leads to say that the predicted segmentation obtained from the UNETR trained without the use of contrast could be potentially used in future studies. In is worth noting that the T1-w sequence used in this thesis is a very common sequence, that is acquired routinely in almost every brain scan. On the contrary, cT1-w sequence is acquired only in few specific protocols, where the injection is required from a clinical question. Therefore, using a DNN to estimate the volume of the ChP from a standard T1-w sequence with small error when compared to the manual segmentation depicted on the cT1-w sequence is a very promising result. The trained network, in principle, could be also used with retrospectively acquired or public dataset, that are made from even thousands of MRI scan, fast, rapidly and with small errors. On the contrary, the FLAIR sequence seems to provide scarcer results than the T1-w sequence when estimating the ChP Volume, as proved by the Percentage Volume Difference, although the Dice Coefficient is better, above all when combined with the T1-w one. This suggests that the FLAIR sequence can be an alternative to consider when good delineation of the ChP is needed and a quantitative biomarker is not the main requirement. However, the aim of this thesis was to investigate the performance in volume estimation rather than the ability to segment the ChP, since ChP volume is the possible quantitative biomarker that can help investigating neurodegenerative disease like MS.

To conclude, this thesis has demonstrated the concrete possibility to perform an accurate and reliable ChP segmentation with DNNs. In addition, the novelty introduced is the estimation of the Percentage Volume Difference for both the DNNs trained with and without contrast injection. The UNETR seems to be overall the best architecture, for this dataset, for both training performed on T1-w or cT1-w contrast promoting its use in this task for future applications.

6.1 FUTURE DIRECTIONS

This study has performed the analysis of a single-center dataset. This is a weakness that should be mitigated validating the results on others independent dataset. Moreover, the trained networks should be tested with fine tuning of other datasets. Regarding this point, an interesting task can be to establish the minimum number of subjects of a new dataset to be incorporated in a training procedure to obtain results in line the performance of the best DNNs achieved in this thesis.

Moreover, it could be interesting investigating the truthfulness of the threshold (21,4%) used to select the best configurations for this dataset, that was the only data available in the literature to draw a comparison. Lastly, the UNETR has been selected as the best candidate to estimate ChP volume from MRI data. Novel approaches suggest that self-supervised learning over the UNETR can improve its results, this direction should be pursued.

APPENDIX A

This Appendix reports the complete preliminary analysis done on the sub-dataset in order to choose the interpolation method to co-register the images to the T1-w one and to discard one of the three initial MRI sequences (T1-w, FLAIR, DIR) to limit the time-consuming work done by the radiologist. The nine patients were: 000091, 001889, 001922, 002043, 002045, 002050, 002056, 002059, 002060.

Two interpolation method were analyzed: the Nearest-Neighbor Interpolation and the Linear Interpolation (threshold to be selected). The results and the discussion are reported in the following paragraphs.

A.1 NEAREST-NEIGHBOR INTERPOLATION

The results obtained considering the Nearest-Neighbor Interpolation technique are illustrated in the Figures and Tables below.

Dice Coefficient: reference cT1-w

With regards to the Dice Coefficient, FLAIR is the best sequence because of the higher mean value and the lower variability between subjects, while the worst sequence is DIR (Table A 1, Figure A 1).

SUBJECTS	T1-w	FLAIR	DIR
2060	0.687	0.672	0.689
2059	0.680	0.688	0.588
2056	0.694	0.745	0.685
2051	0.579	0.551	0.545
2050	0.623	0.637	0.649
2045	0.688	0.712	0.667
1922	0.737	0.690	0.707
1889	0.638	0.659	0.668
91	0.732	0.690	0.743
MEAN	0.673	0.672	0.660
SD	0.051	0.055	0.060

Table A 1: Dice Coefficient of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

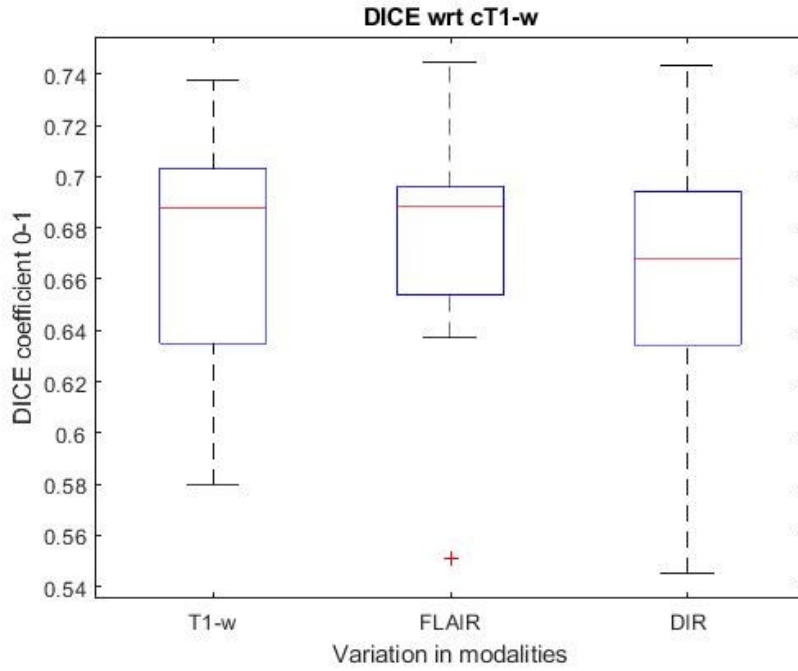


Figure A 1: Boxplot of the Dice Coefficient median and variation intra-modality, calculated between each image modality (T1-w, FLAIR, DIR) and the gold-standard cT1-w sequence.

Hausdorff Distance: reference cT1-w

The best sequence in terms of Hausdorff Distance is DIR but the mean value is like that of FLAIR one. Patient 2059 is the outlier for this metric (Table A 2, Figure A 2).

SUBJECTS	T1-w	FLAIR	DIR
91	2	2.236	2.236
1889	5.099	3.3167	3.317
1922	4.898	2.236	2
2045	10.029	7.810	2.236
2050	12.083	2.828	2.236
2051	11.358	7.810	6.164
2056	2.449	2	2.449
2059	16.910	5.678	21.932
2060	4.123	2.236	2.236
MEAN	7.661	4.017	4.978
SD	5.127	2.424	6.489

Table A 2: 95% Hausdorff Distance of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

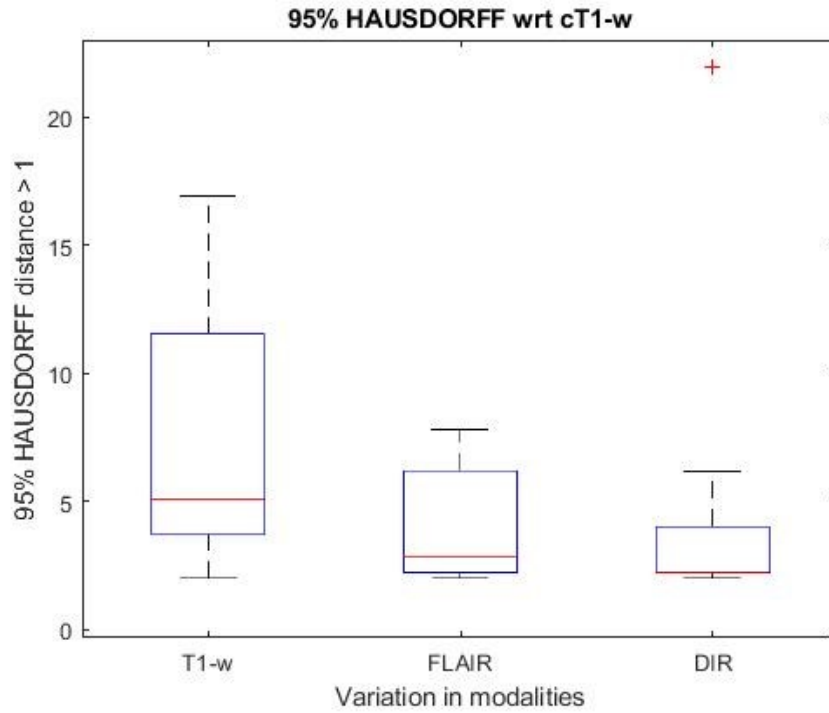


Figure A 2: Boxplot of the 95% Hausdorff Distance median and variation intra-modality, calculated between each image modality (T1-w, FLAIR, DIR) and the gold-standard cT1-w sequence.

Volume Analysis: reference cT1-w

In reference to volume analysis, FLAIR and T1-w sequences mean volume values are more similar to that of cT1-w, in contrast to DIR sequence (Table A 3).

SUBJECTS	cT1-w	T1-w	FLAIR	DIR
2060	4556	4007	5035	5323
2059	4100	4123	4607	3830
2056	3204	3441	3484	5149
2051	3510	2960	4013	5704
2050	2569	2407	2557	4020
2045	2670	3102	3026	3966
1922	3145	3436	3973	4419
1889	3079	2386	2650	3066
91	3019	3056	3415	3473
MEAN	3316.89	3213.11	3640	4327.78
SD	647.24	611.84	846.76	891.86

Table A 3: Segmentation Volume (1 voxel = 1 mm³) calculated for each patient for each available manual segmentation sequence and the gold-standard cT1-w sequence. Subjects are reported with the original ID, mean and standard deviation of each image modality are reported on the bottom rows.

However, DIR sequence has the lower RMSE value (Table A 4).

SEQUENCE	RMSE	MSE
T1-w	1285.484	1652469.778
FLAIR	1383.046	1912816.445
DIR	1039.107	1079742.667

Table A 4: RMSE and MSE for each image modality (T1-w, FLAIR, DIR) with respect to the gold-standard cT1-w sequence.

Looking at the OLS analysis, DIR is the sequence that gives the best approximation of the gold-standard segmentation (*Figure A 3*).

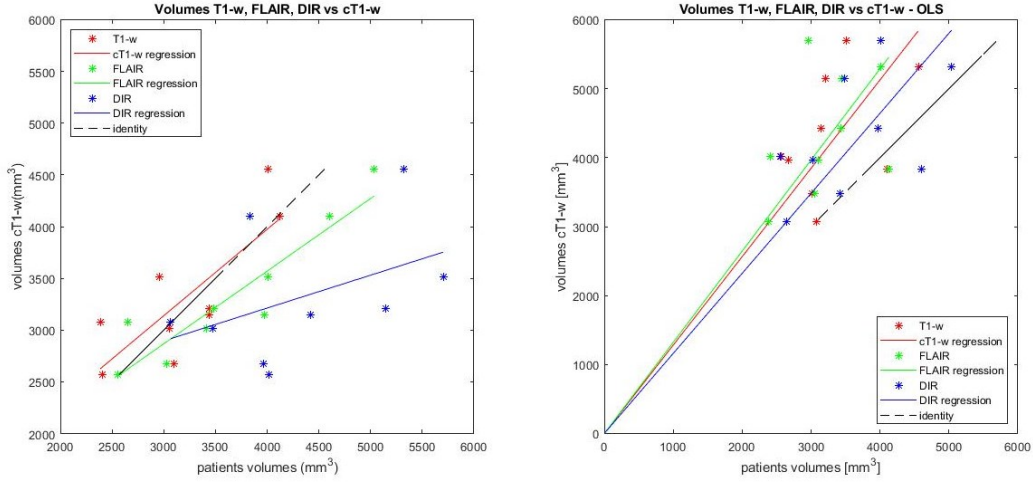


Figure A 3: Panel left: Linear Regression of subject's volume for each modality; panel right: OLS volume representation for each modality.

FLAIR sequence seems to be more correlated to cT1-w considering the Pearson's Correlation Coefficient, DIR the worst (*Table A 5*).

P'S CORR	cT1-w	T1-w	FLAIR	DIR
cT1-w	1	0.789	0.919	0.437
T1-w		1	0.899	0.381
FLAIR			1	0.549
DIR				1

Table A 5: Pearson's Correlation Analysis coefficients between each image modality (T1-w, FLAIR, DIR) and the gold-standard cT1-w sequence. The significant correlation coefficients ($\alpha=0,05$) are highlighted: the others are not significant.

A.1.1 Discussion

The conclusions drawn at the end of this analysis show that, with regards to the Dice Coefficient, FLAIR is the best sequence because of the higher mean value and the lower variability between subjects, while the worst sequence is DIR. Patient 2051 has the lower Dice value in all sequences, so it is the main outlier of the dataset. This is confirmed by the visual inspection of the patient's sequences. In fact, the patient has some cysts inside the ChP. The best sequence in terms of Hausdorff Distance is DIR but the mean value is like that of FLAIR one, even if the DIR sequence has a lower variability between the subjects (if patient 2059 is excluded). Patient 2059 is the outlier for this metric. With reference to the volume analysis, FLAIR and T1-w sequences mean volume values are more similar to that of cT1-w, in contrast to DIR sequence. However, DIR sequence is that with the lower RMSE. Once again, FLAIR seems to be more correlated to cT1-w

considering the Pearson’s Correlation Coefficient. On the contrary, looking at the OLS analysis, DIR is the best sequence.

These results are ineffective because, as that obtained by visual inspection, they mark again the seeming equality between FLAIR and DIR. Therefore, it was decided to change the interpolation method. The choice fell on linear interpolation.

A.2 LINEAR INTERPOLATION

The primary objective was to identify the threshold that represents the best match point of the three sequences FLAIR, DIR, T1-w with cT1-w. In this case the compared thresholds are: [0.10 0.15 0.20 0.30 0.40 0.50 0.55 0.6 0.7 0.8 0.85 0.9 0.95 0.99]

The *Tables* and *Figures* below show the results obtained with the linear interpolation co-registration technique.

Dice Coefficient: reference cT1-w

Table A 6 shows the mean Dice Coefficient values for each selected threshold in each sequence, while the *Figure A 4-A 6* below show the variability of the metric for each threshold and for each modality.

Th	0.10	0.15	0.20	0.30	0.40	0.50	0.55	0.60	0.70	0.80	0.85	0.90	0.95	0.99
T1-w	0.66	0.67	0.68	0.69	0.69	0.69	0.68	0.67	0.64	0.61	0.58	0.54	0.49	0.42
FLAIR	0.74	0.74	0.73	0.72	0.71	0.70	0.69	0.67	0.66	0.64	0.63	0.60	0.57	0.53
DIR	0.73	0.73	0.72	0.71	0.70	0.67	0.66	0.64	0.62	0.59	0.58	0.56	0.53	0.48

Table A 6: Dice Coefficient of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence for each threshold Th. Only mean values are reported.

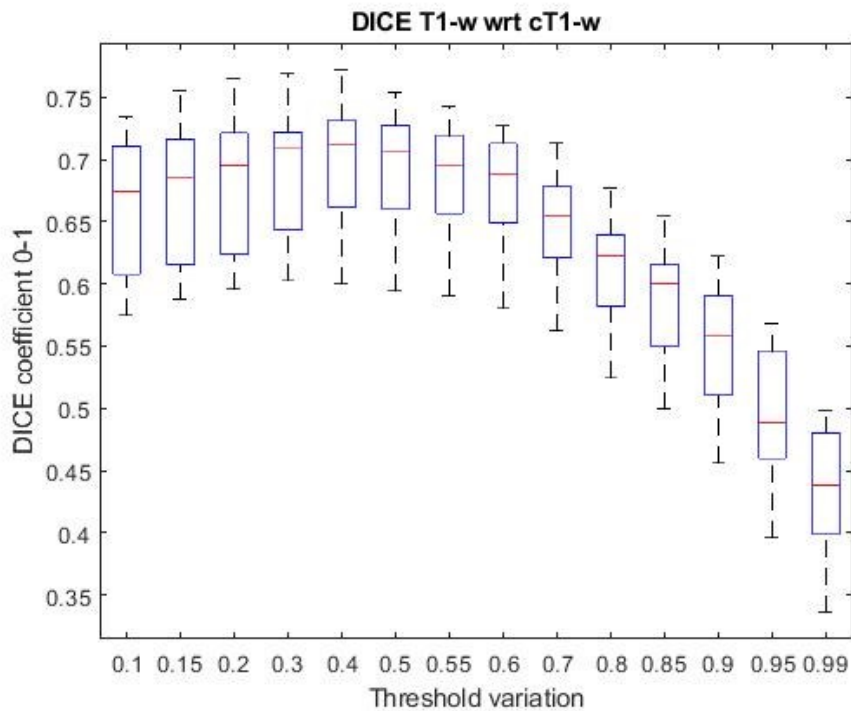


Figure A 4: Boxplot of the Dice Coefficient median and variation in T1-w modality for each threshold Th.

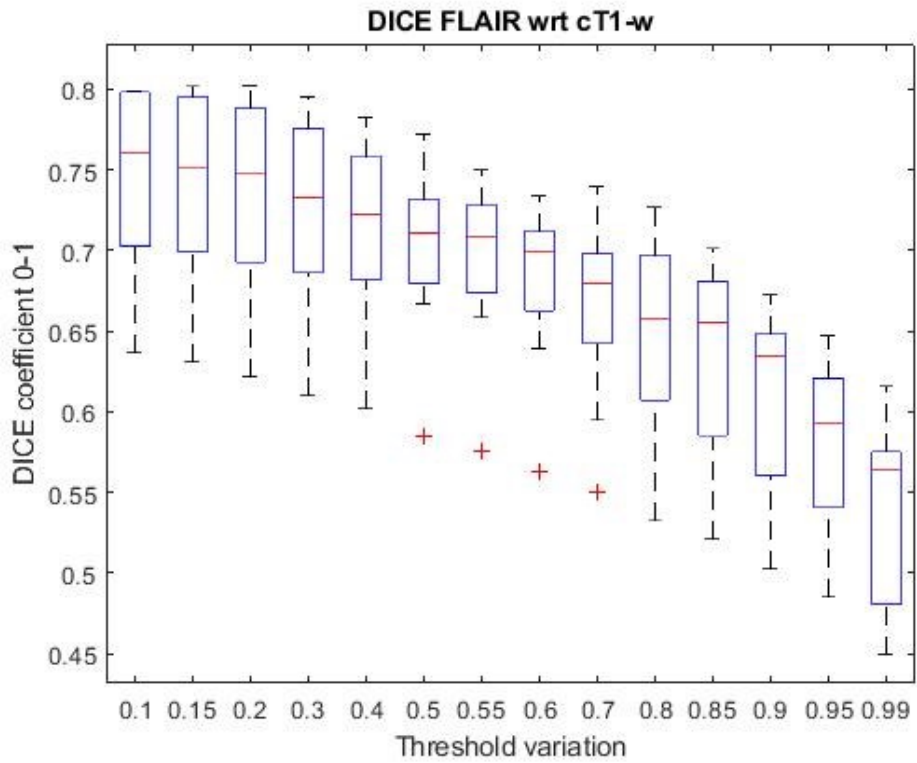


Figure A 5: Boxplot of the Dice Coefficient median and variation in FLAIR modality for each threshold Th .

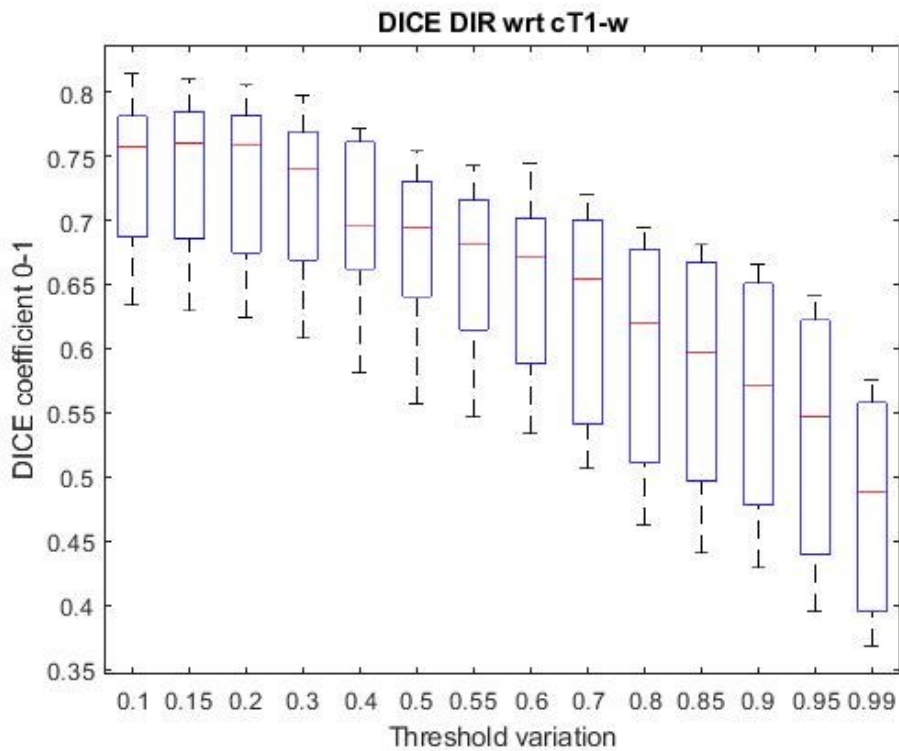


Figure A 6: Boxplot of the Dice Coefficient median and variation in DIR modality for each threshold Th .

Figure A 7 shows the mean Dice Coefficient values for each threshold and for each sequence type.

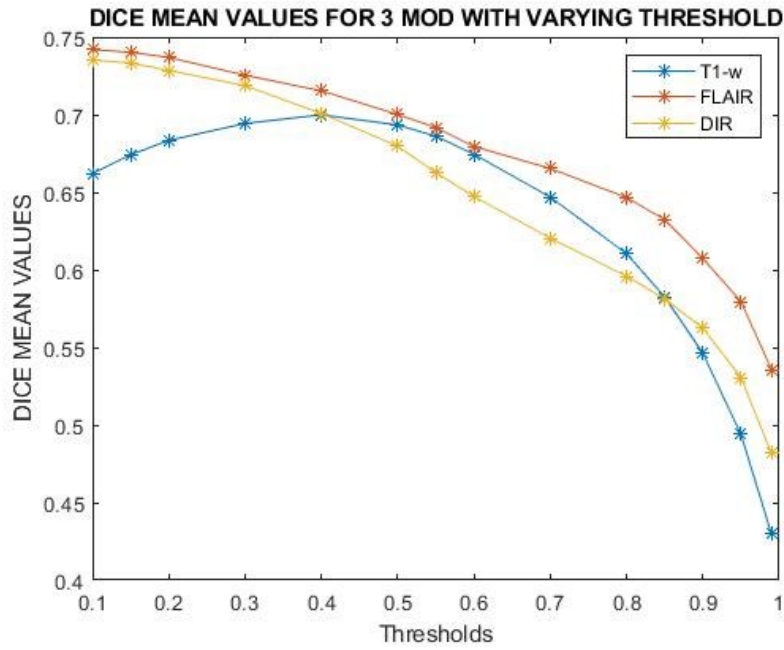


Figure A 7: Mean Dice Coefficient for each threshold Th for each modality

Hausdorff Distance: reference cT1-w

Table A 7 shows the mean 95% Hausdorff Distance values for each selected threshold in each sequence, while the Figure A 8-A 10 below show the variability of the metric for each threshold and for each modality.

Th	0.10	0.15	0.20	0.30	0.40	0.50	0.55	0.60	0.70	0.80	0.85	0.90	0.95	0.99
T1-w	8.97	8.81	8.78	8.43	8.05	7.57	7.24	6.83	5.40	3.82	3.41	3.15	3.06	3.11
FLAIR	4.53	4.57	4.42	4.35	4.30	4.16	4.18	4.15	4.16	4.00	3.87	3.88	3.58	3.35
DIR	5.27	5.28	5.27	5.15	5.13	5.08	5.03	4.99	4.92	4.93	4.97	4.96	5.01	5.08

Table A 7: 95% Hausdorff Distance of the manual segmentation calculated for each patient between each available sequence and the gold-standard cT1-w sequence for each threshold Th . Only mean values are reported.

