UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Dipartimento di Ingegneria dell'Informazione

Corso di laurea magistrale in Bioingegneria per le Neuroscienze

ANALYSIS AND ADAPTATION OF NN-UNET (NO-NEW-UNET) TO PROCESS SINGLE MODALITIES MAGNETIC RESONANCE IMAGES FOR BRAIN TUMORS AND ISCHEMIC STROKE LESIONS SEGMENTATION

Relatore: Prof. Manfredo Atzori

Correlatore: Dott. Louis Fabrice Tshimanga

Laureando/a: Andrea Viberti

ANNO ACCADEMICO 2021 – 2022

Data di laurea 28/11/2022

Index

Abstract

Automatic segmentation of brain tumors and stroke lesions from medical images using deep learning algorithms is crucial for prognosis, clinical assessment and treatment planning, and provides valuable clinical information. The analysis of the current state-of-the-art both for brain tumors and stroke lesions segmentations pointed out the most efficient techniques in these fields. It also underlined that most of them treat input MR images acquired with distinct modalities without caring of the divergence between the intensities with which the different cerebral and tumoral subregions are represented in these different modalities. Moreover, it was highlighted the almost complete absence of deep learning algorithms dealing with both brain tumors and stroke lesions segmentation. The main objective of this thesis was therefore to separate input images into the different available modalities, so that features extracted from images with divergent intensities may not be fused at earl levels, and to develop an Inter-pathology Learning technique between brain tumor and stroke lesion segmentation models, to transfer knowledge between those fields. The most promising and efficient model was identified in nnUNet, a self-adaptive framework that automatically adapts its network architecture to the specific task and dataset.

Two different methods to separate input images of different MR modalities where thus developed: the first based on ensemble of models, while the second consisting in a multi-path network, modifying nnUNet by creating one different encoder for each input modality including dense connections, and resulting in the creation of Dense Multi-path nnUNet, which was the most promising one. The Dense Multi-path nnUNet models were then trained and evaluated using FeTS 2022 dataset for Brain Tumor Segmentation, while using ISLES 2022 dataset for Stroke Lesion Segmentation, being able to obtain Dice scores for the three tumoral regions (ED, NCR and ET) of 0.886, 0.823 and 0.903, with an average of 0.871; while the Dice score obtained for the segmentation of the stroke lesion was 0.660, overcoming the performances of nnUNet in both cases.

An Inter-pathology Learning technique was also developed between the brain tumor segmentation model trained with FLAIR images, and the corresponding stroke lesion segmentation model trained with FLAIR images, showing superior performances of the basic model trained to segment stroke lesions.

1. Introduction

Before presenting the goals of this thesis, a short introduction about brain tumors and stroke lesions is performed. Their treatment is strictly dependent from the accurate and quick identification of the lesion from medical images, and the consequent resection (of the tumoral mass) or analysis and care (of the stroke lesion). For these reasons, the acquired imaging modalities in these aims are analyzed, and a description of the most used methods for lesions segmentation is realized.

Tumors of the nervous system are growing masses of neoplastic and abnormal cells in the brain or spinal cord which can have different origins. Brain Tumors are tumoral forms that affects the Central Nervous System (CNS), and can be classified in: primary tumors, which develop directly from cells of the central nervous system; and secondary tumors, or metastasis, that arise from tumors that originate in other organs and later diffuse in the nervous system. Primary brain tumors cause 2% of all cancer deaths in Europe. The most diffuse type of primary brain tumors are Gliomas, whose name derives from their origin in the brain's glial cells (which provide support and stability to neurons to keep them healthy and guide their development), and it represents 81% of all malignant brain tumors (Ostrom et al., 2014) and 45% of all primary brain tumors (Liu et al., 2016). Moreover, the World Health Organization (WHO) has classified Gliomas in 4 grades based on their degree of severity and their level of malignancy or benignancy (Tiwari et al., 2020): I) adult-type diffuse gliomas, II) pediatric-type diffuse low-grade gliomas (LGG), III) pediatric-type diffuse high-grade gliomas (HGG), and IV) circumscribed astrocytic gliomas. The difference between low-grade gliomas and high-grade gliomas is that LGG often grow slower, can be benign or malignant but can evolve in HGG; they can be treated with radiotherapy, chemotherapy and surgery. While HGG are malignant, with high mortality, they grow rapidly and aggressively and they are often forming a necrotic core, with surrounding oedema and swelling (Kamnitsas, Bai, et al., 2017). Moreover, gliomas can be classified also based on the glial cells affected; among those, Glioblastoma represents the most aggressive and malignant form.