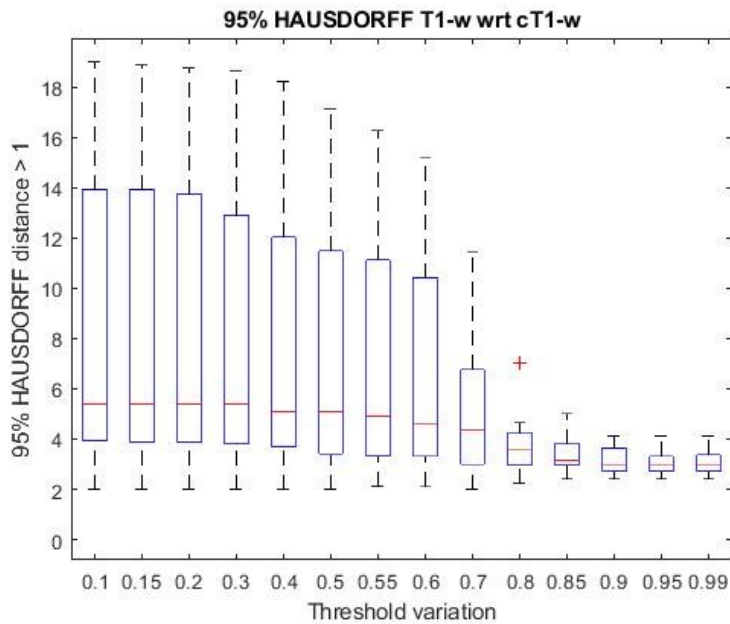


Figure A 8: Boxplot of the 95% Hausdorff Distance median and variation in T1-w modality for each threshold Th .

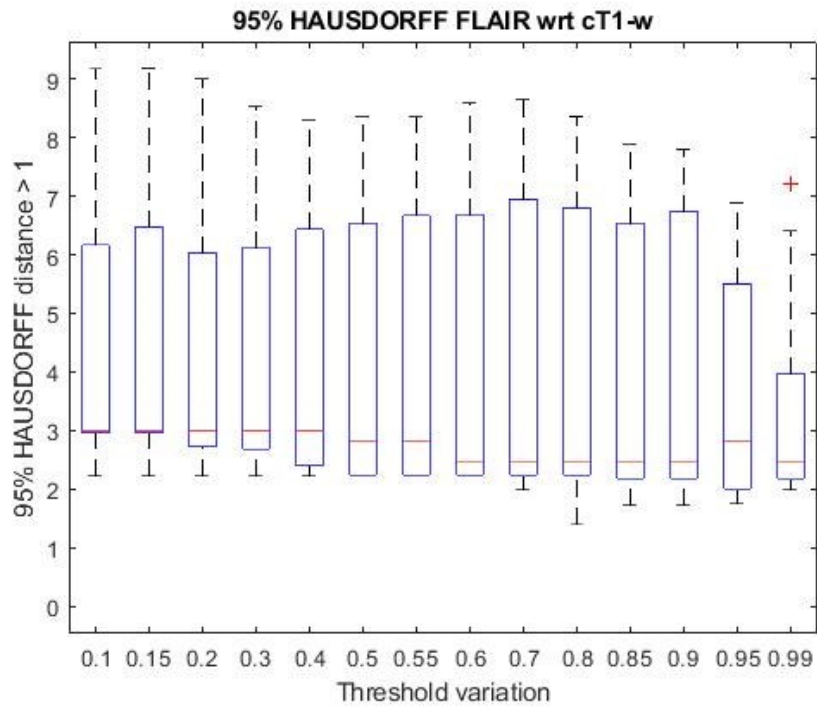


Figure A 9: Boxplot of the 95% Hausdorff Distance median and variation in FLAIR modality for each threshold Th .

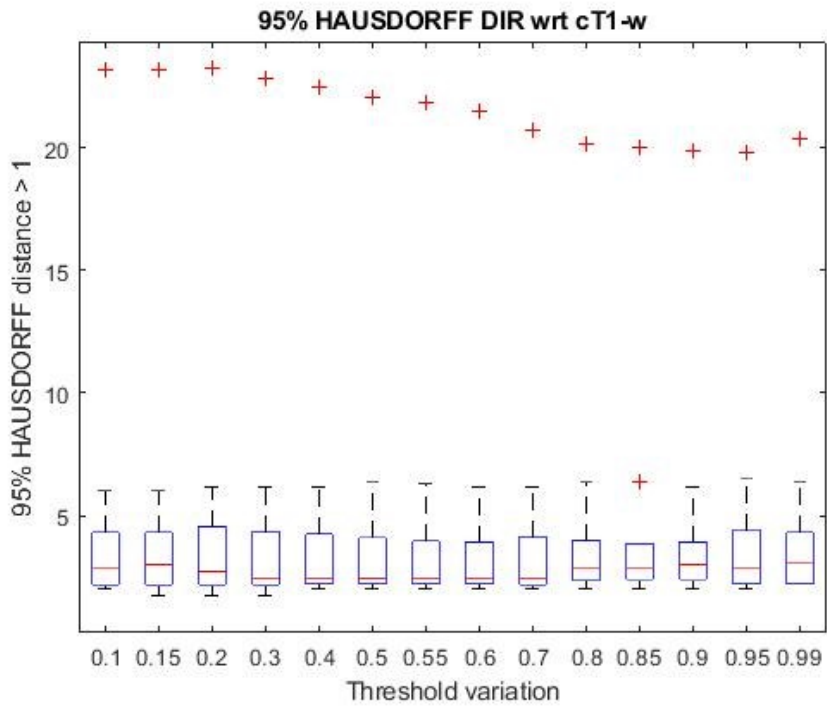


Figure A 10: Boxplot of the 95% Hausdorff Distance median and variation in DIR modality for each threshold Th .

Figure A 11 shows the mean 95% Hausdorff Distance values for each threshold and for each sequence type.

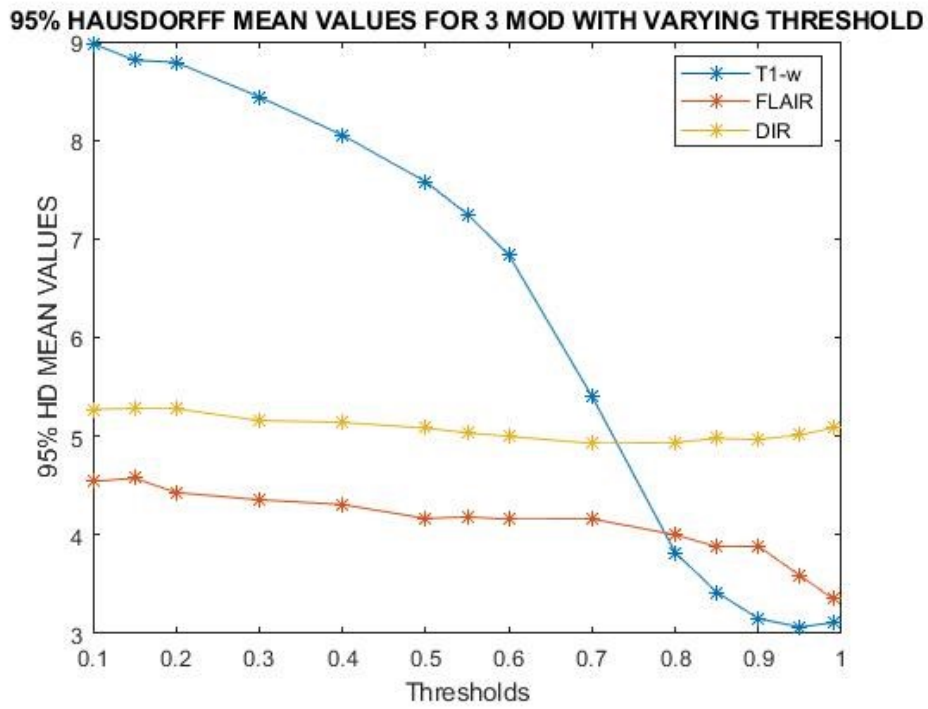


Figure A 11: Mean 95% Hausdorff Distance for each threshold Th for each modality.

Volume Analysis: reference cT1-w

In this case, the volume values are not reported. Only the Pearson’s Correlation Analysis for some exemplificative thresholds and RMSE are reported in this section.

Figure A 12 shows the mean RMSE values for each threshold for each sequence type.

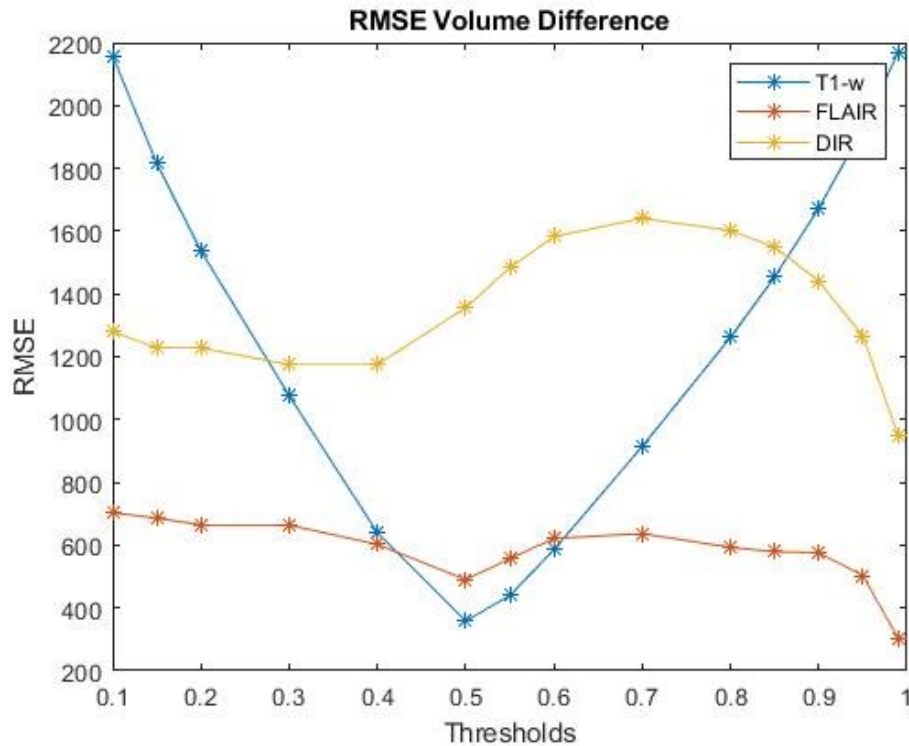


Figure A 12: RMSE for each threshold Th for each modality.

Appendix A

The following *Figures (A 13, A 14, A 15)* shows the Pearson's Correlation coefficient and relative p-values for three selected thresholds: 0,1; 0,5; 0,9.

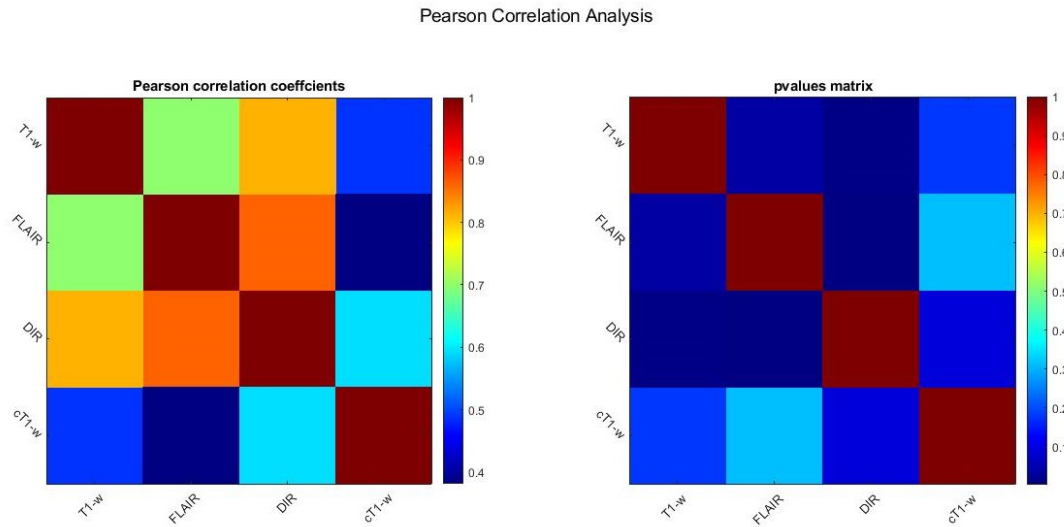


Figure A 13: $Th=0.1$. Left panel: Pearson's Correlation Coefficients; right panel: p-values matrix.

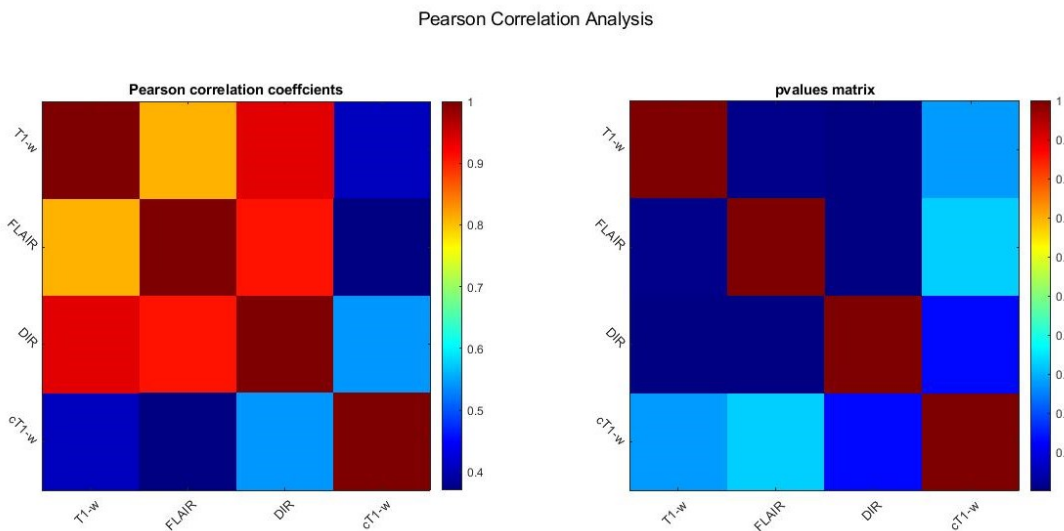


Figure A 14: $Th=0.5$. Figure: Left panel: Pearson's Correlation Coefficients; right panel: p-values matrix.

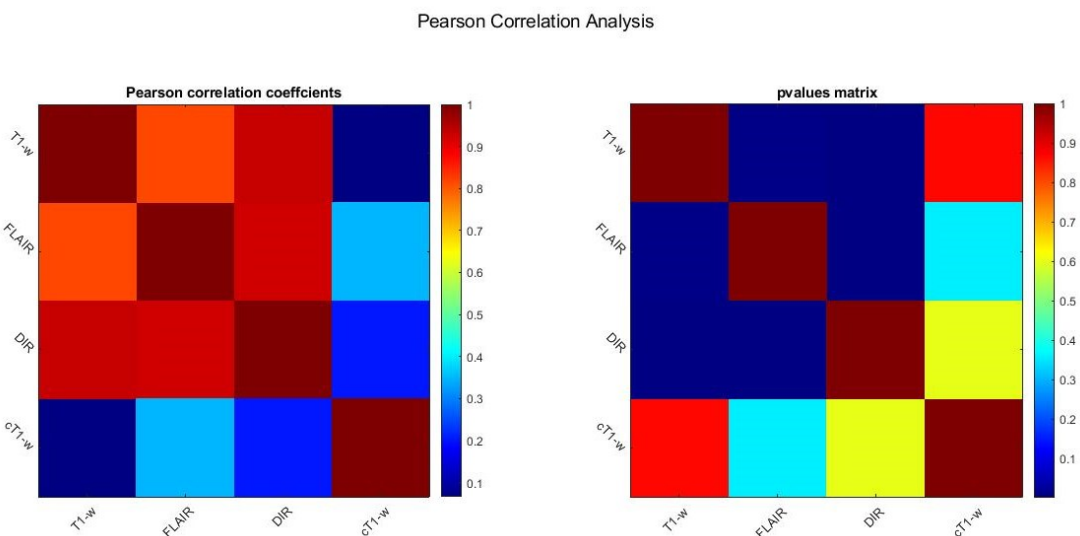


Figure A 15: $Th=0.9$. Figure: Left panel: Pearson's Correlation Coefficients; right panel: p-values matrix.

A.2.1 Discussion

The analysis of the obtained results shows that the best threshold is 0.50. This value allows to find a balance between a high mean Dice Coefficient value and a low data variability. Moreover, T1-w has its maximum value around 0.50. Looking at the mean Dice values for every tested threshold in every sequence, FLAIR seems to be again the best sequence to use to increment the performances. Considering the Hausdorff Distance, 0.50 seems to be again a good threshold value. Moreover, looking at the mean Hausdorff Distance, FLAIR and DIR seem to be the best sequences to use because they reach the minimum value, while T1-w achieves very high values (except for the range 0.80-0.99). However, this metric has a big variability between data. With regards to the Pearson's Correlation Analysis, considering not only the Pearson's Coefficient but also the p-values, FLAIR and DIR are once again the sequences more correlated with the cT1-w volume for all the tested thresholds. However, considering the mean RMSE for all the thresholds in every investigated sequence, T1-w seems to be the best sequence thresholding all other sequences at 0.50, but FLAIR is second one. Moreover, the graph curves for these two sequences show an elbow point which could be interpreted as the best value to select.

Considering all these results, it has been chosen to threshold the images at 0.50.

The preliminary analysis was performed co-registering the images cT1-w, FLAIR and DIR to the T1-w one using the linear interpolation technique with threshold 0.50. The results are shown and explained in the Results chapter and in the Discussion one.

APPENDIX B

This Appendix reports the results of the performance indices calculated over the validation set.

The following Tables report the mean values (and standard deviation ones) of the performance indices or the single value over all subjects (RMSE). The MSE and the Volume Difference are not reported because of the presence of the RMSE and the Percentage Volume Difference.

Sections B.2 and B.3 reports the general Tables with all 96 combinations for each input-ground truth combination. The other sections report the Table-per-DNN for each input-ground truth combination ordered by Percentage Volume Difference.

B.1 LEGEND

The DNNs are named considering the training variable parameters used as follows:

DNN_PatchSize_LossFunction_DataAugmentation

where DNN, PatchSize, LossFunction and DataAugmentation represent:

- DNN (Deep Neural Network architecture): 3DUNET (3D U-Net), DynUNET (nnU-Net), VNET (V-Net), UNETR
- PatchSize (Patch Size): 64 (64x64x64), 96 (96x96x96), 128 (128x128x128)
- LossFunction (Loss Function): Dice (Generalized Dice Loss), DiceCE (Combination of Dice Loss and Cross-Entropy Loss), CE (Cross-Entropy Loss), wCE (Weighted-Cross-Entropy)
- DataAugmentation (Data Augmentation): DA (Data Augmentation Transforms applied), noDA (no application of Data Augmentation Transforms)

B.2 DNN VALIDATION SET RESULTS: TRAINING WITH GROUND TRUTH WITH CONTRAST

The performance indices are calculated making the comparison between the predicted segmentation obtained with each DNN and the manual segmentation obtained from the images with contrast, that is considered henceforth the gold standard ground truth (GT) (cT1-w). Three are the input – manual segmentation (MSeg) examined combinations: T1-w – cT1-w; FLAIR – cT1-w; T1-w+FLAIR – cT1-w.

Appendix B

Table B 1: Metrics on validation set for DNN trained over T1-w images with GT cT1-w MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	%Vol Diff mean	%Vol Diff sd	RMSE
UNETR_128_Dice_DA	0,75	0,04	1,66	0,50	0,60	0,05	7,54	3,79	271
DynUNET_96_DiceCE_DA	0,77	0,03	1,47	0,33	0,62	0,04	7,86	6,60	311
DynUNET_96_CE_DA	0,76	0,04	1,47	0,33	0,62	0,05	7,94	5,20	312
DynUNET_128_CE_DA	0,76	0,04	1,55	0,59	0,61	0,05	8,00	5,07	301
UNETR_96_CE_DA	0,75	0,04	1,54	0,33	0,60	0,05	8,30	3,83	289
DynUNET_64_CE_DA	0,76	0,04	1,59	0,77	0,62	0,05	8,58	5,26	320
UNETR_128_DiceCE_noDA	0,73	0,05	1,92	0,89	0,57	0,06	8,72	6,69	308
3DUNET_64_DiceCE_DA	0,73	0,04	1,83	0,62	0,57	0,05	8,74	5,98	352
UNETR_128_CE_DA	0,74	0,04	1,90	1,12	0,59	0,06	8,98	4,78	331
DynUNET_128_DiceCE_DA	0,76	0,04	1,83	1,77	0,62	0,05	9,04	4,87	323
DynUNET_96_CE_noDA	0,75	0,04	1,64	0,56	0,60	0,05	9,15	6,64	338
UNETR_64_CE_DA	0,75	0,04	1,65	0,37	0,60	0,05	9,22	6,18	359
DynUNET_96_DiceCE_noDA	0,75	0,05	1,67	0,70	0,60	0,06	9,31	6,19	346
DynUNET_128_Dice_DA	0,77	0,04	1,70	1,35	0,62	0,05	9,33	7,68	368
DynUNET_96_Dice_noDA	0,76	0,04	1,50	0,53	0,61	0,05	9,34	6,34	364
DynUNET_96_Dice_DA	0,77	0,04	1,59	0,72	0,62	0,05	9,36	7,34	349
UNETR_64_Dice_DA	0,74	0,04	1,64	0,32	0,59	0,05	9,46	5,19	348
UNETR_64_DiceCE_DA	0,74	0,04	1,69	0,51	0,59	0,05	9,49	5,45	341
DynUNET_128_CE_noDA	0,76	0,04	1,49	0,33	0,61	0,05	9,84	6,53	351
3DUNET_96_CE_noDA	0,72	0,05	2,17	1,82	0,56	0,06	9,85	4,48	345
UNETR_128_CE_noDA	0,72	0,05	2,01	1,14	0,56	0,06	9,93	4,21	339
DynUNET_64_Dice_noDA	0,76	0,04	9,76	30,07	0,61	0,06	10,12	8,06	382
UNETR_96_DiceCE_DA	0,75	0,04	1,69	0,59	0,60	0,05	10,13	9,93	398
3DUNET_96_CE_DA	0,74	0,05	2,20	1,97	0,58	0,06	10,19	4,39	339
DynUNET_128_DiceCE_noDA	0,75	0,04	1,87	1,29	0,60	0,06	10,20	5,50	374
3DUNET_96_DiceCE_DA	0,73	0,04	1,72	0,49	0,57	0,05	10,29	7,60	371
3DUNET_64_DiceCE_noDA	0,72	0,05	2,05	1,03	0,56	0,05	10,37	6,32	399
UNETR_64_CE_noDA	0,74	0,05	1,84	0,99	0,59	0,06	10,38	6,76	389
DynUNET_64_CE_noDA	0,75	0,04	1,97	1,61	0,60	0,06	10,41	5,66	371
3DUNET_128_Dice_noDA	0,70	0,05	2,06	1,07	0,55	0,05	10,45	6,24	371
UNETR_64_DiceCE_noDA	0,74	0,05	1,90	1,27	0,59	0,06	10,49	6,64	391
3DUNET_96_DiceCE_noDA	0,70	0,05	2,16	1,32	0,55	0,06	10,60	5,35	368
3DUNET_128_CE_DA	0,72	0,06	2,42	2,71	0,57	0,07	10,61	4,93	373
DynUNET_128_Dice_noDA	0,76	0,04	1,70	1,01	0,61	0,05	10,68	8,16	403
DynUNET_64_DiceCE_DA	0,75	0,04	1,66	0,49	0,61	0,06	10,73	6,11	386
DynUNET_64_DiceCE_noDA	0,74	0,04	1,69	0,54	0,59	0,05	10,74	6,04	408
UNETR_64_Dice_noDA	0,74	0,05	13,18	42,39	0,58	0,06	10,80	6,59	387
3DUNET_96_Dice_noDA	0,72	0,04	1,87	0,63	0,57	0,05	10,98	8,71	395
UNETR_128_Dice_noDA	0,72	0,06	2,33	1,76	0,57	0,07	11,00	4,75	370
3DUNET_128_DiceCE_noDA	0,69	0,05	2,36	1,66	0,53	0,06	11,04	5,50	394
UNETR_96_CE_noDA	0,72	0,05	1,81	0,58	0,57	0,06	11,09	7,76	425
DynUNET_64_Dice_DA	0,75	0,04	25,54	43,70	0,60	0,05	11,12	10,32	429
3DUNET_64_CE_noDA	0,73	0,05	1,80	0,83	0,57	0,05	11,14	7,58	443

Appendix B

UNETR_96_DiceCE_noDA	0,73	0,05	2,28	2,00	0,58	0,06	11,20	5,64	378
3DUNET_128_DiceCE_DA	0,72	0,05	2,12	1,64	0,56	0,06	11,45	8,79	413
3DUNET_64_Dice_DA	0,72	0,04	26,90	50,06	0,57	0,05	11,53	8,26	414
3DUNET_64_Dice_noDA	0,73	0,04	1,75	0,53	0,57	0,05	11,58	8,63	430
UNETR_128_DiceCE_DA	0,75	0,04	1,85	1,15	0,60	0,06	11,63	8,80	413
3DUNET_128_CE_noDA	0,70	0,05	2,44	2,17	0,55	0,06	11,70	7,41	389
3DUNET_96_Dice_DA	0,74	0,04	1,58	0,32	0,59	0,05	13,81	9,59	474
VNET_96_DiceCE_DA	0,59	0,06	29,32	61,57	0,42	0,06	14,59	8,84	533
3DUNET_128_Dice_DA	0,74	0,04	1,58	0,32	0,59	0,05	14,67	10,26	493
3DUNET_64_CE_DA	0,72	0,05	2,05	0,78	0,56	0,05	15,35	9,09	572
UNETR_96_Dice_DA	0,75	0,05	1,98	1,66	0,60	0,06	16,12	9,97	530
VNET_128_CE_DA	0,62	0,06	3,81	3,59	0,45	0,06	16,18	12,30	598
VNET_128_DiceCE_DA	0,62	0,05	3,32	1,95	0,45	0,05	17,00	11,69	566
DynUNET_128_wCE_noDA	0,74	0,04	1,62	0,49	0,59	0,05	17,18	11,53	567
VNET_64_DiceCE_DA	0,58	0,05	4,45	2,42	0,41	0,05	17,82	11,21	664
VNET_128_DiceCE_noDA	0,59	0,06	3,75	1,81	0,42	0,06	18,49	15,98	736
UNETR_96_Dice_noDA	0,73	0,05	19,87	46,47	0,58	0,06	18,51	11,55	589
VNET_128_CE_noDA	0,61	0,08	3,59	2,75	0,44	0,08	18,54	15,68	762
DynUNET_96_wCE_noDA	0,75	0,04	1,47	0,34	0,60	0,05	19,55	11,09	617
VNET_96_CE_noDA	0,58	0,07	3,49	1,87	0,42	0,07	21,97	14,14	816
VNET_64_DiceCE_noDA	0,59	0,05	3,97	2,30	0,42	0,05	23,10	13,72	850
VNET_96_DiceCE_noDA	0,55	0,06	15,62	33,67	0,38	0,06	27,00	19,58	903
UNETR_128_wCE_noDA	0,71	0,05	1,95	0,72	0,56	0,06	28,51	13,26	861
VNET_96_CE_DA	0,55	0,06	6,04	3,30	0,39	0,06	31,27	13,98	1080
DynUNET_64_wCE_noDA	0,73	0,04	1,57	0,37	0,58	0,05	32,83	12,79	984
3DUNET_128_wCE_noDA	0,69	0,05	2,40	1,77	0,52	0,06	34,37	17,45	1030
VNET_64_CE_noDA	0,53	0,09	7,63	4,10	0,36	0,08	36,08	17,31	1257
UNETR_96_wCE_noDA	0,71	0,05	1,94	0,77	0,55	0,06	41,32	16,14	1226
UNETR_64_wCE_noDA	0,72	0,04	1,76	0,55	0,56	0,05	42,09	15,27	1259
DynUNET_128_wCE_DA	0,73	0,04	1,59	0,39	0,58	0,05	42,73	14,90	1256
3DUNET_64_wCE_DA	0,69	0,04	1,98	0,75	0,53	0,05	47,34	21,60	1426
3DUNET_96_wCE_noDA	0,69	0,05	2,09	0,67	0,53	0,06	50,51	17,66	1468
UNETR_64_wCE_DA	0,71	0,04	1,92	0,51	0,56	0,05	51,60	14,70	1523
DynUNET_96_wCE_DA	0,72	0,04	1,65	0,47	0,57	0,05	52,74	14,41	1543
UNETR_128_wCE_DA	0,71	0,04	1,93	0,65	0,56	0,05	53,11	14,81	1566
DynUNET_64_wCE_DA	0,72	0,04	1,73	0,42	0,56	0,05	53,74	15,46	1598
UNETR_96_wCE_DA	0,71	0,05	2,00	0,58	0,55	0,06	54,49	16,67	1633
VNET_64_CE_DA	0,41	0,10	15,60	6,91	0,27	0,08	55,25	17,37	1768
3DUNET_128_wCE_DA	0,69	0,05	2,31	1,49	0,53	0,06	56,85	19,73	1673
3DUNET_64_wCE_noDA	0,69	0,05	1,86	0,50	0,53	0,06	57,64	19,42	1670
3DUNET_96_wCE_DA	0,68	0,05	1,98	0,53	0,52	0,06	67,55	17,67	1961
VNET_96_wCE_noDA	0,60	0,06	2,70	1,08	0,43	0,06	80,02	35,72	2373
VNET_64_wCE_noDA	0,63	0,05	2,47	0,72	0,46	0,05	80,67	30,08	2356
VNET_128_wCE_noDA	0,57	0,07	3,13	1,90	0,40	0,07	85,11	43,61	2585
VNET_96_wCE_DA	0,58	0,06	2,94	1,39	0,41	0,06	101,20	42,80	2990
VNET_64_wCE_DA	0,55	0,06	3,04	0,53	0,38	0,06	112,82	48,88	3351

Appendix B

VNET_128_wCE_DA	0,55	0,06	3,31	1,82	0,38	0,05	114,79	42,91	3353
VNET_128_Dice_noDA	0,22	0,19	92,78	52,35	0,14	0,13	2177,04	2699,50	93091
VNET_128_Dice_DA	0,20	0,18	110,27	38,39	0,12	0,12	2472,75	3336,03	111559
VNET_96_Dice_noDA	0,01	0,00	129,18	9,51	0,00	0,00	15605,3	9801,13	497015
VNET_96_Dice_DA	0,05	0,10	125,45	30,64	0,03	0,06	20547,2	16236,4	749952
VNET_64_Dice_noDA	0,00	0,00	158,56	5,17	0,00	0,00	53830,9	24123,1	1685820
VNET_64_Dice_DA	0,00	0,00	161,94	3,99	0,00	0,00	124102	30142,7	3576246

Table B 2: Metrics on validation set for DNN trained over FLAIR images with GT cT1-w MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_64_CE_DA	0,75	0,05	1,84	0,97	0,60	0,07	8,66	7,91	344
3DUNET_64_DiceCE_DA	0,73	0,05	1,79	0,73	0,58	0,06	9,40	6,08	330
DynUNET_96_Dice_DA	0,75	0,05	1,93	1,22	0,61	0,06	9,41	7,66	357
DynUNET_64_Dice_DA	0,74	0,05	1,95	0,90	0,59	0,06	9,75	7,32	369
DynUNET_128_Dice_noDA	0,75	0,05	1,72	0,69	0,60	0,06	9,92	7,58	362
UNETR_128_DiceCE_DA	0,72	0,06	2,00	1,06	0,57	0,07	9,92	5,45	337
UNETR_128_Dice_DA	0,73	0,05	1,86	0,69	0,58	0,06	9,95	7,50	360
DynUNET_64_Dice_noDA	0,75	0,05	9,97	29,97	0,60	0,07	10,13	7,05	361
DynUNET_96_Dice_noDA	0,75	0,05	1,80	1,12	0,61	0,06	10,22	8,21	376
3DUNET_64_Dice_DA	0,73	0,04	9,21	27,11	0,57	0,05	10,28	8,67	376
DynUNET_96_DiceCE_noDA	0,74	0,05	1,87	0,79	0,59	0,06	10,49	8,94	397
DynUNET_96_CE_DA	0,75	0,05	1,76	0,81	0,60	0,07	10,86	8,10	376
DynUNET_96_DiceCE_DA	0,74	0,06	2,14	1,27	0,59	0,07	10,89	8,09	393
3DUNET_96_Dice_DA	0,74	0,05	1,75	0,62	0,59	0,06	11,04	9,74	388
UNETR_96_CE_DA	0,73	0,06	1,85	0,72	0,58	0,07	11,22	7,71	391
3DUNET_96_DiceCE_DA	0,74	0,04	2,04	0,92	0,59	0,05	11,26	7,76	377
UNETR_64_Dice_noDA	0,72	0,07	2,44	2,61	0,57	0,08	11,41	10,18	435
DynUNET_128_CE_DA	0,75	0,06	2,19	2,15	0,60	0,07	11,48	9,14	406
UNETR_96_Dice_noDA	0,72	0,06	2,00	0,78	0,57	0,07	11,58	6,74	364
UNETR_128_CE_DA	0,72	0,06	2,15	1,18	0,57	0,07	11,75	7,47	386
3DUNET_128_Dice_DA	0,73	0,05	1,80	0,73	0,58	0,06	11,76	9,73	400
DynUNET_128_DiceCE_noDA	0,73	0,06	2,02	1,30	0,58	0,08	11,79	10,11	473
DynUNET_128_DiceCE_DA	0,73	0,06	2,30	2,05	0,58	0,07	12,04	10,03	470
DynUNET_128_wCE_noDA	0,73	0,05	1,73	0,64	0,57	0,06	12,10	8,70	393
DynUNET_128_Dice_DA	0,76	0,05	1,81	1,06	0,61	0,06	12,10	8,91	415
UNETR_128_Dice_noDA	0,70	0,07	2,61	2,10	0,54	0,08	12,32	11,15	462
3DUNET_64_DiceCE_noDA	0,72	0,07	2,58	2,75	0,56	0,08	12,37	10,24	454
DynUNET_96_CE_noDA	0,74	0,06	1,88	0,97	0,59	0,08	12,47	11,74	510
DynUNET_64_CE_noDA	0,73	0,06	2,21	1,84	0,58	0,07	12,62	8,30	465
UNETR_64_Dice_DA	0,71	0,06	2,33	0,74	0,55	0,07	12,67	9,93	483
DynUNET_128_CE_noDA	0,73	0,06	2,01	1,06	0,58	0,07	12,83	9,60	472
3DUNET_128_DiceCE_DA	0,72	0,06	2,37	1,92	0,57	0,07	12,84	9,68	452
3DUNET_128_CE_DA	0,73	0,05	2,22	1,59	0,58	0,06	12,93	9,07	450
3DUNET_64_Dice_noDA	0,73	0,05	1,85	0,74	0,58	0,06	12,98	8,55	421
3DUNET_96_DiceCE_noDA	0,72	0,05	1,98	0,91	0,57	0,06	13,23	8,81	465

Appendix B

3DUNET_128_Dice_noDA	0,72	0,05	1,95	0,67	0,57	0,06	13,43	8,35	412
3DUNET_96_Dice_noDA	0,73	0,05	1,92	0,79	0,58	0,06	13,46	9,15	436
DynUNET_64_DiceCE_DA	0,72	0,06	2,27	0,98	0,56	0,07	13,55	10,15	515
3DUNET_96_CE_noDA	0,72	0,07	2,39	2,07	0,57	0,08	13,68	11,85	522
UNETR_96_DiceCE_noDA	0,71	0,09	2,68	2,53	0,56	0,10	13,69	13,70	538
3DUNET_64_CE_DA	0,73	0,06	2,06	0,91	0,57	0,07	13,73	9,70	483
UNETR_96_Dice_DA	0,74	0,06	1,92	0,97	0,59	0,07	13,76	9,58	476
UNETR_128_DiceCE_noDA	0,69	0,09	2,75	2,58	0,53	0,10	13,85	12,96	556
3DUNET_96_CE_DA	0,73	0,06	2,49	2,39	0,57	0,07	14,44	10,52	525
UNETR_64_CE_noDA	0,72	0,08	2,44	1,86	0,56	0,09	14,56	14,79	595
UNETR_64_DiceCE_noDA	0,72	0,08	2,58	2,47	0,56	0,09	14,78	14,02	576
UNETR_96_CE_noDA	0,70	0,09	3,03	3,14	0,54	0,10	14,92	15,11	605
UNETR_128_CE_noDA	0,67	0,09	3,28	2,87	0,51	0,09	15,33	14,44	632
UNETR_96_DiceCE_DA	0,73	0,06	1,84	0,82	0,58	0,07	15,48	10,44	527
UNETR_64_DiceCE_DA	0,72	0,07	2,18	0,83	0,56	0,08	15,59	11,17	562
3DUNET_128_DiceCE_noDA	0,70	0,05	1,99	0,65	0,54	0,06	15,60	9,58	487
3DUNET_64_CE_noDA	0,72	0,07	2,73	2,91	0,56	0,08	15,91	13,27	612
UNETR_64_CE_DA	0,72	0,07	2,23	1,14	0,57	0,08	16,27	11,90	580
DynUNET_64_DiceCE_noDA	0,72	0,06	2,51	1,14	0,56	0,07	16,33	11,87	616
3DUNET_128_CE_noDA	0,70	0,08	2,46	1,62	0,54	0,09	16,71	14,00	649
DynUNET_96_wCE_noDA	0,73	0,05	1,75	0,75	0,58	0,06	18,77	13,22	604
VNET_96_CE_DA	0,69	0,10	4,82	6,55	0,53	0,11	20,64	19,87	820
VNET_128_CE_noDA	0,66	0,13	3,77	3,81	0,50	0,13	20,77	22,01	881
VNET_128_DiceCE_DA	0,66	0,09	4,04	3,26	0,50	0,09	20,88	17,50	800
VNET_64_DiceCE_DA	0,67	0,12	4,38	6,62	0,52	0,12	21,40	19,70	826
VNET_96_CE_noDA	0,62	0,11	14,17	33,99	0,46	0,11	22,15	19,32	870
UNETR_128_wCE_noDA	0,69	0,06	2,33	1,14	0,53	0,07	24,48	13,33	762
VNET_64_DiceCE_noDA	0,63	0,15	14,74	37,37	0,48	0,14	25,46	23,10	1019
VNET_128_CE_DA	0,62	0,17	19,62	34,61	0,46	0,15	25,60	26,02	1068
VNET_128_DiceCE_noDA	0,60	0,18	6,04	7,60	0,44	0,15	27,37	26,08	1078
VNET_96_DiceCE_DA	0,57	0,16	126,66	86,76	0,41	0,13	27,74	22,01	1011
DynUNET_64_wCE_noDA	0,73	0,04	2,01	1,75	0,58	0,06	29,03	11,70	864
UNETR_96_wCE_noDA	0,70	0,06	2,60	2,35	0,54	0,07	30,40	15,41	940
UNETR_64_wCE_DA	0,71	0,05	1,95	0,74	0,55	0,06	30,91	17,00	975
UNETR_128_wCE_DA	0,70	0,05	2,34	2,08	0,55	0,06	33,45	17,53	1067
UNETR_64_wCE_noDA	0,71	0,04	2,14	1,38	0,55	0,05	37,89	20,00	1193
UNETR_96_wCE_DA	0,70	0,05	2,12	0,98	0,54	0,06	38,38	19,91	1212
VNET_64_CE_noDA	0,56	0,21	9,65	10,94	0,41	0,17	39,24	28,00	1393
VNET_96_DiceCE_noDA	0,51	0,19	57,16	62,59	0,36	0,14	39,55	23,90	1384
DynUNET_128_wCE_DA	0,72	0,05	1,94	1,00	0,56	0,06	40,71	13,02	1210
3DUNET_128_wCE_noDA	0,69	0,04	2,22	0,89	0,53	0,05	40,96	20,51	1229
VNET_64_CE_DA	0,55	0,17	37,33	63,04	0,40	0,14	41,63	22,78	1411
3DUNET_96_wCE_noDA	0,70	0,05	2,24	1,06	0,54	0,06	42,77	20,51	1290
DynUNET_64_wCE_DA	0,72	0,04	2,01	1,01	0,56	0,05	42,91	16,05	1291
3DUNET_64_wCE_noDA	0,70	0,04	2,29	1,62	0,54	0,05	43,50	21,60	1330
3DUNET_128_wCE_DA	0,70	0,05	1,94	0,65	0,55	0,06	47,57	18,15	1384

Appendix B

3DUNET_96_wCE_DA	0,70	0,04	1,98	0,83	0,54	0,05	50,37	18,13	1484
DynUNET_96_wCE_DA	0,71	0,04	1,76	0,54	0,56	0,05	55,23	13,62	1621
VNET_96_wCE_noDA	0,64	0,06	2,91	2,16	0,47	0,06	55,95	23,73	1711
VNET_128_wCE_noDA	0,64	0,06	2,50	0,89	0,47	0,06	56,56	23,44	1658
3DUNET_64_wCE_DA	0,70	0,04	1,95	0,46	0,54	0,05	57,48	15,78	1664
VNET_64_wCE_noDA	0,66	0,04	2,35	1,14	0,50	0,05	61,48	27,76	1869
VNET_96_wCE_DA	0,65	0,05	2,24	0,49	0,49	0,05	69,50	24,24	2066
VNET_64_wCE_DA	0,61	0,06	2,56	0,61	0,45	0,06	82,53	43,22	2604
VNET_128_wCE_DA	0,59	0,05	2,88	0,54	0,42	0,05	107,08	37,25	3206
VNET_64_Dice_DA	0,41	0,28	50,30	48,36	0,30	0,22	1759,00	3180,07	106126
VNET_96_Dice_noDA	0,08	0,16	99,41	30,66	0,05	0,11	24051,6	22581,3	939764
VNET_96_Dice_DA	0,08	0,17	119,58	21,14	0,05	0,12	36454,2	35646,1	1433891
VNET_128_Dice_noDA	0,03	0,08	130,25	13,67	0,02	0,05	37493,8	37093,2	1471086
VNET_128_Dice_DA	0,00	0,00	135,84	6,93	0,00	0,00	78709,7	18423,3	2305780
VNET_64_Dice_noDA	0,00	0,00	159,38	3,92	0,00	0,00	110436	29265,3	3196278

Table B 3: Metrics on validation set for DNN trained over 2 Inputs images (merging of T1-w and FLAIR) with GT cT1-w MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_128_DiceCE_DA	0,76	0,04	1,50	0,38	0,61	0,05	8,23	5,81	322
DynUNET_128_Dice_noDA	0,76	0,04	1,58	0,56	0,62	0,06	8,34	5,92	313
UNETR_128_DiceCE_DA	0,75	0,05	1,73	0,74	0,61	0,06	8,34	5,33	288
DynUNET_96_DiceCE_noDA	0,75	0,05	1,66	0,55	0,60	0,06	8,47	6,23	313
DynUNET_96_Dice_DA	0,77	0,04	1,56	0,72	0,62	0,05	8,62	6,33	329
DynUNET_64_CE_noDA	0,76	0,04	1,79	1,33	0,62	0,06	8,75	5,22	309
DynUNET_128_Dice_DA	0,76	0,04	1,46	0,38	0,62	0,05	9,08	6,75	336
UNETR_128_Dice_noDA	0,74	0,05	1,69	0,39	0,59	0,06	9,25	5,06	307
DynUNET_64_CE_DA	0,76	0,05	1,52	0,65	0,62	0,06	9,37	5,87	325
UNETR_96_DiceCE_noDA	0,74	0,05	2,00	1,65	0,59	0,07	9,39	6,22	324
DynUNET_96_CE_DA	0,76	0,05	1,68	1,33	0,61	0,06	9,43	6,07	351
UNETR_96_Dice_noDA	0,74	0,06	1,92	1,36	0,59	0,07	9,44	6,77	355
UNETR_128_DiceCE_noDA	0,73	0,06	1,76	0,58	0,58	0,07	9,55	5,73	308
3DUNET_128_DiceCE_DA	0,74	0,05	1,61	0,39	0,59	0,06	9,67	7,46	330
3DUNET_64_DiceCE_DA	0,74	0,05	1,74	0,50	0,58	0,06	9,67	4,52	332
UNETR_64_DiceCE_noDA	0,75	0,05	1,86	1,16	0,60	0,06	9,77	6,25	344
DynUNET_128_CE_DA	0,77	0,04	1,48	0,43	0,62	0,06	9,91	6,61	333
DynUNET_128_CE_noDA	0,76	0,05	1,62	0,72	0,61	0,06	9,93	5,47	342
DynUNET_96_DiceCE_DA	0,75	0,05	1,64	0,56	0,60	0,06	9,99	7,84	372
DynUNET_96_Dice_noDA	0,76	0,05	2,06	1,84	0,62	0,06	10,00	7,53	366
3DUNET_64_Dice_DA	0,75	0,05	1,62	0,35	0,60	0,06	10,04	7,93	342
UNETR_96_DiceCE_DA	0,76	0,04	1,57	0,39	0,61	0,05	10,07	6,95	346
3DUNET_64_Dice_noDA	0,75	0,04	1,65	0,48	0,60	0,05	10,44	6,46	339
UNETR_96_CE_DA	0,75	0,05	2,05	1,51	0,60	0,06	10,46	7,08	391
UNETR_64_DiceCE_DA	0,75	0,04	1,70	0,48	0,60	0,05	10,68	8,56	381
3DUNET_96_DiceCE_DA	0,74	0,05	1,64	0,55	0,59	0,06	10,68	6,08	346