On the other hand, stroke is one of the most common cerebrovascular diseases, and one of the main causes of long-term disabilities and mortality worldwide, affecting one in six adults, with an estimated 3-6 million cases annually (Praveen et al., 2018). There are two types of stroke: hemorrhagic stroke, which is due to bleeding inside the brain caused by the rupture of a blood

vessel; and ischemic stroke, which is due to the blockage of a blood vessel which causes the reduction or loss of blood supply to a region of the brain. In either case, parts of the brain become damaged or die, causing brain damage, long-term disability or even death. Ischemic stroke represents 80-85 % of all strokes (Praveen et al., 2018) and can be caused by atherosclerosis, hypertension, physical inactivity, while age, gender and ethnicity represent risk factors. Moreover, stroke can be divided by the duration of the episode, in acute (0 – 24 h), sub-acute (24h – 2 weeks), and chronic (more than 2 weeks). Clinical decisions and timing of interventions are essential for treatment at any stage and to limit subsequent damage: at the acute stage, early intervention can facilitate short-term functional recovery, while at sub-acute or chronic stages, treatment can promote long-term recovery.

Medical imaging is essential for diagnosis, quantitative evaluation and treatment planning of both ischemic strokes and gliomas. Magnetic resonance imaging (MRI) is the most effective technique to generate multi-modal images to identify and analyze different tumor regions; the most used MRI modalities for Glioma diagnosis are: T2-weighted fluid attenuated inversion recovery (FLAIR), T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), and T2-weighted (T2) (ben naceur et al., 2020). MRI represents also the preferred imaging method for the treatment of ischemic stroke, but the lesion should be located and quantified within 3 hours from the onset of the stroke event, and MRI is not suitable for this scope because it's a slow imaging technique. For acute strokes, computed tomography perfusion (CTP) is preferred, because it's faster and cheaper (Y. Zhang et al., 2022).

Lesion segmentation allows accurate delineation of brain tumor region and subregions, and of ischemic stroke lesion: it represents the identification of lesions' outline on medical images, by classifying each voxel as lesion or non-lesion. The goal of brain tumor segmentation is tumor resection, that is to remove as much of the tumor as possible to minimize the chance of recurrence avoiding injuries to vital brain areas; while ischemic stroke lesion segmentation is crucial to estimate the location and volume of the lesion, to assess brain damage and possible risk factors. The preferred way is to have a specialized neuro-radiologist performing manual segmentation of lesions, but this is a tedious, time-consuming procedure, with high inter-rater disagreement. Moreover, it becomes really difficult in the case of gliomas, which have different intra-tumoral structures: Necrotic and Non-Enhancing tumor, Peritumoral Edema, Enhancing tumor (ben naceur et al., 2020).

The introduction of deep learning represents a breakthrough for artificial intelligence. Deep Learning is a machine learning branch based on Artificial Neural Networks (ANNs), whose structure was inspired by information processing and communication between neurons inside

the brain: briefly, a network is composed by layers of nodes (neurons), in which each node receives many inputs from neurons of the previous layers, performs a weighted sum of them and its output is obtained by the application of an activation function. The weights are learnt by the network during training thanks to the backpropagation process, in which the gradient is computed to minimize a specific loss function over the weights. Automatic lesion segmentation is then performed by deep learning algorithms: they allow segmentation of unlabeled medical images by using a model trained on manually annotated lesions, used as training cases; this can better assist diagnosis, prognosis, treatment planning and evaluation, playing a crucial role in image understanding, feature extraction, analysis and interpretation (Wadhwa et al., 2019). In recent years, Convolutional Neural Networks (CNNs) are becoming the most common type of deep learning frameworks used for segmentation; they consist in a combination of convolutional layers, normalization layers and pooling layers that allow to automatically extract statistical rules from the data and use these rules to predict and analyze unseen data. Among CNNs, the state-of-the-art models for image segmentation are variants of encoder-decoder architectures like U-Nets.

In this thesis, the most recent and best techniques used for brain tumors and ischemic stroke lesions segmentation are analyzed, implemented, modified and tested in different conditions on some of the most recent and comprehensive datasets in this fields, which are mainly Brain Tumor Segmentation (BraTS) challenge datasets and Ischemic Stroke Lesion Segmentation (ISLES) challenge datasets.

Among these methods, I focused on some, considered the most interesting ones, like nnUNet, a deep learning framework that allows to achieve 3D semantic segmentation on a lot of different biomedical applications.

It is also known that gliomas can have the same appearance and shape than ischemic stroke in MRI data (Zhao et al., 2018), but the number of techniques developed for the segmentation of both ischemic stroke lesions and glioma brain tumors is really limited, so is it possible to achieve the new state-of-the-art with this dual aim?

2. Related works

A state-of-the-art analysis was performed in the fields of Brain Tumor and Stroke Lesion Segmentation, with the goals of identifying the newest techniques and algorithms developed in these scopes, the ones achieving the highest performances, and also those implementing peculiar and innovative structures. The aim of this study was to identify and point out possible candidates that could be used as baselines for further research.

Methods were identified searching articles published after 2015, year in which UNet network was introduced, which marked a breakthrough in images segmentation task.