Appendix B

DynUNET_96_CE_noDA	0,76	0,04	1,59	0,60	0,61	0,05	10,69	6,36	391
DynUNET_128_DiceCE_noDA	0,74	0,05	2,24	1,36	0,59	0,06	10,73	7,83	437
3DUNET_128_Dice_DA	0,75	0,05	1,60	0,48	0,60	0,06	10,79	8,03	361
3DUNET_96_CE_DA	0,74	0,05	1,71	0,56	0,59	0,07	11,04	4,57	343
3DUNET_96_Dice_DA	0,76	0,05	1,50	0,46	0,61	0,06	11,16	8,81	372
UNETR_128_CE_noDA	0,73	0,06	1,81	0,56	0,57	0,07	11,27	7,53	369
3DUNET_128_CE_DA	0,74	0,05	1,96	1,08	0,59	0,06	11,34	7,34	408
UNETR_128_CE_DA	0,75	0,05	2,07	1,78	0,60	0,06	11,36	7,90	419
UNETR_64_Dice_noDA	0,76	0,04	1,59	0,53	0,61	0,06	11,53	7,97	382
UNETR_96_Dice_DA	0,75	0,05	11,31	35,52	0,60	0,07	11,53	7,73	395
3DUNET_128_Dice_noDA	0,73	0,05	1,76	0,49	0,57	0,06	11,55	5,80	386
DynUNET_64_DiceCE_noDA	0,75	0,06	2,05	1,55	0,60	0,07	11,73	8,22	435
3DUNET_128_DiceCE_noDA	0,72	0,06	1,76	0,49	0,56	0,07	11,76	6,66	407
UNETR_64_CE_DA	0,75	0,05	1,64	0,66	0,61	0,06	11,96	5,79	404
UNETR_128_Dice_DA	0,76	0,04	1,76	0,80	0,61	0,06	11,98	7,82	391
3DUNET_64_CE_DA	0,74	0,06	1,79	0,64	0,59	0,07	11,98	5,54	398
UNETR_64_CE_noDA	0,74	0,05	2,03	1,19	0,59	0,06	12,05	7,69	452
UNETR_96_CE_noDA	0,73	0,06	1,99	1,27	0,58	0,07	12,19	7,74	388
3DUNET_96_DiceCE_noDA	0,73	0,05	1,85	0,67	0,58	0,06	12,27	7,20	411
3DUNET_128_CE_noDA	0,71	0,06	2,10	1,18	0,56	0,07	12,40	8,49	460
3DUNET_96_Dice_noDA	0,74	0,05	2,24	2,00	0,59	0,06	12,71	7,70	439
DynUNET_128_wCE_noDA	0,73	0,04	1,59	0,39	0,58	0,05	12,79	9,20	414
3DUNET_64_CE_noDA	0,74	0,06	1,96	1,36	0,58	0,07	12,83	7,44	438
3DUNET_96_CE_noDA	0,73	0,06	2,17	1,53	0,57	0,07	13,11	8,22	469
3DUNET_64_DiceCE_noDA	0,74	0,05	1,74	0,75	0,59	0,06	13,91	10,03	450
DynUNET_64_Dice_noDA	0,70	0,06	69,78	61,08	0,54	0,07	14,42	10,36	524
DynUNET_64_Dice_DA	0,72	0,05	92,02	49,76	0,56	0,07	14,74	14,67	602
DynUNET_96_wCE_noDA	0,75	0,05	1,77	1,30	0,60	0,06	15,30	10,70	503
DynUNET_64_DiceCE_DA	0,71	0,05	3,07	1,46	0,56	0,06	15,80	8,56	582
UNETR_64_Dice_DA	0,75	0,05	34,53	56,71	0,60	0,07	16,90	11,75	554
VNET_96_DiceCE_DA	0,59	0,09	20,92	32,98	0,42	0,08	21,36	16,54	778
UNETR_96_wCE_noDA	0,73	0,05	1,65	0,52	0,58	0,06	22,53	12,85	694
UNETR_128_wCE_noDA	0,72	0,05	1,99	1,13	0,57	0,06	23,14	12,59	718
VNET_128_DiceCE_noDA	0,60	0,10	34,86	54,10	0,44	0,10	24,15	17,03	920
VNET_128_CE_DA	0,63	0,09	5,08	3,16	0,47	0,09	25,54	12,90	894
VNET_128_wCE_DA	0,62	0,08	6,53	6,36	0,46	0,08	25,58	19,68	940
VNET_128_CE_noDA	0,61	0,12	7,69	6,30	0,45	0,11	27,64	19,36	1018
VNET_64_DiceCE_noDA	0,58	0,11	43,47	61,22	0,42	0,10	27,83	15,99	1000
VNET_64_CE_noDA	0,62	0,12	15,10	35,91	0,46	0,11	28,53	17,25	1039
DynUNET_64_wCE_noDA	0,74	0,04	1,71	0,74	0,59	0,05	29,02	13,85	889
VNET_128_DiceCE_DA	0,57	0,10	22,67	31,32	0,41	0,09	29,85	16,29	1042
VNET_64_DiceCE_DA	0,56	0,10	38,89	52,67	0,39	0,09	33,05	23,41	1138
3DUNET_96_wCE_noDA	0,72	0,05	1,94	0,82	0,57	0,06	33,13	19,28	1028
3DUNET_128_wCE_noDA	0,71	0,04	1,84	0,60	0,55	0,05	33,21	16,17	992
UNETR_64_wCE_noDA	0,73	0,04	1,69	0,63	0,58	0,05	38,00	14,42	1126
VNET_96_DiceCE_noDA	0,50	0,10	59,71	56,91	0,34	0,09	38,07	19,83	1329

Appendix B

3DUNET_64_wCE_noDA	0,72	0,04	1,74	0,50	0,56	0,05	40,72	19,42	1232
VNET_96_wCE_DA	0,63	0,08	24,36	47,00	0,47	0,08	41,38	22,17	1270
UNETR_128_wCE_DA	0,72	0,04	1,94	1,28	0,57	0,05	43,35	15,92	1290
VNET_64_CE_DA	0,48	0,13	59,92	60,60	0,33	0,11	44,91	16,70	1452
VNET_96_CE_noDA	0,56	0,12	10,24	7,23	0,40	0,11	45,19	12,28	1398
DynUNET_128_wCE_DA	0,73	0,05	1,60	0,48	0,57	0,06	45,33	15,67	1335
VNET_96_wCE_noDA	0,66	0,06	2,80	1,45	0,50	0,06	48,22	22,89	1449
3DUNET_64_wCE_DA	0,72	0,04	1,65	0,46	0,56	0,05	48,89	18,40	1438
3DUNET_128_wCE_DA	0,71	0,05	1,71	0,55	0,55	0,06	49,29	15,87	1433
UNETR_96_wCE_DA	0,72	0,04	1,68	0,57	0,56	0,05	50,19	15,50	1470
VNET_64_wCE_DA	0,65	0,05	2,34	0,51	0,48	0,05	51,48	20,64	1548
VNET_96_CE_DA	0,46	0,14	36,17	43,97	0,31	0,11	51,94	16,28	1645
3DUNET_96_wCE_DA	0,71	0,05	1,76	0,56	0,56	0,06	52,53	16,51	1528
UNETR_64_wCE_DA	0,71	0,04	1,81	0,56	0,55	0,05	54,11	17,43	1585
DynUNET_96_wCE_DA	0,72	0,04	1,66	0,54	0,56	0,05	54,14	13,50	1561
VNET_64_wCE_noDA	0,65	0,06	12,27	36,44	0,49	0,06	56,27	23,07	1656
DynUNET_64_wCE_DA	0,71	0,04	13,23	42,93	0,55	0,05	60,91	15,47	1776
VNET_128_wCE_noDA	0,62	0,06	2,79	0,84	0,45	0,06	76,11	35,95	2294
VNET_128_Dice_noDA	0,23	0,25	100,75	46,50	0,15	0,19	9658,64	13272,96	469307
VNET_128_Dice_DA	0,20	0,26	100,89	43,97	0,14	0,19	10130,22	13294,22	469056
VNET_96_Dice_DA	0,01	0,02	130,42	14,53	0,01	0,01	24499,26	19079,00	825575
VNET_96_Dice_noDA	0,00	0,00	141,37	3,77	0,00	0,00	59173,24	19769,19	1721963
VNET_64_Dice_noDA	0,00	0,00	153,70	3,40	0,00	0,00	162433,1	40130,07	4659140
VNET_64_Dice_DA	0,00	0,00	150,56	5,93	0,00	0,00	179706,3	40151,78	5124270

B.3 DNN VALIDATION SET RESULTS: TRAINING WITH GROUND TRUTH WITHOUT CONTRAST, PERFORMANCE INDICES CALCULATED WITH RESPECT TO THE CONTRAST GROUND TRUTH

The performance indices are calculated making the comparison between the predicted segmentation obtained with each DNN architecture, trained with the ground truth manual segmentation obtained from the images without contrast (T1-w, FLAIR), and the gold-standard manual segmentation (cT1-w). Four are the input – MSeg examined combinations: T1-w – T1-w; FLAIR – FLAIR; T1-w+FLAIR – T1-w; T1-w+FLAIR - FLAIR. Considering the bad performance obtained with the V-Net, it was excluded from this comparison.

Appendix B

Table B 4: Metrics on validation set for DNN trained over T1-w images with GT T1-w MSeg, reference cT1-w gold-standard MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_CE_DA	0,71	0,04	3,48	1,87	0,55	0,05	7,40	5,91	309
DynUNET_128_CE_DA	0,70	0,04	3,09	1,12	0,55	0,04	7,93	5,43	322
3DUNET_64_DiceCE_DA	0,67	0,04	3,41	2,24	0,51	0,05	8,22	6,03	331
DynUNET_96_DiceCE_noDA	0,70	0,04	3,73	1,59	0,54	0,05	8,29	6,24	341
DynUNET_64_CE_noDA	0,69	0,04	3,33	1,37	0,53	0,05	8,44	5,63	329
UNETR_96_CE_DA	0,70	0,04	2,80	1,28	0,54	0,05	8,47	4,11	281
UNETR_96_DiceCE_noDA	0,69	0,05	4,03	2,86	0,53	0,06	8,47	5,85	292
UNETR_96_DiceCE_DA	0,70	0,04	3,57	2,47	0,54	0,05	8,61	6,67	324
UNETR_64_CE_DA	0,70	0,05	2,70	0,41	0,54	0,05	8,61	6,85	331
DynUNET_96_Dice_noDA	0,70	0,04	2,86	0,97	0,54	0,05	8,68	6,60	347
UNETR_128_DiceCE_noDA	0,67	0,05	2,90	0,76	0,50	0,05	8,90	6,15	322
DynUNET_128_CE_noDA	0,69	0,04	2,90	0,87	0,53	0,05	8,94	6,47	324
DynUNET_128_DiceCE_DA	0,71	0,04	3,05	0,98	0,55	0,05	8,99	7,45	365
DynUNET_64_CE_DA	0,70	0,04	3,05	1,17	0,54	0,05	9,08	7,22	366
3DUNET_96_CE_noDA	0,68	0,05	3,82	2,34	0,51	0,06	9,19	6,57	340
DynUNET_96_CE_noDA	0,69	0,05	2,79	0,89	0,53	0,05	9,28	5,37	326
UNETR_96_Dice_DA	0,69	0,04	20,09	42,87	0,53	0,05	9,29	5,83	357
UNETR_128_CE_DA	0,69	0,05	3,13	1,14	0,53	0,05	9,52	4,57	320
DynUNET_64_Dice_noDA	0,70	0,04	2,97	1,05	0,54	0,05	9,85	8,60	393
UNETR_64_Dice_DA	0,70	0,04	3,05	1,04	0,53	0,05	10,02	8,33	384
DynUNET_128_DiceCE_noDA	0,70	0,04	2,80	0,83	0,54	0,05	10,21	6,89	378
UNETR_96_CE_noDA	0,68	0,05	3,16	1,53	0,52	0,06	10,25	5,17	339
3DUNET_96_DiceCE_noDA	0,67	0,05	3,45	1,95	0,51	0,05	10,39	6,36	359
DynUNET_96_DiceCE_DA	0,70	0,04	4,32	1,90	0,53	0,05	11,01	7,17	431
3DUNET_128_CE_noDA	0,65	0,06	3,73	1,88	0,49	0,06	11,33	7,79	398
3DUNET_128_Dice_DA	0,68	0,04	2,86	0,90	0,52	0,05	11,49	7,08	388
3DUNET_64_DiceCE_noDA	0,67	0,04	3,09	1,33	0,50	0,04	11,67	7,71	432
UNETR_64_DiceCE_DA	0,69	0,05	3,15	1,08	0,53	0,05	11,77	8,00	447
3DUNET_128_CE_DA	0,67	0,05	3,48	1,34	0,51	0,06	11,79	8,11	393
DynUNET_64_DiceCE_noDA	0,69	0,04	3,12	0,68	0,52	0,05	11,95	7,72	446
UNETR_128_CE_noDA	0,67	0,05	3,06	1,01	0,50	0,06	12,03	8,79	408
3DUNET_64_CE_noDA	0,68	0,05	3,70	2,38	0,51	0,05	12,08	8,14	452
3DUNET_96_Dice_noDA	0,67	0,05	3,77	2,15	0,50	0,05	12,19	7,87	485
3DUNET_64_Dice_noDA	0,67	0,04	3,39	1,80	0,51	0,05	12,33	7,91	422
3DUNET_64_Dice_DA	0,68	0,04	11,23	31,19	0,52	0,04	12,36	10,96	449
UNETR_64_CE_noDA	0,68	0,05	3,20	1,46	0,52	0,05	12,37	7,72	442
UNETR_64_DiceCE_noDA	0,68	0,05	3,90	2,62	0,52	0,06	12,39	7,29	457
3DUNET_96_CE_DA	0,68	0,05	4,03	2,31	0,52	0,05	12,43	7,13	444
DynUNET_128_Dice_noDA	0,70	0,05	3,18	1,16	0,54	0,05	12,52	10,73	476
3DUNET_128_DiceCE_noDA	0,64	0,05	4,03	2,20	0,48	0,05	12,75	9,32	433
DynUNET_128_Dice_DA	0,71	0,04	2,99	0,97	0,55	0,05	12,92	10,40	478
3DUNET_96_Dice_DA	0,68	0,04	13,09	35,29	0,52	0,04	13,13	8,68	436
UNETR_128_DiceCE_DA	0,70	0,04	2,52	0,67	0,54	0,05	13,19	8,34	422

Appendix B

3DUNET_96_DiceCE_DA	0,68	0,04	3,92	2,14	0,52	0,05	13,30	8,35	435
DynUNET_64_DiceCE_DA	0,68	0,04	3,64	0,98	0,51	0,04	13,41	7,93	511
UNETR_96_Dice_noDA	0,70	0,04	3,58	2,65	0,53	0,05	14,20	8,88	444
3DUNET_64_CE_DA	0,67	0,05	4,29	2,27	0,51	0,06	14,53	8,34	511
UNETR_128_Dice_noDA	0,68	0,05	2,90	0,77	0,52	0,06	14,65	10,83	479
3DUNET_128_DiceCE_DA	0,66	0,04	3,40	1,34	0,50	0,05	15,03	8,96	475
DynUNET_128_wCE_noDA	0,68	0,05	3,00	1,19	0,52	0,05	15,35	11,43	505
DynUNET_96_Dice_DA	0,70	0,04	21,53	47,65	0,54	0,05	17,46	11,63	596
UNETR_128_Dice_DA	0,69	0,04	2,51	0,79	0,53	0,05	19,00	11,71	605
DynUNET_96_wCE_noDA	0,68	0,04	3,21	1,36	0,52	0,05	21,88	11,90	670
UNETR_64_Dice_noDA	0,67	0,04	100,48	68,38	0,50	0,05	22,77	15,32	722
3DUNET_128_Dice_noDA	0,62	0,05	4,53	1,82	0,45	0,05	23,85	12,41	870
UNETR_64_wCE_DA	0,68	0,04	2,63	1,04	0,51	0,04	27,90	16,77	901
UNETR_128_wCE_noDA	0,66	0,05	3,40	1,76	0,49	0,06	31,25	13,51	917
DynUNET_64_Dice_DA	0,66	0,06	63,57	62,99	0,49	0,06	32,81	23,96	1088
DynUNET_64_wCE_noDA	0,67	0,04	2,66	1,03	0,51	0,05	33,56	14,76	1009
UNETR_96_wCE_noDA	0,66	0,05	3,32	1,36	0,50	0,05	36,47	16,11	1073
UNETR_96_wCE_DA	0,67	0,04	2,80	0,58	0,51	0,04	43,31	14,15	1281
UNETR_64_wCE_noDA	0,67	0,04	2,61	0,65	0,51	0,05	43,77	18,20	1293
DynUNET_128_wCE_DA	0,67	0,04	3,05	1,07	0,51	0,05	45,14	16,43	1359
3DUNET_128_wCE_noDA	0,63	0,05	3,00	0,73	0,46	0,06	45,62	21,92	1351
3DUNET_96_wCE_DA	0,66	0,04	2,91	0,70	0,49	0,05	49,32	15,65	1434
3DUNET_96_wCE_noDA	0,64	0,05	2,77	0,62	0,48	0,05	50,15	18,66	1455
UNETR_128_wCE_DA	0,66	0,04	3,07	0,91	0,49	0,05	51,27	17,26	1501
3DUNET_64_wCE_noDA	0,64	0,04	3,11	0,92	0,48	0,05	57,95	20,34	1687
3DUNET_128_wCE_DA	0,64	0,05	3,15	0,85	0,47	0,05	61,29	17,71	1769
DynUNET_96_wCE_DA	0,66	0,04	3,19	0,96	0,49	0,05	63,01	15,66	1838
DynUNET_64_wCE_DA	0,65	0,04	3,68	1,12	0,48	0,05	65,11	20,42	1933
3DUNET_64_wCE_DA	0,63	0,05	3,41	1,19	0,46	0,05	70,00	24,16	2032

Table B 5: Metrics on validation set for DNN trained over FLAIR images with GT FLAIR MSeg, reference cT1-w gold-standard MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_DiceCE_DA	0,70	0,06	3,54	2,63	0,54	0,07	11,78	9,73	420
UNETR_64_CE_noDA	0,69	0,06	2,91	2,10	0,53	0,07	14,65	10,78	510
UNETR_96_CE_DA	0,70	0,06	3,17	2,56	0,54	0,07	14,88	12,17	526
UNETR_64_DiceCE_DA	0,69	0,06	2,33	0,83	0,54	0,07	15,44	11,74	534
UNETR_64_Dice_DA	0,69	0,06	17,32	29,77	0,53	0,07	15,80	11,03	535
3DUNET_64_DiceCE_DA	0,71	0,05	2,14	0,98	0,56	0,06	16,75	10,98	550
3DUNET_64_Dice_DA	0,70	0,05	21,13	37,91	0,54	0,06	17,10	14,25	621
DynUNET_64_DiceCE_DA	0,72	0,05	2,10	1,19	0,56	0,06	17,50	11,64	589
UNETR_128_CE_noDA	0,68	0,06	2,50	0,98	0,52	0,07	18,00	7,98	549
DynUNET_64_CE_DA	0,72	0,05	2,16	1,04	0,56	0,06	18,16	11,76	591
DynUNET_128_CE_DA	0,71	0,05	25,40	46,24	0,55	0,06	18,88	11,88	622
3DUNET_64_CE_noDA	0,71	0,05	2,03	0,70	0,55	0,07	19,02	11,63	602

Appendix B

DynUNET_96_CE_noDA	0,71	0,05	1,89	0,52	0,56	0,06	19,18	9,84	600
3DUNET_64_DiceCE_noDA	0,70	0,05	2,74	3,01	0,54	0,06	19,20	11,43	615
UNETR_96_Dice_noDA	0,70	0,06	2,30	0,96	0,54	0,07	19,32	11,29	616
3DUNET_96_DiceCE_DA	0,72	0,05	2,19	0,94	0,56	0,06	19,79	11,84	626
DynUNET_64_DiceCE_noDA	0,71	0,05	2,31	1,31	0,55	0,07	19,96	12,82	643
UNETR_64_CE_DA	0,70	0,05	2,50	1,27	0,54	0,06	20,03	14,03	649
UNETR_96_DiceCE_DA	0,71	0,05	2,13	0,91	0,55	0,06	20,46	12,20	646
DynUNET_128_DiceCE_DA	0,71	0,07	2,36	1,62	0,55	0,08	20,67	11,44	651
3DUNET_64_CE_DA	0,71	0,05	1,91	0,60	0,55	0,06	20,89	10,77	646
3DUNET_96_DiceCE_noDA	0,71	0,05	2,03	0,79	0,55	0,06	20,95	9,80	632
DynUNET_96_DiceCE_noDA	0,71	0,05	2,01	0,67	0,56	0,07	21,14	13,68	698
UNETR_96_CE_noDA	0,69	0,07	2,76	1,85	0,53	0,08	21,48	10,54	653
3DUNET_128_CE_DA	0,71	0,06	2,45	1,84	0,55	0,07	22,09	12,94	693
DynUNET_96_Dice_noDA	0,72	0,05	1,96	0,49	0,56	0,06	22,20	10,81	694
DynUNET_64_CE_noDA	0,72	0,05	1,91	0,97	0,56	0,06	22,63	12,47	731
UNETR_96_Dice_DA	0,71	0,06	2,27	1,14	0,55	0,07	22,86	13,44	748
3DUNET_96_CE_DA	0,71	0,05	2,10	0,67	0,55	0,06	22,92	11,79	706
DynUNET_96_Dice_DA	0,72	0,05	2,10	1,05	0,57	0,06	23,22	12,02	742
3DUNET_128_DiceCE_DA	0,71	0,05	1,97	0,78	0,55	0,06	23,99	10,87	726
UNETR_128_DiceCE_DA	0,71	0,05	2,10	0,80	0,55	0,06	24,40	11,60	750
DynUNET_128_Dice_noDA	0,71	0,05	1,89	0,59	0,56	0,06	24,76	11,18	750
UNETR_128_DiceCE_noDA	0,67	0,07	2,84	1,57	0,51	0,08	24,97	12,11	758
UNETR_64_Dice_noDA	0,70	0,05	2,44	0,95	0,54	0,06	25,07	14,06	789
DynUNET_64_Dice_noDA	0,72	0,05	1,96	0,52	0,56	0,06	25,44	11,57	792
UNETR_128_CE_DA	0,70	0,05	2,32	0,80	0,54	0,06	25,60	11,32	783
3DUNET_128_DiceCE_noDA	0,69	0,06	2,27	1,04	0,53	0,07	26,06	13,80	803
DynUNET_128_DiceCE_noDA	0,71	0,04	1,91	0,47	0,55	0,05	26,18	14,06	816
3DUNET_128_Dice_noDA	0,70	0,05	2,13	0,72	0,54	0,06	27,04	15,19	854
3DUNET_128_CE_noDA	0,69	0,05	2,20	0,75	0,53	0,06	27,26	14,79	847
DynUNET_128_CE_noDA	0,71	0,05	2,03	0,75	0,55	0,06	27,40	13,46	836
UNETR_64_DiceCE_noDA	0,69	0,06	2,65	1,81	0,53	0,07	27,53	14,22	835
DynUNET_64_Dice_DA	0,71	0,05	15,76	34,50	0,56	0,06	27,65	14,18	880
DynUNET_128_Dice_DA	0,72	0,05	2,27	0,84	0,56	0,06	28,60	12,48	867
3DUNET_96_CE_noDA	0,70	0,06	2,21	0,97	0,54	0,07	29,21	14,96	901
DynUNET_96_CE_DA	0,71	0,05	2,42	1,10	0,55	0,06	30,05	12,13	896
UNETR_128_Dice_noDA	0,68	0,06	2,66	1,05	0,51	0,07	31,48	14,85	968
3DUNET_64_Dice_noDA	0,71	0,05	1,97	0,59	0,55	0,06	31,85	12,34	943
UNETR_96_DiceCE_noDA	0,68	0,06	2,71	1,56	0,52	0,07	33,40	16,19	1046
3DUNET_128_Dice_DA	0,70	0,06	2,17	0,86	0,54	0,07	35,48	13,40	1050
3DUNET_96_Dice_DA	0,70	0,06	20,72	47,59	0,54	0,07	38,15	15,69	1146
UNETR_128_Dice_DA	0,70	0,05	2,27	0,68	0,54	0,06	38,57	12,63	1153
3DUNET_96_Dice_noDA	0,70	0,05	2,11	0,73	0,54	0,06	39,19	15,59	1170
DynUNET_128_wCE_noDA	0,68	0,05	2,13	0,73	0,52	0,06	42,19	13,05	1235
DynUNET_96_wCE_noDA	0,69	0,04	1,97	0,50	0,53	0,05	45,38	12,34	1330
UNETR_128_wCE_noDA	0,65	0,06	2,55	0,63	0,48	0,06	52,65	19,24	1576
DynUNET_64_wCE_noDA	0,68	0,04	2,17	0,79	0,52	0,05	54,17	14,58	1598

Appendix B

UNETR_96_wCE_noDA	0,65	0,05	2,63	0,86	0,49	0,06	56,33	21,29	1715
UNETR_64_wCE_DA	0,65	0,04	2,55	0,93	0,48	0,05	65,11	24,24	1969
3DUNET_128_wCE_noDA	0,65	0,05	2,46	0,78	0,49	0,06	66,60	20,31	1945
UNETR_64_wCE_noDA	0,65	0,05	3,06	1,72	0,49	0,05	66,62	24,26	2007
3DUNET_96_wCE_noDA	0,66	0,05	2,48	0,85	0,49	0,06	69,51	22,43	2052
UNETR_128_wCE_DA	0,65	0,05	2,73	0,94	0,48	0,06	71,47	18,84	2107
UNETR_96_wCE_DA	0,65	0,05	2,84	1,01	0,48	0,05	72,89	17,63	2140
DynUNET_128_wCE_DA	0,66	0,05	2,25	0,62	0,49	0,06	73,74	17,97	2139
3DUNET_96_wCE_DA	0,65	0,05	2,57	0,85	0,49	0,06	74,64	16,92	2148
3DUNET_128_wCE_DA	0,65	0,05	2,57	1,04	0,49	0,06	76,15	19,22	2199
DynUNET_96_wCE_DA	0,66	0,05	2,65	1,37	0,49	0,06	78,59	18,63	2285
3DUNET_64_wCE_noDA	0,64	0,04	2,47	0,70	0,47	0,05	80,44	20,98	2344
3DUNET_64_wCE_DA	0,64	0,05	2,77	0,89	0,47	0,05	83,70	18,12	2407
DynUNET_64_wCE_DA	0,65	0,04	2,83	1,13	0,48	0,05	85,52	16,61	2486

Table B 6: Metrics on validation set for DNN trained over 2 Inputs images (merging of T1-w and FLAIR) with GT T1-w MSeg, reference cT1-w gold-standard MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_Dice_noDA	0,70	0,05	2,91	1,40	0,53	0,05	8,34	7,78	356
DynUNET_128_CE_noDA	0,69	0,04	2,80	0,97	0,53	0,05	8,47	7,47	353
DynUNET_96_CE_DA	0,70	0,05	2,49	0,65	0,54	0,05	8,55	6,36	315
UNETR_96_CE_DA	0,69	0,05	2,48	0,37	0,53	0,06	8,68	7,74	367
UNETR_64_DiceCE_DA	0,69	0,05	2,64	0,62	0,53	0,05	8,75	7,03	342
UNETR_64_DiceCE_noDA	0,69	0,05	3,85	2,61	0,53	0,06	8,94	7,29	340
3DUNET_96_DiceCE_DA	0,69	0,04	3,97	2,42	0,53	0,05	8,96	8,18	346
DynUNET_128_CE_DA	0,71	0,05	2,50	0,55	0,55	0,05	8,98	6,74	353
3DUNET_96_Dice_DA	0,69	0,05	3,07	0,93	0,53	0,05	9,21	6,72	346
3DUNET_128_Dice_DA	0,69	0,04	3,45	1,69	0,53	0,05	9,32	7,48	353
DynUNET_128_DiceCE_DA	0,70	0,04	3,02	1,73	0,55	0,05	9,34	6,97	361
DynUNET_64_CE_DA	0,71	0,04	13,65	41,58	0,55	0,05	9,38	6,85	372
UNETR_96_DiceCE_DA	0,70	0,04	2,17	0,50	0,54	0,05	9,67	6,56	338
UNETR_128_CE_DA	0,69	0,06	3,55	2,09	0,52	0,06	9,84	10,13	415
UNETR_64_CE_noDA	0,68	0,05	3,36	1,97	0,52	0,06	9,89	9,38	410
UNETR_64_CE_DA	0,69	0,05	2,78	0,95	0,53	0,06	9,93	9,43	423
DynUNET_64_DiceCE_noDA	0,70	0,04	2,49	0,77	0,54	0,05	9,95	9,23	387
UNETR_96_Dice_DA	0,70	0,05	2,66	0,76	0,54	0,05	9,98	7,18	370
UNETR_128_Dice_noDA	0,68	0,05	3,43	1,93	0,52	0,05	10,22	7,96	355
DynUNET_128_Dice_noDA	0,70	0,04	3,22	1,31	0,54	0,05	10,30	6,46	383
UNETR_96_Dice_noDA	0,69	0,05	3,53	1,99	0,53	0,05	10,53	8,38	391
DynUNET_96_DiceCE_noDA	0,69	0,06	3,30	1,78	0,53	0,06	10,64	8,36	412
3DUNET_96_CE_DA	0,69	0,05	3,67	1,36	0,53	0,06	10,69	8,04	420
3DUNET_128_CE_DA	0,70	0,04	2,83	1,03	0,53	0,05	10,82	9,52	392
3DUNET_64_CE_DA	0,69	0,05	2,92	0,90	0,53	0,06	10,86	9,77	414
UNETR_128_CE_noDA	0,67	0,05	4,24	2,26	0,51	0,06	10,92	9,15	434
DynUNET_96_Dice_DA	0,70	0,05	3,46	2,41	0,55	0,06	10,95	8,73	401

Appendix B

3DUNET_96_Dice_noDA	0,69	0,05	3,36	1,45	0,53	0,05	11,01	8,25	423
DynUNET_96_DiceCE_DA	0,70	0,05	3,54	1,46	0,54	0,06	11,09	9,85	449
UNETR_128_Dice_DA	0,69	0,04	3,15	1,07	0,53	0,05	11,14	7,15	368
UNETR_96_DiceCE_noDA	0,68	0,06	3,92	2,22	0,52	0,06	11,32	7,59	405
3DUNET_64_DiceCE_noDA	0,68	0,05	3,37	1,91	0,52	0,05	11,38	7,29	383
3DUNET_64_Dice_noDA	0,69	0,04	3,47	2,11	0,53	0,05	11,44	8,29	399
3DUNET_64_CE_noDA	0,68	0,05	3,57	2,09	0,52	0,06	11,49	8,01	413
3DUNET_128_Dice_noDA	0,68	0,05	3,47	1,91	0,52	0,05	11,61	8,14	408
3DUNET_128_DiceCE_DA	0,69	0,04	3,47	2,10	0,53	0,05	11,67	8,00	374
DynUNET_64_CE_noDA	0,69	0,04	2,60	0,73	0,53	0,05	11,73	8,88	403
3DUNET_96_DiceCE_noDA	0,68	0,05	3,46	2,00	0,52	0,05	11,83	8,39	455
UNETR_96_CE_noDA	0,67	0,06	3,80	2,50	0,51	0,07	11,85	9,69	449
3DUNET_64_DiceCE_DA	0,68	0,04	31,09	54,28	0,52	0,05	11,94	8,46	396
3DUNET_128_DiceCE_noDA	0,67	0,05	4,08	2,56	0,51	0,05	11,97	9,66	450
DynUNET_96_CE_noDA	0,68	0,05	3,22	1,90	0,52	0,06	12,09	9,09	465
3DUNET_128_CE_noDA	0,67	0,05	3,36	1,35	0,51	0,06	12,11	10,24	463
3DUNET_64_Dice_DA	0,70	0,04	2,82	0,94	0,54	0,05	12,31	9,21	419
DynUNET_128_Dice_DA	0,70	0,04	2,83	0,72	0,54	0,05	12,32	9,84	456
UNETR_64_Dice_noDA	0,70	0,04	2,86	1,44	0,54	0,05	12,47	8,90	422
DynUNET_128_DiceCE_noDA	0,70	0,05	3,47	1,73	0,54	0,06	12,53	6,83	430
3DUNET_96_CE_noDA	0,68	0,05	3,87	2,14	0,52	0,06	12,71	9,72	483
UNETR_128_DiceCE_DA	0,69	0,04	3,57	1,79	0,53	0,05	12,93	8,12	420
UNETR_64_Dice_DA	0,69	0,06	35,53	58,29	0,53	0,06	13,05	12,30	495
UNETR_128_DiceCE_noDA	0,67	0,06	3,28	1,16	0,51	0,07	13,51	9,93	462
DynUNET_64_Dice_DA	0,68	0,05	48,92	54,35	0,51	0,06	14,19	12,77	531
DynUNET_64_DiceCE_DA	0,68	0,05	8,35	2,11	0,51	0,06	15,19	9,55	577
DynUNET_64_Dice_noDA	0,67	0,05	76,72	68,54	0,50	0,05	15,83	11,31	522
DynUNET_96_wCE_noDA	0,69	0,04	2,37	0,85	0,53	0,05	20,12	13,30	658
DynUNET_128_wCE_noDA	0,68	0,04	3,22	1,82	0,52	0,04	22,50	14,34	722
UNETR_128_wCE_noDA	0,67	0,04	3,03	1,26	0,50	0,05	24,92	13,61	758
UNETR_96_wCE_noDA	0,67	0,05	2,80	0,80	0,50	0,05	29,35	13,70	878
DynUNET_64_wCE_noDA	0,68	0,04	2,83	1,01	0,52	0,05	30,62	15,35	935
3DUNET_128_wCE_noDA	0,66	0,04	2,75	0,75	0,49	0,05	38,53	20,17	1152
UNETR_64_wCE_noDA	0,66	0,04	3,22	1,25	0,50	0,04	40,15	18,65	1215
3DUNET_96_wCE_noDA	0,67	0,04	2,80	0,90	0,51	0,05	41,83	16,94	1226
UNETR_96_wCE_DA	0,68	0,04	2,65	0,74	0,51	0,05	42,48	16,92	1270
DynUNET_128_wCE_DA	0,68	0,04	2,85	0,75	0,51	0,04	42,74	14,63	1259
UNETR_128_wCE_DA	0,66	0,04	2,87	0,98	0,50	0,05	46,83	16,54	1374
3DUNET_64_wCE_noDA	0,66	0,04	3,04	1,31	0,49	0,04	47,73	20,88	1434
3DUNET_64_wCE_DA	0,66	0,04	2,55	0,70	0,49	0,04	51,62	19,82	1528
3DUNET_128_wCE_DA	0,66	0,04	2,73	0,95	0,50	0,05	52,14	16,44	1500
UNETR_64_wCE_DA	0,67	0,04	2,80	0,72	0,50	0,04	54,16	17,58	1596
3DUNET_96_wCE_DA	0,66	0,04	2,81	0,83	0,49	0,04	54,81	17,75	1596
DynUNET_96_wCE_DA	0,67	0,04	3,18	1,03	0,50	0,05	55,31	18,41	1663
DynUNET_64_wCE_DA	0,66	0,04	14,57	42,92	0,50	0,05	55,92	15,02	1641

Appendix B

Table B 7: Metrics on validation set for DNN trained over 2 Inputs images (merging of T1-w and FLAIR) with GT FLAIR MSeg, reference cT1-w gold-standard MSeg. The DNN are ordered by increasing Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95%HD mean	95%HD sd	Jaccard mean	Jaccard sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_128_Dice_noDA	0,71	0,05	1,94	0,50	0,56	0,06	15,51	10,61	512
DynUNET_96_DiceCE_DA	0,72	0,05	1,84	0,57	0,57	0,06	17,17	11,80	565
DynUNET_64_DiceCE_DA	0,70	0,06	12,62	35,99	0,54	0,07	17,28	13,58	593
3DUNET_96_Dice_noDA	0,72	0,05	1,82	0,52	0,57	0,06	18,29	9,93	573
3DUNET_128_CE_DA	0,72	0,06	1,85	0,55	0,56	0,07	18,71	13,06	627
DynUNET_96_DiceCE_noDA	0,71	0,06	2,29	0,76	0,55	0,07	18,81	10,37	589
3DUNET_128_DiceCE_noDA	0,71	0,05	1,94	0,47	0,55	0,06	19,04	9,53	580
UNETR_128_Dice_noDA	0,70	0,06	1,99	0,60	0,55	0,07	19,09	10,27	601
3DUNET_96_CE_noDA	0,71	0,06	2,42	1,88	0,56	0,07	19,32	10,52	598
UNETR_128_CE_DA	0,71	0,05	2,07	0,63	0,55	0,06	19,65	13,15	641
DynUNET_128_DiceCE_DA	0,72	0,05	1,80	0,54	0,57	0,06	19,97	13,38	668
UNETR_96_CE_noDA	0,70	0,06	2,11	0,64	0,54	0,07	20,00	10,56	622
DynUNET_128_Dice_noDA	0,72	0,05	1,80	0,57	0,56	0,06	20,12	15,10	687
DynUNET_128_CE_DA	0,72	0,05	1,73	0,58	0,57	0,06	20,12	11,59	644
UNETR_64_DiceCE_DA	0,72	0,05	1,99	0,72	0,56	0,06	20,29	15,02	684
DynUNET_64_DiceCE_noDA	0,70	0,06	42,35	66,49	0,54	0,06	20,98	15,05	698
DynUNET_64_CE_noDA	0,72	0,05	2,13	0,94	0,56	0,07	21,77	14,72	723
3DUNET_64_CE_noDA	0,71	0,05	2,08	0,64	0,56	0,06	21,82	14,06	709
DynUNET_64_Dice_DA	0,67	0,07	79,59	65,79	0,51	0,08	22,13	17,70	778
3DUNET_128_Dice_DA	0,72	0,06	2,50	1,93	0,56	0,07	22,18	15,35	720
3DUNET_64_DiceCE_DA	0,71	0,05	12,69	37,99	0,56	0,06	22,42	13,98	713
3DUNET_128_CE_noDA	0,71	0,05	2,11	0,53	0,55	0,06	22,53	11,63	690
DynUNET_96_Dice_noDA	0,72	0,05	1,82	0,69	0,57	0,07	22,58	11,53	707
DynUNET_128_DiceCE_noDA	0,72	0,05	2,01	0,71	0,56	0,06	22,84	15,91	757
3DUNET_64_Dice_noDA	0,72	0,05	1,85	0,55	0,57	0,06	22,99	11,62	705
3DUNET_64_Dice_DA	0,72	0,05	1,87	0,83	0,56	0,06	23,54	14,35	749
DynUNET_128_Dice_DA	0,72	0,05	1,75	0,56	0,57	0,06	24,32	15,73	798
DynUNET_64_CE_DA	0,72	0,05	1,90	0,56	0,56	0,06	24,48	13,32	769
3DUNET_64_CE_DA	0,70	0,06	2,00	0,62	0,55	0,07	24,75	15,75	796
UNETR_64_CE_noDA	0,71	0,05	2,15	0,80	0,55	0,06	24,82	13,27	776
DynUNET_96_Dice_DA	0,72	0,06	15,56	36,66	0,57	0,07	24,86	14,40	804
UNETR_64_DiceCE_noDA	0,71	0,05	2,03	0,86	0,56	0,06	24,89	13,02	773
DynUNET_128_CE_noDA	0,71	0,05	1,97	0,67	0,55	0,06	25,06	13,38	785
DynUNET_96_CE_DA	0,72	0,06	2,10	1,09	0,56	0,07	25,18	11,01	758
3DUNET_96_Dice_DA	0,72	0,05	10,04	30,75	0,57	0,06	25,25	13,17	768
3DUNET_64_DiceCE_noDA	0,72	0,05	1,88	0,52	0,56	0,06	25,38	12,29	771
3DUNET_96_CE_DA	0,71	0,06	1,92	0,65	0,56	0,07	25,57	14,16	792
UNETR_96_DiceCE_DA	0,71	0,05	10,02	30,70	0,56	0,06	25,81	14,92	809
DynUNET_96_CE_noDA	0,71	0,05	2,02	0,64	0,56	0,06	26,27	14,21	821
UNETR_64_CE_DA	0,72	0,05	1,86	0,58	0,56	0,06	26,43	14,44	830
3DUNET_128_DiceCE_DA	0,71	0,06	1,93	0,58	0,55	0,07	27,64	13,90	833
DynUNET_64_Dice_noDA	0,70	0,05	63,66	62,34	0,54	0,06	28,41	18,00	913

Appendix B

UNETR_128_CE_noDA	0,70	0,05	2,21	0,56	0,54	0,06	28,48	13,06	851
UNETR_128_DiceCE_DA	0,71	0,06	2,25	1,67	0,55	0,07	28,63	14,22	876
3DUNET_96_DiceCE_DA	0,72	0,05	1,83	0,57	0,56	0,06	28,68	12,71	853
UNETR_96_Dice_noDA	0,71	0,05	2,10	0,63	0,55	0,06	29,07	12,20	870
UNETR_96_CE_DA	0,71	0,05	2,00	0,61	0,56	0,06	29,36	13,37	881
3DUNET_96_DiceCE_noDA	0,70	0,05	1,98	0,52	0,54	0,06	29,37	13,92	883
UNETR_96_Dice_DA	0,72	0,05	1,87	0,55	0,56	0,06	29,50	14,68	909
UNETR_64_Dice_noDA	0,72	0,05	2,02	0,64	0,56	0,06	31,41	14,01	954
UNETR_96_DiceCE_noDA	0,70	0,05	2,07	0,62	0,55	0,06	32,43	12,72	965
UNETR_128_Dice_DA	0,71	0,06	2,23	0,79	0,55	0,07	32,89	15,02	993
UNETR_128_DiceCE_noDA	0,69	0,05	2,10	0,58	0,53	0,06	32,94	12,76	968
UNETR_64_Dice_DA	0,68	0,07	119,52	50,75	0,52	0,07	36,69	22,71	1158
DynUNET_96_wCE_noDA	0,70	0,05	2,04	0,66	0,54	0,05	39,09	12,95	1159
UNETR_128_wCE_noDA	0,68	0,05	2,27	0,61	0,52	0,06	45,78	13,42	1340
DynUNET_128_wCE_noDA	0,68	0,05	2,11	0,65	0,52	0,05	46,99	16,49	1389
UNETR_96_wCE_noDA	0,68	0,05	2,31	0,70	0,52	0,06	49,12	15,70	1437
3DUNET_128_wCE_noDA	0,67	0,05	2,25	0,71	0,51	0,05	63,05	16,72	1822
DynUNET_64_wCE_noDA	0,67	0,05	11,88	35,40	0,50	0,05	66,62	16,76	1957
3DUNET_96_wCE_noDA	0,67	0,05	2,26	0,68	0,51	0,05	66,67	17,30	1924
UNETR_96_wCE_DA	0,67	0,05	2,47	1,03	0,51	0,06	68,13	21,48	1994
UNETR_128_wCE_DA	0,66	0,05	2,40	0,69	0,50	0,06	68,68	18,60	2003
UNETR_64_wCE_noDA	0,66	0,05	2,56	1,00	0,49	0,05	71,45	17,78	2086
UNETR_64_wCE_DA	0,67	0,05	2,39	0,79	0,50	0,05	74,70	20,64	2178
DynUNET_128_wCE_DA	0,66	0,06	2,49	1,07	0,49	0,06	74,92	20,38	2187
DynUNET_96_wCE_DA	0,66	0,05	2,42	0,74	0,49	0,06	77,91	17,91	2272
3DUNET_128_wCE_DA	0,66	0,05	2,29	0,70	0,49	0,06	78,99	22,97	2308
3DUNET_64_wCE_noDA	0,65	0,04	2,70	1,17	0,49	0,05	79,75	19,19	2309
3DUNET_96_wCE_DA	0,65	0,05	2,56	1,01	0,48	0,06	85,94	21,36	2487
3DUNET_64_wCE_DA	0,64	0,05	2,62	0,94	0,48	0,05	89,60	21,62	2610
DynUNET_64_wCE_DA	0,64	0,05	14,98	43,75	0,47	0,06	94,54	24,30	2756

B.4 DNN VALIDATION SET RESULTS: SINGLE-DNN TABLES PERFORMANCE INDICES FOR TRAINING WITH GROUND TRUTH WITH CONTRAST

In the following tables, only these performance indices are reported: Dice Coefficient, 95% Hausdorff Distance, Percentage Volume Difference.

B.4.1 Input T1-w, GT cT1-w

Table B 8: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_64_DiceCE_DA	0,73	0,04	1,83	0,62	8,74	5,98	352
3DUNET_96_CE_noDA	0,72	0,05	2,17	1,82	9,85	4,48	345
3DUNET_96_CE_DA	0,74	0,05	2,20	1,97	10,19	4,39	339
3DUNET_96_DiceCE_DA	0,73	0,04	1,72	0,49	10,29	7,60	371
3DUNET_64_DiceCE_noDA	0,72	0,05	2,05	1,03	10,37	6,32	399
3DUNET_128_Dice_noDA	0,70	0,05	2,06	1,07	10,45	6,24	371
3DUNET_96_DiceCE_noDA	0,70	0,05	2,16	1,32	10,60	5,35	368
3DUNET_128_CE_DA	0,72	0,06	2,42	2,71	10,61	4,93	373
3DUNET_96_Dice_noDA	0,72	0,04	1,87	0,63	10,98	8,71	395
3DUNET_128_DiceCE_noDA	0,69	0,05	2,36	1,66	11,04	5,50	394
3DUNET_64_CE_noDA	0,73	0,05	1,80	0,83	11,14	7,58	443
3DUNET_128_DiceCE_DA	0,72	0,05	2,12	1,64	11,45	8,79	413
3DUNET_64_Dice_DA	0,72	0,04	26,90	50,06	11,53	8,26	414
3DUNET_64_Dice_noDA	0,73	0,04	1,75	0,53	11,58	8,63	430
3DUNET_128_CE_noDA	0,70	0,05	2,44	2,17	11,70	7,41	389
3DUNET_96_Dice_DA	0,74	0,04	1,58	0,32	13,81	9,59	474
3DUNET_128_Dice_DA	0,74	0,04	1,58	0,32	14,67	10,26	493
3DUNET_64_CE_DA	0,72	0,05	2,05	0,78	15,35	9,09	572
3DUNET_128_wCE_noDA	0,69	0,05	2,40	1,77	34,37	17,45	1030
3DUNET_64_wCE_DA	0,69	0,04	1,98	0,75	47,34	21,60	1426
3DUNET_96_wCE_noDA	0,69	0,05	2,09	0,67	50,51	17,66	1468
3DUNET_128_wCE_DA	0,69	0,05	2,31	1,49	56,85	19,73	1673
3DUNET_64_wCE_noDA	0,69	0,05	1,86	0,50	57,64	19,42	1670
3DUNET_96_wCE_DA	0,68	0,05	1,98	0,53	67,55	17,67	1961

Table B 9: Metrics of nnU-Net (DynUNET) combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_DiceCE_DA	0,77	0,03	1,47	0,33	7,86	6,60	311
DynUNET_96_CE_DA	0,76	0,04	1,47	0,33	7,94	5,20	312
DynUNET_128_CE_DA	0,76	0,04	1,55	0,59	8,00	5,07	301
DynUNET_64_CE_DA	0,76	0,04	1,59	0,77	8,58	5,26	320
DynUNET_128_DiceCE_DA	0,76	0,04	1,83	1,77	9,04	4,87	323
DynUNET_96_CE_noDA	0,75	0,04	1,64	0,56	9,15	6,64	338
DynUNET_96_DiceCE_noDA	0,75	0,05	1,67	0,70	9,31	6,19	346
DynUNET_128_Dice_DA	0,77	0,04	1,70	1,35	9,33	7,68	368
DynUNET_96_Dice_noDA	0,76	0,04	1,50	0,53	9,34	6,34	364
DynUNET_96_Dice_DA	0,77	0,04	1,59	0,72	9,36	7,34	349
DynUNET_128_CE_noDA	0,76	0,04	1,49	0,33	9,84	6,53	351
DynUNET_64_Dice_noDA	0,76	0,04	9,76	30,07	10,12	8,06	382
DynUNET_128_DiceCE_noDA	0,75	0,04	1,87	1,29	10,20	5,50	374

Appendix B

DynUNET_64_CE_noDA	0,75	0,04	1,97	1,61	10,41	5,66	371
DynUNET_128_Dice_noDA	0,76	0,04	1,70	1,01	10,68	8,16	403
DynUNET_64_DiceCE_DA	0,75	0,04	1,66	0,49	10,73	6,11	386
DynUNET_64_DiceCE_noDA	0,74	0,04	1,69	0,54	10,74	6,04	408
DynUNET_64_Dice_DA	0,75	0,04	25,54	43,70	11,12	10,32	429
DynUNET_128_wCE_noDA	0,74	0,04	1,62	0,49	17,18	11,53	567
DynUNET_96_wCE_noDA	0,75	0,04	1,47	0,34	19,55	11,09	617
DynUNET_64_wCE_noDA	0,73	0,04	1,57	0,37	32,83	12,79	984
DynUNET_128_wCE_DA	0,73	0,04	1,59	0,39	42,73	14,90	1256
DynUNET_96_wCE_DA	0,72	0,04	1,65	0,47	52,74	14,41	1543
DynUNET_64_wCE_DA	0,72	0,04	1,73	0,42	53,74	15,46	1598

Table B 10: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_128_Dice_DA	0,75	0,04	1,66	0,50	7,54	3,79	271
UNETR_96_CE_DA	0,75	0,04	1,54	0,33	8,30	3,83	289
UNETR_128_DiceCE_noDA	0,73	0,05	1,92	0,89	8,72	6,69	308
UNETR_128_CE_DA	0,74	0,04	1,90	1,12	8,98	4,78	331
UNETR_64_CE_DA	0,75	0,04	1,65	0,37	9,22	6,18	359
UNETR_64_Dice_DA	0,74	0,04	1,64	0,32	9,46	5,19	348
UNETR_64_DiceCE_DA	0,74	0,04	1,69	0,51	9,49	5,45	341
UNETR_128_CE_noDA	0,72	0,05	2,01	1,14	9,93	4,21	339
UNETR_96_DiceCE_DA	0,75	0,04	1,69	0,59	10,13	9,93	398
UNETR_64_CE_noDA	0,74	0,05	1,84	0,99	10,38	6,76	389
UNETR_64_DiceCE_noDA	0,74	0,05	1,90	1,27	10,49	6,64	391
UNETR_64_Dice_noDA	0,74	0,05	13,18	42,39	10,80	6,59	387
UNETR_128_Dice_noDA	0,72	0,06	2,33	1,76	11,00	4,75	370
UNETR_96_CE_noDA	0,72	0,05	1,81	0,58	11,09	7,76	425
UNETR_96_DiceCE_noDA	0,73	0,05	2,28	2,00	11,20	5,64	378
UNETR_128_DiceCE_DA	0,75	0,04	1,85	1,15	11,63	8,80	413
UNETR_96_Dice_DA	0,75	0,05	1,98	1,66	16,12	9,97	530
UNETR_96_Dice_noDA	0,73	0,05	19,87	46,47	18,51	11,55	589
UNETR_128_wCE_noDA	0,71	0,05	1,95	0,72	28,51	13,26	861
UNETR_96_wCE_noDA	0,71	0,05	1,94	0,77	41,32	16,14	1226
UNETR_64_wCE_noDA	0,72	0,04	1,76	0,55	42,09	15,27	1259
UNETR_64_wCE_DA	0,71	0,04	1,92	0,51	51,60	14,70	1523
UNETR_128_wCE_DA	0,71	0,04	1,93	0,65	53,11	14,81	1566
UNETR_96_wCE_DA	0,71	0,05	2,00	0,58	54,49	16,67	1633

Appendix B

Table B 11: Metrics of V-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
VNET_96_DiceCE_DA	0,59	0,06	29,32	61,57	14,59	8,84	533
VNET_128_CE_DA	0,62	0,06	3,81	3,59	16,18	12,30	598
VNET_128_DiceCE_DA	0,62	0,05	3,32	1,95	17,00	11,69	566
VNET_64_DiceCE_DA	0,58	0,05	4,45	2,42	17,82	11,21	664
VNET_128_DiceCE_noDA	0,59	0,06	3,75	1,81	18,49	15,98	736
VNET_128_CE_noDA	0,61	0,08	3,59	2,75	18,54	15,68	762
VNET_96_CE_noDA	0,58	0,07	3,49	1,87	21,97	14,14	816
VNET_64_DiceCE_noDA	0,59	0,05	3,97	2,30	23,10	13,72	850
VNET_96_DiceCE_noDA	0,55	0,06	15,62	33,67	27,00	19,58	903
VNET_96_CE_DA	0,55	0,06	6,04	3,30	31,27	13,98	1080
VNET_64_CE_noDA	0,53	0,09	7,63	4,10	36,08	17,31	1257
VNET_64_CE_DA	0,41	0,10	15,60	6,91	55,25	17,37	1768
VNET_96_wCE_noDA	0,60	0,06	2,70	1,08	80,02	35,72	2373
VNET_64_wCE_noDA	0,63	0,05	2,47	0,72	80,67	30,08	2356
VNET_128_wCE_noDA	0,57	0,07	3,13	1,90	85,11	43,61	2585
VNET_96_wCE_DA	0,58	0,06	2,94	1,39	101,20	42,80	2990
VNET_64_wCE_DA	0,55	0,06	3,04	0,53	112,82	48,88	3351
VNET_128_wCE_DA	0,55	0,06	3,31	1,82	114,79	42,91	3353
VNET_128_Dice_noDA	0,22	0,19	92,78	52,35	2177,04	2699,50	93091
VNET_128_Dice_DA	0,20	0,18	110,27	38,39	2472,75	3336,03	111559
VNET_96_Dice_noDA	0,01	0,00	129,18	9,51	15605,31	9801,13	497015
VNET_96_Dice_DA	0,05	0,10	125,45	30,64	20547,22	16236,40	749952
VNET_64_Dice_noDA	0,00	0,00	158,56	5,17	53830,86	24123,14	1685820
VNET_64_Dice_DA	0,00	0,00	161,94	3,99	124101,8	30142,74	3576246

B.4.2 Input FLAIR, GT cT1-w

Table B 12: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_64_DiceCE_DA	0,73	0,05	1,79	0,73	9,40	6,08	330
3DUNET_64_Dice_DA	0,73	0,04	9,21	27,11	10,28	8,67	376
3DUNET_96_Dice_DA	0,74	0,05	1,75	0,62	11,04	9,74	388
3DUNET_96_DiceCE_DA	0,74	0,04	2,04	0,92	11,26	7,76	377
3DUNET_128_Dice_DA	0,73	0,05	1,80	0,73	11,76	9,73	400
3DUNET_64_DiceCE_noDA	0,72	0,07	2,58	2,75	12,37	10,24	454
3DUNET_128_DiceCE_DA	0,72	0,06	2,37	1,92	12,84	9,68	452
3DUNET_128_CE_DA	0,73	0,05	2,22	1,59	12,93	9,07	450
3DUNET_64_Dice_noDA	0,73	0,05	1,85	0,74	12,98	8,55	421
3DUNET_96_DiceCE_noDA	0,72	0,05	1,98	0,91	13,23	8,81	465
3DUNET_128_Dice_noDA	0,72	0,05	1,95	0,67	13,43	8,35	412
3DUNET_96_Dice_noDA	0,73	0,05	1,92	0,79	13,46	9,15	436

Appendix B

3DUNET_96_CE_noDA	0,72	0,07	2,39	2,07	13,68	11,85	522
3DUNET_64_CE_DA	0,73	0,06	2,06	0,91	13,73	9,70	483
3DUNET_96_CE_DA	0,73	0,06	2,49	2,39	14,44	10,52	525
3DUNET_128_DiceCE_noDA	0,70	0,05	1,99	0,65	15,60	9,58	487
3DUNET_64_CE_noDA	0,72	0,07	2,73	2,91	15,91	13,27	612
3DUNET_128_CE_noDA	0,70	0,08	2,46	1,62	16,71	14,00	649
3DUNET_128_wCE_noDA	0,69	0,04	2,22	0,89	40,96	20,51	1229
3DUNET_96_wCE_noDA	0,70	0,05	2,24	1,06	42,77	20,51	1290
3DUNET_64_wCE_noDA	0,70	0,04	2,29	1,62	43,50	21,60	1330
3DUNET_128_wCE_DA	0,70	0,05	1,94	0,65	47,57	18,15	1384
3DUNET_96_wCE_DA	0,70	0,04	1,98	0,83	50,37	18,13	1484
3DUNET_64_wCE_DA	0,70	0,04	1,95	0,46	57,48	15,78	1664

Table B 13: Metrics of nnU-Net (DynUNET) combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_64_CE_DA	0,75	0,05	1,84	0,97	8,66	7,91	344
DynUNET_96_Dice_DA	0,75	0,05	1,93	1,22	9,41	7,66	357
DynUNET_64_Dice_DA	0,74	0,05	1,95	0,90	9,75	7,32	369
DynUNET_128_Dice_noDA	0,75	0,05	1,72	0,69	9,92	7,58	362
DynUNET_64_Dice_noDA	0,75	0,05	9,97	29,97	10,13	7,05	361
DynUNET_96_Dice_noDA	0,75	0,05	1,80	1,12	10,22	8,21	376
DynUNET_96_DiceCE_noDA	0,74	0,05	1,87	0,79	10,49	8,94	397
DynUNET_96_CE_DA	0,75	0,05	1,76	0,81	10,86	8,10	376
DynUNET_96_DiceCE_DA	0,74	0,06	2,14	1,27	10,89	8,09	393
DynUNET_128_CE_DA	0,75	0,06	2,19	2,15	11,48	9,14	406
DynUNET_128_DiceCE_noDA	0,73	0,06	2,02	1,30	11,79	10,11	473
DynUNET_128_DiceCE_DA	0,73	0,06	2,30	2,05	12,04	10,03	470
DynUNET_128_wCE_noDA	0,73	0,05	1,73	0,64	12,10	8,70	393
DynUNET_128_Dice_DA	0,76	0,05	1,81	1,06	12,10	8,91	415
DynUNET_96_CE_noDA	0,74	0,06	1,88	0,97	12,47	11,74	510
DynUNET_64_CE_noDA	0,73	0,06	2,21	1,84	12,62	8,30	465
DynUNET_128_CE_noDA	0,73	0,06	2,01	1,06	12,83	9,60	472
DynUNET_64_DiceCE_DA	0,72	0,06	2,27	0,98	13,55	10,15	515
DynUNET_64_DiceCE_noDA	0,72	0,06	2,51	1,14	16,33	11,87	616
DynUNET_96_wCE_noDA	0,73	0,05	1,75	0,75	18,77	13,22	604
DynUNET_64_wCE_noDA	0,73	0,04	2,01	1,75	29,03	11,70	864
DynUNET_128_wCE_DA	0,72	0,05	1,94	1,00	40,71	13,02	1210
DynUNET_64_wCE_DA	0,72	0,04	2,01	1,01	42,91	16,05	1291
DynUNET_96_wCE_DA	0,71	0,04	1,76	0,54	55,23	13,62	1621

Appendix B

Table B 14: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_128_DiceCE_DA	0,72	0,06	2,00	1,06	9,92	5,45	337
UNETR_128_Dice_DA	0,73	0,05	1,86	0,69	9,95	7,50	360
UNETR_96_CE_DA	0,73	0,06	1,85	0,72	11,22	7,71	391
UNETR_64_Dice_noDA	0,72	0,07	2,44	2,61	11,41	10,18	435
UNETR_96_Dice_noDA	0,72	0,06	2,00	0,78	11,58	6,74	364
UNETR_128_CE_DA	0,72	0,06	2,15	1,18	11,75	7,47	386
UNETR_128_Dice_noDA	0,70	0,07	2,61	2,10	12,32	11,15	462
UNETR_64_Dice_DA	0,71	0,06	2,33	0,74	12,67	9,93	483
UNETR_96_DiceCE_noDA	0,71	0,09	2,68	2,53	13,69	13,70	538
UNETR_96_Dice_DA	0,74	0,06	1,92	0,97	13,76	9,58	476
UNETR_128_DiceCE_noDA	0,69	0,09	2,75	2,58	13,85	12,96	556
UNETR_64_CE_noDA	0,72	0,08	2,44	1,86	14,56	14,79	595
UNETR_64_DiceCE_noDA	0,72	0,08	2,58	2,47	14,78	14,02	576
UNETR_96_CE_noDA	0,70	0,09	3,03	3,14	14,92	15,11	605
UNETR_128_CE_noDA	0,67	0,09	3,28	2,87	15,33	14,44	632
UNETR_96_DiceCE_DA	0,73	0,06	1,84	0,82	15,48	10,44	527
UNETR_64_DiceCE_DA	0,72	0,07	2,18	0,83	15,59	11,17	562
UNETR_64_CE_DA	0,72	0,07	2,23	1,14	16,27	11,90	580
UNETR_128_wCE_noDA	0,69	0,06	2,33	1,14	24,48	13,33	762
UNETR_96_wCE_noDA	0,70	0,06	2,60	2,35	30,40	15,41	940
UNETR_64_wCE_DA	0,71	0,05	1,95	0,74	30,91	17,00	975
UNETR_128_wCE_DA	0,70	0,05	2,34	2,08	33,45	17,53	1067
UNETR_64_wCE_noDA	0,71	0,04	2,14	1,38	37,89	20,00	1193
UNETR_96_wCE_DA	0,70	0,05	2,12	0,98	38,38	19,91	1212

Table B 15: Metrics of V-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
VNET_96_CE_DA	0,69	0,10	4,82	6,55	20,64	19,87	820
VNET_128_CE_noDA	0,66	0,13	3,77	3,81	20,77	22,01	881
VNET_128_DiceCE_DA	0,66	0,09	4,04	3,26	20,88	17,50	800
VNET_64_DiceCE_DA	0,67	0,12	4,38	6,62	21,40	19,70	826
VNET_96_CE_noDA	0,62	0,11	14,17	33,99	22,15	19,32	870
VNET_64_DiceCE_noDA	0,63	0,15	14,74	37,37	25,46	23,10	1019
VNET_128_CE_DA	0,62	0,17	19,62	34,61	25,60	26,02	1068
VNET_128_DiceCE_noDA	0,60	0,18	6,04	7,60	27,37	26,08	1078
VNET_96_DiceCE_DA	0,57	0,16	126,66	86,76	27,74	22,01	1011
VNET_64_CE_noDA	0,56	0,21	9,65	10,94	39,24	28,00	1393
VNET_96_DiceCE_noDA	0,51	0,19	57,16	62,59	39,55	23,90	1384
VNET_64_CE_DA	0,55	0,17	37,33	63,04	41,63	22,78	1411
VNET_96_wCE_noDA	0,64	0,06	2,91	2,16	55,95	23,73	1711
VNET_128_wCE_noDA	0,64	0,06	2,50	0,89	56,56	23,44	1658

Appendix B

VNET_64_wCE_noDA	0,66	0,04	2,35	1,14	61,48	27,76	1869
VNET_96_wCE_DA	0,65	0,05	2,24	0,49	69,50	24,24	2066
VNET_64_wCE_DA	0,61	0,06	2,56	0,61	82,53	43,22	2604
VNET_128_wCE_DA	0,59	0,05	2,88	0,54	107,08	37,25	3206
VNET_64_Dice_DA	0,41	0,28	50,30	48,36	1759,00	3180,07	106126
VNET_96_Dice_noDA	0,08	0,16	99,41	30,66	24051,64	22581,34	939764
VNET_96_Dice_DA	0,08	0,17	119,58	21,14	36454,17	35646,11	1433891
VNET_128_Dice_noDA	0,03	0,08	130,25	13,67	37493,84	37093,18	1471086
VNET_128_Dice_DA	0,00	0,00	135,84	6,93	78709,71	18423,26	2305780
VNET_64_Dice_noDA	0,00	0,00	159,38	3,92	110436	29265,30	3196278

B.4.3 Input T1-w+FLAIR, GT cT1-w

Table B 16: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_128_DiceCE_DA	0,74	0,05	1,61	0,39	9,67	7,46	330
3DUNET_64_DiceCE_DA	0,74	0,05	1,74	0,50	9,67	4,52	332
3DUNET_64_Dice_DA	0,75	0,05	1,62	0,35	10,04	7,93	342
3DUNET_64_Dice_noDA	0,75	0,04	1,65	0,48	10,44	6,46	339
3DUNET_96_DiceCE_DA	0,74	0,05	1,64	0,55	10,68	6,08	346
3DUNET_128_Dice_DA	0,75	0,05	1,60	0,48	10,79	8,03	361
3DUNET_96_CE_DA	0,74	0,05	1,71	0,56	11,04	4,57	343
3DUNET_96_Dice_DA	0,76	0,05	1,50	0,46	11,16	8,81	372
3DUNET_128_CE_DA	0,74	0,05	1,96	1,08	11,34	7,34	408
3DUNET_128_Dice_noDA	0,73	0,05	1,76	0,49	11,55	5,80	386
3DUNET_128_DiceCE_noDA	0,72	0,06	1,76	0,49	11,76	6,66	407
3DUNET_64_CE_DA	0,74	0,06	1,79	0,64	11,98	5,54	398
3DUNET_96_DiceCE_noDA	0,73	0,05	1,85	0,67	12,27	7,20	411
3DUNET_128_CE_noDA	0,71	0,06	2,10	1,18	12,40	8,49	460
3DUNET_96_Dice_noDA	0,74	0,05	2,24	2,00	12,71	7,70	439
3DUNET_64_CE_noDA	0,74	0,06	1,96	1,36	12,83	7,44	438
3DUNET_96_CE_noDA	0,73	0,06	2,17	1,53	13,11	8,22	469
3DUNET_64_DiceCE_noDA	0,74	0,05	1,74	0,75	13,91	10,03	450
3DUNET_96_wCE_noDA	0,72	0,05	1,94	0,82	33,13	19,28	1028
3DUNET_128_wCE_noDA	0,71	0,04	1,84	0,60	33,21	16,17	992
3DUNET_64_wCE_noDA	0,72	0,04	1,74	0,50	40,72	19,42	1232
3DUNET_64_wCE_DA	0,72	0,04	1,65	0,46	48,89	18,40	1438
3DUNET_128_wCE_DA	0,71	0,05	1,71	0,55	49,29	15,87	1433
3DUNET_96_wCE_DA	0,71	0,05	1,76	0,56	52,53	16,51	1528

Appendix B

Table B 17: Metrics of nnU-Net (DynUNET) combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_128_DiceCE_DA	0,76	0,04	1,50	0,38	8,23	5,81	322
DynUNET_128_Dice_noDA	0,76	0,04	1,58	0,56	8,34	5,92	313
DynUNET_96_DiceCE_noDA	0,75	0,05	1,66	0,55	8,47	6,23	313
DynUNET_96_Dice_DA	0,77	0,04	1,56	0,72	8,62	6,33	329
DynUNET_64_CE_noDA	0,76	0,04	1,79	1,33	8,75	5,22	309
DynUNET_128_Dice_DA	0,76	0,04	1,46	0,38	9,08	6,75	336
DynUNET_64_CE_DA	0,76	0,05	1,52	0,65	9,37	5,87	325
DynUNET_96_CE_DA	0,76	0,05	1,68	1,33	9,43	6,07	351
DynUNET_128_CE_DA	0,77	0,04	1,48	0,43	9,91	6,61	333
DynUNET_128_CE_noDA	0,76	0,05	1,62	0,72	9,93	5,47	342
DynUNET_96_DiceCE_DA	0,75	0,05	1,64	0,56	9,99	7,84	372
DynUNET_96_Dice_noDA	0,76	0,05	2,06	1,84	10,00	7,53	366
DynUNET_96_CE_noDA	0,76	0,04	1,59	0,60	10,69	6,36	391
DynUNET_128_DiceCE_noDA	0,74	0,05	2,24	1,36	10,73	7,83	437
DynUNET_64_DiceCE_noDA	0,75	0,06	2,05	1,55	11,73	8,22	435
DynUNET_128_wCE_noDA	0,73	0,04	1,59	0,39	12,79	9,20	414
DynUNET_64_Dice_noDA	0,70	0,06	69,78	61,08	14,42	10,36	524
DynUNET_64_Dice_DA	0,72	0,05	92,02	49,76	14,74	14,67	602
DynUNET_96_wCE_noDA	0,75	0,05	1,77	1,30	15,30	10,70	503
DynUNET_64_DiceCE_DA	0,71	0,05	3,07	1,46	15,80	8,56	582
DynUNET_64_wCE_noDA	0,74	0,04	1,71	0,74	29,02	13,85	889
DynUNET_128_wCE_DA	0,73	0,05	1,60	0,48	45,33	15,67	1335
DynUNET_96_wCE_DA	0,72	0,04	1,66	0,54	54,14	13,50	1561
DynUNET_64_wCE_DA	0,71	0,04	13,23	42,93	60,91	15,47	1776

Table B 18: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_128_DiceCE_DA	0,75	0,05	1,73	0,74	8,34	5,33	288
UNETR_128_Dice_noDA	0,74	0,05	1,69	0,39	9,25	5,06	307
UNETR_96_DiceCE_noDA	0,74	0,05	2,00	1,65	9,39	6,22	324
UNETR_96_Dice_noDA	0,74	0,06	1,92	1,36	9,44	6,77	355
UNETR_128_DiceCE_noDA	0,73	0,06	1,76	0,58	9,55	5,73	308
UNETR_64_DiceCE_noDA	0,75	0,05	1,86	1,16	9,77	6,25	344
UNETR_96_DiceCE_DA	0,76	0,04	1,57	0,39	10,07	6,95	346
UNETR_96_CE_DA	0,75	0,05	2,05	1,51	10,46	7,08	391
UNETR_64_DiceCE_DA	0,75	0,04	1,70	0,48	10,68	8,56	381
UNETR_128_CE_noDA	0,73	0,06	1,81	0,56	11,27	7,53	369
UNETR_128_CE_DA	0,75	0,05	2,07	1,78	11,36	7,90	419
UNETR_64_Dice_noDA	0,76	0,04	1,59	0,53	11,53	7,97	382
UNETR_96_Dice_DA	0,75	0,05	11,31	35,52	11,53	7,73	395
UNETR_64_CE_DA	0,75	0,05	1,64	0,66	11,96	5,79	404

Appendix B

UNETR_128_Dice_DA	0,76	0,04	1,76	0,80	11,98	7,82	391
UNETR_64_CE_noDA	0,74	0,05	2,03	1,19	12,05	7,69	452
UNETR_96_CE_noDA	0,73	0,06	1,99	1,27	12,19	7,74	388
UNETR_64_Dice_DA	0,75	0,05	34,53	56,71	16,90	11,75	554
UNETR_96_wCE_noDA	0,73	0,05	1,65	0,52	22,53	12,85	694
UNETR_128_wCE_noDA	0,72	0,05	1,99	1,13	23,14	12,59	718
UNETR_64_wCE_noDA	0,73	0,04	1,69	0,63	38,00	14,42	1126
UNETR_128_wCE_DA	0,72	0,04	1,94	1,28	43,35	15,92	1290
UNETR_96_wCE_DA	0,72	0,04	1,68	0,57	50,19	15,50	1470
UNETR_64_wCE_DA	0,71	0,04	1,81	0,56	54,11	17,43	1585

Table B 19: Metrics of V-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
VNET_96_DiceCE_DA	0,59	0,09	20,92	32,98	21,36	16,54	778
VNET_128_DiceCE_noDA	0,60	0,10	34,86	54,10	24,15	17,03	920
VNET_128_CE_DA	0,63	0,09	5,08	3,16	25,54	12,90	894
VNET_128_wCE_DA	0,62	0,08	6,53	6,36	25,58	19,68	940
VNET_128_CE_noDA	0,61	0,12	7,69	6,30	27,64	19,36	1018
VNET_64_DiceCE_noDA	0,58	0,11	43,47	61,22	27,83	15,99	1000
VNET_64_CE_noDA	0,62	0,12	15,10	35,91	28,53	17,25	1039
VNET_128_DiceCE_DA	0,57	0,10	22,67	31,32	29,85	16,29	1042
VNET_64_DiceCE_DA	0,56	0,10	38,89	52,67	33,05	23,41	1138
VNET_96_DiceCE_noDA	0,50	0,10	59,71	56,91	38,07	19,83	1329
VNET_96_wCE_DA	0,63	0,08	24,36	47,00	41,38	22,17	1270
VNET_64_CE_DA	0,48	0,13	59,92	60,60	44,91	16,70	1452
VNET_96_CE_noDA	0,56	0,12	10,24	7,23	45,19	12,28	1398
VNET_96_wCE_noDA	0,66	0,06	2,80	1,45	48,22	22,89	1449
VNET_64_wCE_DA	0,65	0,05	2,34	0,51	51,48	20,64	1548
VNET_96_CE_DA	0,46	0,14	36,17	43,97	51,94	16,28	1645
VNET_64_wCE_noDA	0,65	0,06	12,27	36,44	56,27	23,07	1656
VNET_128_wCE_noDA	0,62	0,06	2,79	0,84	76,11	35,95	2294
VNET_128_Dice_noDA	0,23	0,25	100,75	46,50	9658,64	13272,96	469307
VNET_128_Dice_DA	0,20	0,26	100,89	43,97	10130,22	13294,22	469056
VNET_96_Dice_DA	0,01	0,02	130,42	14,53	24499,26	19079,00	825575
VNET_96_Dice_noDA	0,00	0,00	141,37	3,77	59173,24	19769,19	1721963
VNET_64_Dice_noDA	0,00	0,00	153,70	3,40	162433,1	40130,07	4659140
VNET_64_Dice_DA	0,00	0,00	150,56	5,93	179706,3	40151,78	5124270

B.5 DNN VALIDATION SET RESULTS: SINGLE-DNN TABLES PERFORMANCE INDICES FOR TRAINING WITH GROUND TRUTH WITHOUT CONTRAST, PERFORMANCE INDICES CALCULATED WITH RESPECT TO THE CONTRAST GROUND TRUTH

In the following tables, only these performance indices are reported: Dice Coefficient, 95% Hausdorff Distance, Percentage Volume Difference.

B.5.1 Input T1-w, GT T1-w – reference cT1-w

Table B 20: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_64_DiceCE_DA	0,67	0,04	3,41	2,24	8,22	6,03	331
3DUNET_96_CE_noDA	0,68	0,05	3,82	2,34	9,19	6,57	340
3DUNET_96_DiceCE_noDA	0,67	0,05	3,45	1,95	10,39	6,36	359
3DUNET_128_CE_noDA	0,65	0,06	3,73	1,88	11,33	7,79	398
3DUNET_128_Dice_DA	0,68	0,04	2,86	0,90	11,49	7,08	388
3DUNET_64_DiceCE_noDA	0,67	0,04	3,09	1,33	11,67	7,71	432
3DUNET_128_CE_DA	0,67	0,05	3,48	1,34	11,79	8,11	393
3DUNET_64_CE_noDA	0,68	0,05	3,70	2,38	12,08	8,14	452
3DUNET_96_Dice_noDA	0,67	0,05	3,77	2,15	12,19	7,87	485
3DUNET_64_Dice_noDA	0,67	0,04	3,39	1,80	12,33	7,91	422
3DUNET_64_Dice_DA	0,68	0,04	11,23	31,19	12,36	10,96	449
3DUNET_96_CE_DA	0,68	0,05	4,03	2,31	12,43	7,13	444
3DUNET_128_DiceCE_noDA	0,64	0,05	4,03	2,20	12,75	9,32	433
3DUNET_96_Dice_DA	0,68	0,04	13,09	35,29	13,13	8,68	436
3DUNET_96_DiceCE_DA	0,68	0,04	3,92	2,14	13,30	8,35	435
3DUNET_64_CE_DA	0,67	0,05	4,29	2,27	14,53	8,34	511
3DUNET_128_DiceCE_DA	0,66	0,04	3,40	1,34	15,03	8,96	475
3DUNET_128_Dice_noDA	0,62	0,05	4,53	1,82	23,85	12,41	870
3DUNET_128_wCE_noDA	0,63	0,05	3,00	0,73	45,62	21,92	1351
3DUNET_96_wCE_DA	0,66	0,04	2,91	0,70	49,32	15,65	1434
3DUNET_96_wCE_noDA	0,64	0,05	2,77	0,62	50,15	18,66	1455
3DUNET_64_wCE_noDA	0,64	0,04	3,11	0,92	57,95	20,34	1687
3DUNET_128_wCE_DA	0,64	0,05	3,15	0,85	61,29	17,71	1769
3DUNET_64_wCE_DA	0,63	0,05	3,41	1,19	70,00	24,16	2032

Table B 21: Metrics of nnU-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_CE_DA	0,71	0,04	3,48	1,87	7,40	5,91	309
DynUNET_128_CE_DA	0,70	0,04	3,09	1,12	7,93	5,43	322
DynUNET_96_DiceCE_noDA	0,70	0,04	3,73	1,59	8,29	6,24	341

Appendix B

DynUNET_64_CE_noDA	0,69	0,04	3,33	1,37	8,44	5,63	329
DynUNET_96_Dice_noDA	0,70	0,04	2,86	0,97	8,68	6,60	347
DynUNET_128_CE_noDA	0,69	0,04	2,90	0,87	8,94	6,47	324
DynUNET_128_DiceCE_DA	0,71	0,04	3,05	0,98	8,99	7,45	365
DynUNET_64_CE_DA	0,70	0,04	3,05	1,17	9,08	7,22	366
DynUNET_96_CE_noDA	0,69	0,05	2,79	0,89	9,28	5,37	326
DynUNET_64_Dice_noDA	0,70	0,04	2,97	1,05	9,85	8,60	393
DynUNET_128_DiceCE_noDA	0,70	0,04	2,80	0,83	10,21	6,89	378
DynUNET_96_DiceCE_DA	0,70	0,04	4,32	1,90	11,01	7,17	431
DynUNET_64_DiceCE_noDA	0,69	0,04	3,12	0,68	11,95	7,72	446
DynUNET_128_Dice_noDA	0,70	0,05	3,18	1,16	12,52	10,73	476
DynUNET_128_Dice_DA	0,71	0,04	2,99	0,97	12,92	10,40	478
DynUNET_64_DiceCE_DA	0,68	0,04	3,64	0,98	13,41	7,93	511
DynUNET_128_wCE_noDA	0,68	0,05	3,00	1,19	15,35	11,43	505
DynUNET_96_Dice_DA	0,70	0,04	21,53	47,65	17,46	11,63	596
DynUNET_96_wCE_noDA	0,68	0,04	3,21	1,36	21,88	11,90	670
DynUNET_64_Dice_DA	0,66	0,06	63,57	62,99	32,81	23,96	1088
DynUNET_64_wCE_noDA	0,67	0,04	2,66	1,03	33,56	14,76	1009
DynUNET_128_wCE_DA	0,67	0,04	3,05	1,07	45,14	16,43	1359
DynUNET_96_wCE_DA	0,66	0,04	3,19	0,96	63,01	15,66	1838
DynUNET_64_wCE_DA	0,65	0,04	3,68	1,12	65,11	20,42	1933

Table B 22: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_96_CE_DA	0,70	0,04	2,80	1,28	8,47	4,11	281
UNETR_96_DiceCE_noDA	0,69	0,05	4,03	2,86	8,47	5,85	292
UNETR_96_DiceCE_DA	0,70	0,04	3,57	2,47	8,61	6,67	324
UNETR_64_CE_DA	0,70	0,05	2,70	0,41	8,61	6,85	331
UNETR_128_DiceCE_noDA	0,67	0,05	2,90	0,76	8,90	6,15	322
UNETR_96_Dice_DA	0,69	0,04	20,09	42,87	9,29	5,83	357
UNETR_128_CE_DA	0,69	0,05	3,13	1,14	9,52	4,57	320
UNETR_64_Dice_DA	0,70	0,04	3,05	1,04	10,02	8,33	384
UNETR_96_CE_noDA	0,68	0,05	3,16	1,53	10,25	5,17	339
UNETR_64_DiceCE_DA	0,69	0,05	3,15	1,08	11,77	8,00	447
UNETR_128_CE_noDA	0,67	0,05	3,06	1,01	12,03	8,79	408
UNETR_64_CE_noDA	0,68	0,05	3,20	1,46	12,37	7,72	442
UNETR_64_DiceCE_noDA	0,68	0,05	3,90	2,62	12,39	7,29	457
UNETR_128_DiceCE_DA	0,70	0,04	2,52	0,67	13,19	8,34	422
UNETR_96_Dice_noDA	0,70	0,04	3,58	2,65	14,20	8,88	444
UNETR_128_Dice_noDA	0,68	0,05	2,90	0,77	14,65	10,83	479
UNETR_128_Dice_DA	0,69	0,04	2,51	0,79	19,00	11,71	605
UNETR_64_Dice_noDA	0,67	0,04	100,48	68,38	22,77	15,32	722
UNETR_64_wCE_DA	0,68	0,04	2,63	1,04	27,90	16,77	901
UNETR_128_wCE_noDA	0,66	0,05	3,40	1,76	31,25	13,51	917

Appendix B

UNETR_96_wCE_noDA	0,66	0,05	3,32	1,36	36,47	16,11	1073
UNETR_96_wCE_DA	0,67	0,04	2,80	0,58	43,31	14,15	1281
UNETR_64_wCE_noDA	0,67	0,04	2,61	0,65	43,77	18,20	1293
UNETR_128_wCE_DA	0,66	0,04	3,07	0,91	51,27	17,26	1501

B.5.2 Input FLAIR, GT FLAIR – reference cT1-w

Table B 23: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_64_DiceCE_DA	0,71	0,05	2,14	0,98	16,75	10,98	550
3DUNET_64_Dice_DA	0,70	0,05	21,13	37,91	17,10	14,25	621
3DUNET_64_CE_noDA	0,71	0,05	2,03	0,70	19,02	11,63	602
3DUNET_64_DiceCE_noDA	0,70	0,05	2,74	3,01	19,20	11,43	615
3DUNET_96_DiceCE_DA	0,72	0,05	2,19	0,94	19,79	11,84	626
3DUNET_64_CE_DA	0,71	0,05	1,91	0,60	20,89	10,77	646
3DUNET_96_DiceCE_noDA	0,71	0,05	2,03	0,79	20,95	9,80	632
3DUNET_128_CE_DA	0,71	0,06	2,45	1,84	22,09	12,94	693
3DUNET_96_CE_DA	0,71	0,05	2,10	0,67	22,92	11,79	706
3DUNET_128_DiceCE_DA	0,71	0,05	1,97	0,78	23,99	10,87	726
3DUNET_128_DiceCE_noDA	0,69	0,06	2,27	1,04	26,06	13,80	803
3DUNET_128_Dice_noDA	0,70	0,05	2,13	0,72	27,04	15,19	854
3DUNET_128_CE_noDA	0,69	0,05	2,20	0,75	27,26	14,79	847
3DUNET_96_CE_noDA	0,70	0,06	2,21	0,97	29,21	14,96	901
3DUNET_64_Dice_noDA	0,71	0,05	1,97	0,59	31,85	12,34	943
3DUNET_128_Dice_DA	0,70	0,06	2,17	0,86	35,48	13,40	1050
3DUNET_96_Dice_DA	0,70	0,06	20,72	47,59	38,15	15,69	1146
3DUNET_96_Dice_noDA	0,70	0,05	2,11	0,73	39,19	15,59	1170
3DUNET_128_wCE_noDA	0,65	0,05	2,46	0,78	66,60	20,31	1945
3DUNET_96_wCE_noDA	0,66	0,05	2,48	0,85	69,51	22,43	2052
3DUNET_96_wCE_DA	0,65	0,05	2,57	0,85	74,64	16,92	2148
3DUNET_128_wCE_DA	0,65	0,05	2,57	1,04	76,15	19,22	2199
3DUNET_64_wCE_noDA	0,64	0,04	2,47	0,70	80,44	20,98	2344
3DUNET_64_wCE_DA	0,64	0,05	2,77	0,89	83,70	18,12	2407

Table B 24: Metrics of nnU-Net (DynUNET) combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_DiceCE_DA	0,70	0,06	3,54	2,63	11,78	9,73	420
DynUNET_64_DiceCE_DA	0,72	0,05	2,10	1,19	17,50	11,64	589
DynUNET_64_CE_DA	0,72	0,05	2,16	1,04	18,16	11,76	591
DynUNET_128_CE_DA	0,71	0,05	25,40	46,24	18,88	11,88	622
DynUNET_96_CE_noDA	0,71	0,05	1,89	0,52	19,18	9,84	600
DynUNET_64_DiceCE_noDA	0,71	0,05	2,31	1,31	19,96	12,82	643

Appendix B

DynUNET_128_DiceCE_DA	0,71	0,07	2,36	1,62	20,67	11,44	651
DynUNET_96_DiceCE_noDA	0,71	0,05	2,01	0,67	21,14	13,68	698
DynUNET_96_Dice_noDA	0,72	0,05	1,96	0,49	22,20	10,81	694
DynUNET_64_CE_noDA	0,72	0,05	1,91	0,97	22,63	12,47	731
DynUNET_96_Dice_DA	0,72	0,05	2,10	1,05	23,22	12,02	742
DynUNET_128_Dice_noDA	0,71	0,05	1,89	0,59	24,76	11,18	750
DynUNET_64_Dice_noDA	0,72	0,05	1,96	0,52	25,44	11,57	792
DynUNET_128_DiceCE_noDA	0,71	0,04	1,91	0,47	26,18	14,06	816
DynUNET_128_CE_noDA	0,71	0,05	2,03	0,75	27,40	13,46	836
DynUNET_64_Dice_DA	0,71	0,05	15,76	34,50	27,65	14,18	880
DynUNET_128_Dice_DA	0,72	0,05	2,27	0,84	28,60	12,48	867
DynUNET_96_CE_DA	0,71	0,05	2,42	1,10	30,05	12,13	896
DynUNET_128_wCE_noDA	0,68	0,05	2,13	0,73	42,19	13,05	1235
DynUNET_96_wCE_noDA	0,69	0,04	1,97	0,50	45,38	12,34	1330
DynUNET_64_wCE_noDA	0,68	0,04	2,17	0,79	54,17	14,58	1598
DynUNET_128_wCE_DA	0,66	0,05	2,25	0,62	73,74	17,97	2139
DynUNET_96_wCE_DA	0,66	0,05	2,65	1,37	78,59	18,63	2285
DynUNET_64_wCE_DA	0,65	0,04	2,83	1,13	85,52	16,61	2486

Table B 25: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_64_CE_noDA	0,69	0,06	2,91	2,10	14,65	10,78	260543
UNETR_96_CE_DA	0,70	0,06	3,17	2,56	14,88	12,17	276794
UNETR_64_DiceCE_DA	0,69	0,06	2,33	0,83	15,44	11,74	284970
UNETR_64_Dice_DA	0,69	0,06	17,32	29,77	15,80	11,03	286544
UNETR_128_CE_noDA	0,68	0,06	2,50	0,98	18,00	7,98	301903
UNETR_96_Dice_noDA	0,70	0,06	2,30	0,96	19,32	11,29	379780
UNETR_64_CE_DA	0,70	0,05	2,50	1,27	20,03	14,03	421250
UNETR_96_DiceCE_DA	0,71	0,05	2,13	0,91	20,46	12,20	416983
UNETR_96_CE_noDA	0,69	0,07	2,76	1,85	21,48	10,54	426755
UNETR_96_Dice_DA	0,71	0,06	2,27	1,14	22,86	13,44	559705
UNETR_128_DiceCE_DA	0,71	0,05	2,10	0,80	24,40	11,60	561761
UNETR_128_DiceCE_noDA	0,67	0,07	2,84	1,57	24,97	12,11	574089
UNETR_64_Dice_noDA	0,70	0,05	2,44	0,95	25,07	14,06	622050
UNETR_128_CE_DA	0,70	0,05	2,32	0,80	25,60	11,32	612923
UNETR_64_DiceCE_noDA	0,69	0,06	2,65	1,81	27,53	14,22	697932
UNETR_128_Dice_noDA	0,68	0,06	2,66	1,05	31,48	14,85	936467
UNETR_96_DiceCE_noDA	0,68	0,06	2,71	1,56	33,40	16,19	1094083
UNETR_128_Dice_DA	0,70	0,05	2,27	0,68	38,57	12,63	1330220
UNETR_128_wCE_noDA	0,65	0,06	2,55	0,63	52,65	19,24	2484031
UNETR_96_wCE_noDA	0,65	0,05	2,63	0,86	56,33	21,29	2940240
UNETR_64_wCE_DA	0,65	0,04	2,55	0,93	65,11	24,24	3877985
UNETR_64_wCE_noDA	0,65	0,05	3,06	1,72	66,62	24,26	4027893
UNETR_128_wCE_DA	0,65	0,05	2,73	0,94	71,47	18,84	4440152
UNETR_96_wCE_DA	0,65	0,05	2,84	1,01	72,89	17,63	4580237

B.5.3 Input T1-w+FLAIR, GT T1-w – reference cT1-w

Table B 26: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_96_DiceCE_DA	0,69	0,04	3,97	2,42	8,96	8,18	346
3DUNET_96_Dice_DA	0,69	0,05	3,07	0,93	9,21	6,72	346
3DUNET_128_Dice_DA	0,69	0,04	3,45	1,69	9,32	7,48	353
3DUNET_96_CE_DA	0,69	0,05	3,67	1,36	10,69	8,04	420
3DUNET_128_CE_DA	0,70	0,04	2,83	1,03	10,82	9,52	392
3DUNET_64_CE_DA	0,69	0,05	2,92	0,90	10,86	9,77	414
3DUNET_96_Dice_noDA	0,69	0,05	3,36	1,45	11,01	8,25	423
3DUNET_64_DiceCE_noDA	0,68	0,05	3,37	1,91	11,38	7,29	383
3DUNET_64_Dice_noDA	0,69	0,04	3,47	2,11	11,44	8,29	399
3DUNET_64_CE_noDA	0,68	0,05	3,57	2,09	11,49	8,01	413
3DUNET_128_Dice_noDA	0,68	0,05	3,47	1,91	11,61	8,14	408
3DUNET_128_DiceCE_DA	0,69	0,04	3,47	2,10	11,67	8,00	374
3DUNET_96_DiceCE_noDA	0,68	0,05	3,46	2,00	11,83	8,39	455
3DUNET_64_DiceCE_DA	0,68	0,04	31,09	54,28	11,94	8,46	396
3DUNET_128_DiceCE_noDA	0,67	0,05	4,08	2,56	11,97	9,66	450
3DUNET_128_CE_noDA	0,67	0,05	3,36	1,35	12,11	10,24	463
3DUNET_64_Dice_DA	0,70	0,04	2,82	0,94	12,31	9,21	419
3DUNET_96_CE_noDA	0,68	0,05	3,87	2,14	12,71	9,72	483
3DUNET_128_wCE_noDA	0,66	0,04	2,75	0,75	38,53	20,17	1152
3DUNET_96_wCE_noDA	0,67	0,04	2,80	0,90	41,83	16,94	1226
3DUNET_64_wCE_noDA	0,66	0,04	3,04	1,31	47,73	20,88	1434
3DUNET_64_wCE_DA	0,66	0,04	2,55	0,70	51,62	19,82	1528
3DUNET_128_wCE_DA	0,66	0,04	2,73	0,95	52,14	16,44	1500
3DUNET_96_wCE_DA	0,66	0,04	2,81	0,83	54,81	17,75	1596

Table B 27: Metrics of nnU-Net (DynUNET) combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_Dice_noDA	0,70	0,05	2,91	1,40	8,34	7,78	356
DynUNET_128_CE_noDA	0,69	0,04	2,80	0,97	8,47	7,47	353
DynUNET_96_CE_DA	0,70	0,05	2,49	0,65	8,55	6,36	315
DynUNET_128_CE_DA	0,71	0,05	2,50	0,55	8,98	6,74	353
DynUNET_128_DiceCE_DA	0,70	0,04	3,02	1,73	9,34	6,97	361
DynUNET_64_CE_DA	0,71	0,04	13,65	41,58	9,38	6,85	372
DynUNET_64_DiceCE_noDA	0,70	0,04	2,49	0,77	9,95	9,23	387
DynUNET_128_Dice_noDA	0,70	0,04	3,22	1,31	10,30	6,46	383
DynUNET_96_DiceCE_noDA	0,69	0,06	3,30	1,78	10,64	8,36	412
DynUNET_96_Dice_DA	0,70	0,05	3,46	2,41	10,95	8,73	401
DynUNET_96_DiceCE_DA	0,70	0,05	3,54	1,46	11,09	9,85	449
DynUNET_64_CE_noDA	0,69	0,04	2,60	0,73	11,73	8,88	403

Appendix B

DynUNET_96_CE_noDA	0,68	0,05	3,22	1,90	12,09	9,09	465
DynUNET_128_Dice_DA	0,70	0,04	2,83	0,72	12,32	9,84	456
DynUNET_128_DiceCE_noDA	0,70	0,05	3,47	1,73	12,53	6,83	430
DynUNET_64_Dice_DA	0,68	0,05	48,92	54,35	14,19	12,77	531
DynUNET_64_DiceCE_DA	0,68	0,05	8,35	2,11	15,19	9,55	577
DynUNET_64_Dice_noDA	0,67	0,05	76,72	68,54	15,83	11,31	522
DynUNET_96_wCE_noDA	0,69	0,04	2,37	0,85	20,12	13,30	658
DynUNET_128_wCE_noDA	0,68	0,04	3,22	1,82	22,50	14,34	722
DynUNET_64_wCE_noDA	0,68	0,04	2,83	1,01	30,62	15,35	935
DynUNET_128_wCE_DA	0,68	0,04	2,85	0,75	42,74	14,63	1259
DynUNET_96_wCE_DA	0,67	0,04	3,18	1,03	55,31	18,41	1663
DynUNET_64_wCE_DA	0,66	0,04	14,57	42,92	55,92	15,02	1641

Table B 28: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_96_CE_DA	0,69	0,05	2,48	0,37	8,68	7,74	367
UNETR_64_DiceCE_DA	0,69	0,05	2,64	0,62	8,75	7,03	342
UNETR_64_DiceCE_noDA	0,69	0,05	3,85	2,61	8,94	7,29	340
UNETR_96_DiceCE_DA	0,70	0,04	2,17	0,50	9,67	6,56	338
UNETR_128_CE_DA	0,69	0,06	3,55	2,09	9,84	10,13	415
UNETR_64_CE_noDA	0,68	0,05	3,36	1,97	9,89	9,38	410
UNETR_64_CE_DA	0,69	0,05	2,78	0,95	9,93	9,43	423
UNETR_96_Dice_DA	0,70	0,05	2,66	0,76	9,98	7,18	370
UNETR_128_Dice_noDA	0,68	0,05	3,43	1,93	10,22	7,96	355
UNETR_96_Dice_noDA	0,69	0,05	3,53	1,99	10,53	8,38	391
UNETR_128_CE_noDA	0,67	0,05	4,24	2,26	10,92	9,15	434
UNETR_128_Dice_DA	0,69	0,04	3,15	1,07	11,14	7,15	368
UNETR_96_DiceCE_noDA	0,68	0,06	3,92	2,22	11,32	7,59	405
UNETR_96_CE_noDA	0,67	0,06	3,80	2,50	11,85	9,69	449
UNETR_64_Dice_noDA	0,70	0,04	2,86	1,44	12,47	8,90	422
UNETR_128_DiceCE_DA	0,69	0,04	3,57	1,79	12,93	8,12	420
UNETR_64_Dice_DA	0,69	0,06	35,53	58,29	13,05	12,30	495
UNETR_128_DiceCE_noDA	0,67	0,06	3,28	1,16	13,51	9,93	462
UNETR_128_wCE_noDA	0,67	0,04	3,03	1,26	24,92	13,61	758
UNETR_96_wCE_noDA	0,67	0,05	2,80	0,80	29,35	13,70	878
UNETR_64_wCE_noDA	0,66	0,04	3,22	1,25	40,15	18,65	1215
UNETR_96_wCE_DA	0,68	0,04	2,65	0,74	42,48	16,92	1270
UNETR_128_wCE_DA	0,66	0,04	2,87	0,98	46,83	16,54	1374
UNETR_64_wCE_DA	0,67	0,04	2,80	0,72	54,16	17,58	1596

B.5.4 Input T1-w+FLAIR, GT FLAIR – reference cT1-w

Table B 29: Metrics of 3D U-Net combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
3DUNET_128_Dice_noDA	0,71	0,05	1,94	0,50	15,51	10,61	512
3DUNET_96_Dice_noDA	0,72	0,05	1,82	0,52	18,29	9,93	573
3DUNET_128_CE_DA	0,72	0,06	1,85	0,55	18,71	13,06	627
3DUNET_128_DiceCE_noDA	0,71	0,05	1,94	0,47	19,04	9,53	580
3DUNET_96_CE_noDA	0,71	0,06	2,42	1,88	19,32	10,52	598
3DUNET_64_CE_noDA	0,71	0,05	2,08	0,64	21,82	14,06	709
3DUNET_128_Dice_DA	0,72	0,06	2,50	1,93	22,18	15,35	720
3DUNET_64_DiceCE_DA	0,71	0,05	12,69	37,99	22,42	13,98	713
3DUNET_128_CE_noDA	0,71	0,05	2,11	0,53	22,53	11,63	690
3DUNET_64_Dice_noDA	0,72	0,05	1,85	0,55	22,99	11,62	705
3DUNET_64_Dice_DA	0,72	0,05	1,87	0,83	23,54	14,35	749
3DUNET_64_CE_DA	0,70	0,06	2,00	0,62	24,75	15,75	796
3DUNET_96_Dice_DA	0,72	0,05	10,04	30,75	25,25	13,17	768
3DUNET_64_DiceCE_noDA	0,72	0,05	1,88	0,52	25,38	12,29	771
3DUNET_96_CE_DA	0,71	0,06	1,92	0,65	25,57	14,16	792
3DUNET_128_DiceCE_DA	0,71	0,06	1,93	0,58	27,64	13,90	833
3DUNET_96_DiceCE_DA	0,72	0,05	1,83	0,57	28,68	12,71	853
3DUNET_96_DiceCE_noDA	0,70	0,05	1,98	0,52	29,37	13,92	883
3DUNET_128_wCE_noDA	0,67	0,05	2,25	0,71	63,05	16,72	1822
3DUNET_96_wCE_noDA	0,67	0,05	2,26	0,68	66,67	17,30	1924
3DUNET_128_wCE_DA	0,66	0,05	2,29	0,70	78,99	22,97	2308
3DUNET_64_wCE_noDA	0,65	0,04	2,70	1,17	79,75	19,19	2309
3DUNET_96_wCE_DA	0,65	0,05	2,56	1,01	85,94	21,36	2487
3DUNET_64_wCE_DA	0,64	0,05	2,62	0,94	89,60	21,62	2610

Table B 30: Metrics of nnU-Net (DynUNET) combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
DynUNET_96_DiceCE_DA	0,72	0,05	1,84	0,57	17,17	11,80	565
DynUNET_64_DiceCE_DA	0,70	0,06	12,62	35,99	17,28	13,58	593
DynUNET_96_DiceCE_noDA	0,71	0,06	2,29	0,76	18,81	10,37	589
DynUNET_128_DiceCE_DA	0,72	0,05	1,80	0,54	19,97	13,38	668
DynUNET_128_Dice_noDA	0,72	0,05	1,80	0,57	20,12	15,10	687
DynUNET_128_CE_DA	0,72	0,05	1,73	0,58	20,12	11,59	644
DynUNET_64_DiceCE_noDA	0,70	0,06	42,35	66,49	20,98	15,05	698
DynUNET_64_CE_noDA	0,72	0,05	2,13	0,94	21,77	14,72	723
DynUNET_64_Dice_DA	0,67	0,07	79,59	65,79	22,13	17,70	778
DynUNET_96_Dice_noDA	0,72	0,05	1,82	0,69	22,58	11,53	707
DynUNET_128_DiceCE_noDA	0,72	0,05	2,01	0,71	22,84	15,91	757
DynUNET_128_Dice_DA	0,72	0,05	1,75	0,56	24,32	15,73	798

Appendix B

DynUNET_64_CE_DA	0,72	0,05	1,90	0,56	24,48	13,32	769
DynUNET_96_Dice_DA	0,72	0,06	15,56	36,66	24,86	14,40	804
DynUNET_128_CE_noDA	0,71	0,05	1,97	0,67	25,06	13,38	785
DynUNET_96_CE_DA	0,72	0,06	2,10	1,09	25,18	11,01	758
DynUNET_96_CE_noDA	0,71	0,05	2,02	0,64	26,27	14,21	821
DynUNET_64_Dice_noDA	0,70	0,05	63,66	62,34	28,41	18,00	913
DynUNET_96_wCE_noDA	0,70	0,05	2,04	0,66	39,09	12,95	1159
DynUNET_128_wCE_noDA	0,68	0,05	2,11	0,65	46,99	16,49	1389
DynUNET_64_wCE_noDA	0,67	0,05	11,88	35,40	66,62	16,76	1957
DynUNET_128_wCE_DA	0,66	0,06	2,49	1,07	74,92	20,38	2187
DynUNET_96_wCE_DA	0,66	0,05	2,42	0,74	77,91	17,91	2272
DynUNET_64_wCE_DA	0,64	0,05	14,98	43,75	94,54	24,30	2756

Table B 31: Metrics of UNETR combinations sorted by mean Percentage Volume Difference.

Deep Neural Network	Dice mean	Dice sd	95% HD mean	95% HD sd	% Vol Diff mean	% Vol Diff sd	RMSE
UNETR_128_Dice_noDA	0,70	0,06	1,99	0,60	19,09	10,27	601
UNETR_128_CE_DA	0,71	0,05	2,07	0,63	19,65	13,15	641
UNETR_96_CE_noDA	0,70	0,06	2,11	0,64	20,00	10,56	622
UNETR_64_DiceCE_DA	0,72	0,05	1,99	0,72	20,29	15,02	684
UNETR_64_CE_noDA	0,71	0,05	2,15	0,80	24,82	13,27	776
UNETR_64_DiceCE_noDA	0,71	0,05	2,03	0,86	24,89	13,02	773
UNETR_96_DiceCE_DA	0,71	0,05	10,02	30,70	25,81	14,92	809
UNETR_64_CE_DA	0,72	0,05	1,86	0,58	26,43	14,44	830
UNETR_128_CE_noDA	0,70	0,05	2,21	0,56	28,48	13,06	851
UNETR_128_DiceCE_DA	0,71	0,06	2,25	1,67	28,63	14,22	876
UNETR_96_Dice_noDA	0,71	0,05	2,10	0,63	29,07	12,20	870
UNETR_96_CE_DA	0,71	0,05	2,00	0,61	29,36	13,37	881
UNETR_96_Dice_DA	0,72	0,05	1,87	0,55	29,50	14,68	909
UNETR_64_Dice_noDA	0,72	0,05	2,02	0,64	31,41	14,01	954
UNETR_96_DiceCE_noDA	0,70	0,05	2,07	0,62	32,43	12,72	965
UNETR_128_Dice_DA	0,71	0,06	2,23	0,79	32,89	15,02	993
UNETR_128_DiceCE_noDA	0,69	0,05	2,10	0,58	32,94	12,76	968
UNETR_64_Dice_DA	0,68	0,07	119,52	50,75	36,69	22,71	1158
UNETR_128_wCE_noDA	0,68	0,05	2,27	0,61	45,78	13,42	1340
UNETR_96_wCE_noDA	0,68	0,05	2,31	0,70	49,12	15,70	1437
UNETR_96_wCE_DA	0,67	0,05	2,47	1,03	68,13	21,48	1994
UNETR_128_wCE_DA	0,66	0,05	2,40	0,69	68,68	18,60	2003
UNETR_64_wCE_noDA	0,66	0,05	2,56	1,00	71,45	17,78	2086
UNETR_64_wCE_DA	0,67	0,05	2,39	0,79	74,70	20,64	2178

LIST OF ACRONYMS

Acronym	Meaning
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative dataset
AI	Artificial Intelligence
BBB	Brain-Blood-Barrier
BCSFB	Blood-Cerebrospinal-Fluid-Barrier
(C)GMM	(Constrain) Gaussian Mixture Model
ChP	Choroid Plexus
CNN	Connected Neural Networks
CNS	Central Nervous System
CSF	Cerebrospinal Fluid
DC	Dice Metric
Dice	Generalized Dice Loss
DiceCE	Combined Dice and CE Loss
DNN	Deep learning Neural Network
EM	Expectation Maximization algorithm
FCN	Fully Convolutional Neural Networks
FLAIR	Fluid Attenuation Inversion Recovery MRI
FS	FreeSurfer
GD	Gaussian Distribution
GELU	Gaussian Error Linear-Unit operation
GPU	Graphics Processing Units
GT	Ground Truth
HCP	Human Connectome Project dataset
HD	Hausdorff Distance
IS	Immune System
LI	Linear Interpolation
MDD	Major Depressive Disorder
MLP	Multi-Layer Perceptron
MONAI	Medical Open Network for Artificial Intelligence

List of Acronyms

MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
MSeg	Manual Segmentation
MSA	Multi-head Self-Attention layer
NLP	Natural Language Processing
NNI	Nearest-Neighbors Interpolation
OLS	Ordinary Least Square – Linear regression without intercept
PC	Pearson’s Linear Correlation Coefficient
PET	Positron Emission Tomography
PMS	Progressive Multiple Sclerosis
PRELU	Parametric Rectified Linear-Unit operation
RAS	Right, Anterior, Superior orientation specifications
ReLU	Rectified Linear-Unit operation
(R)MSE	(Root) Mean Squared Error
RRMS	Relapsing Remitting Multiple Sclerosis
SGD	Stochastic Gradient Descent method
cT1-w	T1-weighted MRI after contrast injection
T1-w	T1-weighted MRI without contrast injection
(w)CE	(Weighted) Cross-Entropy Loss
WM	White Matter

BIBLIOGRAPHY

- Althubaity, N., Schubert, J., Martins, D., Yousaf, T., Nettis, M. A., Mondelli, V., Pariante, C., Harrison, N. A., Bullmore, E. T., Dima, D., Turkheimer, F. E., & Veronese, M. (2022). Choroid plexus enlargement is associated with neuroinflammation and reduction of blood brain barrier permeability in depression. *NeuroImage: Clinical*, 33. <https://doi.org/10.1016/j.nicl.2021.102926>
- Balafar, M. A. (2014). Gaussian mixture model based segmentation methods for brain MRI images. *Artificial Intelligence Review*, 41(3), 429–439. <https://doi.org/10.1007/s10462-012-9317-3>
- Balusu, S., Brkic, M., Libert, C., & Vandenbroucke, R. E. (2016). The choroid plexus-cerebrospinal fluid interface in Alzheimer’s disease: More than just a barrier. In *Neural Regeneration Research* (Vol. 11, Issue 4, pp. 534–537). Editorial Board of Neural Regeneration Research. <https://doi.org/10.4103/1673-5374.180372>
- Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. (2019). *Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory & Practice*. https://doi.org/10.1007/978-3-030-32245-8_11
- Carloni, S., Bertocchi, A., Mancinelli, S., Bellini, M., Erreni, M., Borreca, A., Braga, D., Giugliano, S., Mozzarelli, A. M., Manganaro, D., Fernandez Perez, D., Colombo, F., di Sabatino, A., Pasini, D., Penna, G., Matteoli, M., Lodato, S., & Rescigno, M. (2021). Identification of a choroid plexus vascular barrier closing during intestinal inflammation. *Science*, 374(6566), 439–448. <https://doi.org/10.1126/science.abc6108>
- Crum, W. R., Camara, O., & Hill, D. L. G. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11), 1451–1461. <https://doi.org/10.1109/TMI.2006.880587>
- Damkier, H. H., Brown, P. D., & Praetorius, J. (2013). Cerebrospinal fluid secretion by the choroid plexus. *Physiological Reviews*, 93(4), 1847–1892. <https://doi.org/10.1152/physrev.00004.2013>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38. <https://www.jstor.org/stable/2984875>
- Dice, L. R. (1945). MEASURES OF THE AMOUNT OF ECOLOGIC ASSOCIATION BETWEEN SPECIES. *Ecology*, 26, 297–302. <http://www.jstor.org/stable/1932409>
- Dosovitskiy, A., Beyler, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Hounsby, N. (2021, June 3). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021*. <http://arxiv.org/abs/2010.11929>
- Dosovitskiy, A., Springenberg, T., Riedmiller, M., & Brox, T. (n.d.). *Discriminative Unsupervised Feature Learning with Convolutional Neural Networks*.
- Engelhardt, B., Wolburg-Buchholz, K., & Wolburg, H. (2001). Involvement of the Choroid Plexus in Central Nervous System Inflammation. *Microscopy Research and Technique*, 52(1), 112–129.

Bibliography

- Fischl, B. (2012). FreeSurfer. In *NeuroImage* (Vol. 62, Issue 2, pp. 774–781).
<https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33(3), 341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)
- Fleischer, V., Gonzalez-Escamilla, G., Ciolac, D., Albrecht, P., Küry, P., Gruchot, J., Dietrich, M., Hecker, C., Müntefering, T., Bock, S., Oshaghi, M., Radetz, A., Cerina, M., Krämer, J., Wachsmuth, L., Faber, C., Lassmann, H., Ruck, T., Meuth, S. G., ... Groppa, S. (2021). Translational value of choroid plexus imaging for tracking neuroinflammation in mice and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 118(36). <https://doi.org/10.1073/pnas.2025000118>
- Ghosh, S., Das, N., Das, I., & Maulik, U. (2019). Understanding deep learning techniques for image segmentation. *ACM Computing Surveys*, 52(4). <https://doi.org/10.1145/3329784>
- Greenspan, H., Ruf, A., & Goldberger, J. (2006). Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9), 1233–1245. <https://doi.org/10.1109/TMI.2006.880668>
- Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., & Xu, D. (2021). *UNETR: Transformers for 3D Medical Image Segmentation*.
- Hendrycks, D., & Gimpel, K. (2016). *Gaussian Error Linear Units (GELUs)*.
<http://arxiv.org/abs/1606.08415>
- Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9), 850–863. <https://doi.org/10.1109/34.232073>
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- Kingma, D. P., & Ba, J. (2015, December 22). Adam: A Method for Stochastic Optimization. *ICLR*. <http://arxiv.org/abs/1412.6980>
- Lassmann, H. (2019). Pathogenic mechanisms associated with different clinical courses of multiple sclerosis. *Frontiers in Immunology*, 9(3116).
<https://doi.org/10.3389/fimmu.2018.03116>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
<https://doi.org/10.1038/nature14539>
- Lizano, P., Lutz, O., Ling, G., Lee, A. M., Eum, S., Bishop, J. R., Kelly, S., Pasternak, O., Clementz, B., Pearlson, G., Sweeney, J. A., Gershon, E., Tamminga, C., & Keshavan, M. (2019). Association of Choroid Plexus Enlargement With Cognitive, Inflammatory, and Structural Phenotypes Across the Psychosis Spectrum. *American Journal of Psychiatry*, 176(7), 564–572. <https://doi.org/10.1176/appi.ajp.2019.18070825>
- Long, J., Shelhamer, E., & Darrell, T. (2014). *Fully Convolutional Networks for Semantic Segmentation*. <http://arxiv.org/abs/1411.4038>
- Loshchilov, I., & Hutter, F. (2019, November 14). Decoupled Weight Decay Regularization. *ICLR*. <http://arxiv.org/abs/1711.05101>

Bibliography

- Maekawa, T., Hori, M., Murata, K., Feiweier, T., Andica, C., Fukunaga, I., Koshino, S., Hagiwara, A., Kamiya, K., Kamagata, K., Wada, A., Abe, O., & Aoki, S. (2019). Choroid plexus cysts analyzed using diffusion-weighted imaging with short diffusion-time. *Magnetic Resonance Imaging*, *57*, 323–327. <https://doi.org/10.1016/j.mri.2018.12.010>
- Manouchehri, N., & Stüve, O. (2021). Choroid plexus volumetrics and brain inflammation in multiple sclerosis. *Proceedings of the National Academy of Sciences*, *118*(40). <https://doi.org/10.1073/pnas.2115221118>
- Marques, F., Sousa, J. C., Brito, M. A., Pahnke, J., Santos, C., Correia-Neves, M., & Palha, J. A. (2017). The choroid plexus in health and in disease: dialogues into and out of the brain. *Neurobiology of Disease*, *107*, 32–40. <https://doi.org/10.1016/j.nbd.2016.08.011>
- Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*.
- Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2021.3059968>
- Müller, J., Sinnecker, T., Wendebourg, M. J., Schläger, R., Kuhle, J., Schädelin, S., Benkert, P., Derfuss, T., Cattin, P., Jud, C., Spiess, F., Amann, M., Lincke, T., Barakovic, M., Cagol, A., Tsagkas, C., Parmar, K., Pröbstel, A.-K., Reimann, S., ... Yaldizli, Ö. (2022). Choroid Plexus Volume in Multiple Sclerosis vs Neuromyelitis Optica Spectrum Disorder. *Neurology - Neuroimmunology Neuroinflammation*, *9*(3), e1147. <https://doi.org/10.1212/NXI.0000000000001147>
- Pandis, N. (2016). Linear regression. *American Journal of Orthodontics and Dentofacial Orthopedics*, *149*(3), 431–434. <https://doi.org/10.1016/j.ajodo.2015.11.019>
- Ricigliano, V. A. G., Morena, E., Colombi, A., Tonietto, M., Hamzaoui, M., Poirion, E., Bottlaender, M., Gervais, P., Louapre, C., Bodini, B., & Stankoff, B. (2021). Choroid Plexus Enlargement in Inflammatory Multiple Sclerosis: 3.0-T MRI and Translocator Protein PET Evaluation. *Radiology*, *301*(1), 166–177. <https://doi.org/10.1148/radiol.2021204426>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.
- Ruby Usha, & Yendapalli Vamsidhar. (2020). Binary cross entropy with deep learning technique for Image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5393–5397. <https://doi.org/10.30534/ijatcse/2020/175942020>
- Schmidt-Mengin, M., Ricigliano, V. A. G., Bodini, B., Morena, E., Colombi, A., Hamzaoui, M., Panah, A. Y., Stankoff, B., & Colliot, O. (2021). *Axial multi-layer perceptron architecture for automatic segmentation of choroid plexus in multiple sclerosis*.
- Schubert, J. J., Veronese, M., Marchitelli, L., Bodini, B., Tonietto, M., Stankoff, B., Brooks, D. J., Bertoldo, A., Edison, P., & Turkheimer, F. E. (2019). Dynamic 11C-PIB PET shows cerebrospinal fluid flow alterations in Alzheimer disease and multiple sclerosis. *Journal of Nuclear Medicine*, *60*(10), 1452–1460. <https://doi.org/10.2967/jnumed.118.223834>
- Spector, R., Keep, R. F., Robert Snodgrass, S., Smith, Q. R., & Johanson, C. E. (2015). A balanced view of choroid plexus structure and function: Focus on adult humans. *Experimental Neurology*, *267*, 78–86. <https://doi.org/10.1016/j.expneurol.2015.02.032>

- Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). *Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations*.
https://doi.org/10.1007/978-3-319-67558-9_28
- Tadayon, E., Moret, B., Sprugnoli, G., Monti, L., Pascual-Leone, A., & Santarnecchi, E. (2020). Improving Choroid Plexus Segmentation in the Healthy and Diseased Brain: Relevance for Tau-PET Imaging in Dementia. *Journal of Alzheimer's Disease*, 74(4), 1057–1068.
<https://doi.org/10.3233/JAD-190706>
- Tadayon, E., Pascual-Leone, A., Press, D., & Santarnecchi, E. (2020). Choroid plexus volume is associated with levels of CSF proteins: relevance for Alzheimer's and Parkinson's disease. *Neurobiology of Aging*, 89, 108–117. <https://doi.org/10.1016/j.neurobiolaging.2020.01.005>
- Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., ... Cohen, J. A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2), 162–173. [https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2)
- Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., & Dosovitskiy, A. (2021). *MLP-Mixer: An all-MLP Architecture for Vision*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention Is All You Need. *NIPS*.
<http://arxiv.org/abs/1706.03762>
- Vercellino, M., Votta, B., Condello, C., Piacentino, C., Romagnolo, A., Merola, A., Capello, E., Mancardi, G. L., Mutani, R., Giordana, M. T., & Cavalla, P. (2008). Involvement of the choroid plexus in multiple sclerosis autoimmune inflammation: A neuropathological study. *Journal of Neuroimmunology*, 199(1–2), 133–141.
<https://doi.org/10.1016/j.jneuroim.2008.04.035>
- Willmott, C. J. (1982). Some Comments on the Evaluation of Model Performance. *Bulletin of the American Meteorological Society*, 63(11), 1309–1313.
[https://doi.org/https://doi.org/10.1175/1520-0477\(1982\)063%3C1309:SCOTEO%3E2.0.CO;2](https://doi.org/https://doi.org/10.1175/1520-0477(1982)063%3C1309:SCOTEO%3E2.0.CO;2)
- Yushkevich, P. A., Gao, Y., & Gerig, G. (2016). ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2016-October*, 3342–3345. <https://doi.org/10.1109/EMBC.2016.7591443>
- Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31(3), 1116–1128.
<https://doi.org/10.1016/j.neuroimage.2006.01.015>
- Zhao, L., Feng, X., Meyer, C. H., & Alsop, D. C. (2020). Choroid Plexus Segmentation Using Optimized 3D U-Net. *Proceedings - International Symposium on Biomedical Imaging, 2020-April*, 381–384. <https://doi.org/10.1109/ISBI45749.2020.9098443>
- Zhou, Y.-F., Huang, J.-C., Zhang, P., Fan, F.-M., Chen, S., Fan, H.-Z., Cui, Y.-M., Luo, X.-G., Tan, S.-P., Wang, Z.-R., Feng, W., Yuan, Y., Yang, F.-D., Savransky, A., Ryan, M., Goldwasser, E., Chiappelli, J., Rowland, L. M., Kochunov, P., ... Hong, L. E. (2020).

Bibliography

Choroid Plexus Enlargement and Allostatic Load in Schizophrenia. *Schizophrenia Bulletin*, 46(3), 722–731. <https://doi.org/10.1093/schbul/sbz100>

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., Wells III, W. M., Jolesz, F. A., Kikinis, R., & St, F. (2004). Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index 1 : Scientific Reports. *Acad Radiol*, 11(2), 178–189. [https://doi.org/10.1016/s1076-6332\(03\)00671-8](https://doi.org/10.1016/s1076-6332(03)00671-8)