First of all, the most performing models in popular challenge were picked out, analyzing challenges in the domains of interest. Moreover, articles were found using specific key words in Google Scholar, or starting from analyzed reviews. A table was produced after each field studied (*Table 2.1, Table 2.2*), in which specific colors were used to highlight disparate aspects:

- Orange was used to underline the most interesting articles, mainly related to the usage of 3D CNNs, which nowadays represent indispensable architectures to perform well on medical images segmentation;
- Yellow was used to point remarkable articles, but involving segmentation of medical images outside the domain of brain tumors and stroke lesions;
- Green was used to highlight limitedly cited/interesting articles, but implying the introduction of structures or architectures judged relevant or unique.

2.1 Brain Tumor Segmentation

In the field of Brain Tumor Segmentation, at first articles were searched starting from the winning and top performing methods in BraTS 2021 challenge, which is the most recent brain tumor segmentation challenge, sanctioning then the current state-of-the-art.

Afterwards, algorithms were identified searching in Google Scholar, using specific keywords like “brain tumor segmentation” or “automatic brain tumor segmentation”.

Finally, many articles were found out starting from the following analyzed reviews: “State of the art survey on MRI brain tumor segmentation” (Gordillo et al., 2013), “A review on brain

tumor segmentation of MRI images" (Wadhwa et al., 2019), "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019" (Tiwari et al., 2020), "Deep Learning for Brain Tumor Segmentation: A Survey of State-of-the-Art" (Magadza & Viriri, 2021).

A brief description is provided below for the most interesting methods, while details about architectures and respective performances can be visualized in *Table2.1a* and *Table 2.1b*.

Starting from the top performing methods in BraTS 2021, Isensee et al. introduced in 2018 nnUNet, a self-adapting framework able to perform semantic segmentation in many different biomedical domains, without needing to design a specific solution and architecture for the given task or dataset. It was the winning method of BraTS 2020 challenge, but it was applied for the segmentation in many different biomedical domains, achieving a new state-of-the-art in several of these (Isensee et al., 2020a). Furthermore, an extended version of nnUNet was introduced by H.M. Luu et al. and took the first place in BraTS 2021 challenge; the architecture remained nnUNet, but with Group Normalization instead of batch normalization, because it works better with a reduced batch size, a larger encoder, and with the use of axial attention in the decoder (Luu & Park, 2021). Further analysis on these architectures will be performed in the following chapters. Siddiquee & Myronenko introduced SegResNet, an encoder (with ResNet blocks and skip connections)-decoder network with instance normalization; the most efficient architecture was identified carrying out an ensemble of the most performing models between the trained networks, and executing an average with equal weights. It ranked second in BraTS 2021.

L. Fidon et al. took part to BraTS 2021 with a network consisting in an efficient test-time ensemble of seven 3D networks, both traditional Unets and their transformer variations, TransUNets (including a vision transformer in the bottleneck), exploring different learning schemes. They introduced the Generalized Wasserstein Dice Loss, able to improve segmentation performances (Fidon et al., 2021). M. Futrega et al. took third place in BraTS 2021 challenge, by realizing an ablation study to select the optimal UNet variant, testing: classic UNet; SegResNetVAE, obtained by adding to SegResNet a Variational Autoencoder (VAE) (Myronenko, 2018); UNETR, a UNet in which the encoder is a generalization of a Vision Transformer (ViT) (Nvidia et al., n.d.); Attention UNet, with an attention gate added in the decoder (Oktay et al., 2018); Residual UNet (He et al., n.d.); finding out that UNet remains the most performing one. They also optimized the model by increasing the encoder path, the number of convolutional filters, etc. (Futrega et al., 2021). H. Jia et al. developed HNF-Netv2, which represents an extension of HNF-Net by adding interscale and intra-scale semantic discrimination enhancing blocks to exploit global semantic discrimination; it ranked 8th in

BraTS 2021 challenge (Jia et al., 2022). A. Hatamizadeh et al. proposed Swin UNETR, a network exploiting UNETR architecture but using a hierarchical Swin Transformer as encoder, which extracts features at five different resolutions, and is connected to a FCNN-based decoder at each resolution step via skip connections (Hatamizadeh et al., 2022). In 2016 K. Kamnitsas et al. introduced DeepMedic, a 3D CNN with eleven layers, that consists in two parallel convolutional paths that work at different scales (the second path works with subsampled images) with residual connections between outputs of every two layers (Kamnitsas et al., n.d.). In 2017 they improved DeepMedic by adding a 3D fully connected Conditional Random Field (CRF), that consists in a post-processing regularization method which removes false positives modeling dependencies between neighboring pixels. It was also developed a specific training pipeline (dense training) which allowed to raise batch size without increasing memory consumption (Kamnitsas, Ledig, et al., 2017). M. Havaei et al. in 2015 studied different CNN architectures, leading to the identification of InputCascadeCNN, a cascade of two 2D CNNs, in which the output of the first CNN becomes the input of the second, by substituting the last fully connected layer with a convolutional output layer. Each network has a local and a global pathway, to extract corresponding features (Havaei et al., 2015). In 2019 Hu et al. extended InputCascadeCNN by proposing a multi-cascaded convolutional neural network (MCCNN), based on the cascade of three cascaded networks trained respectively with axial, coronal and sagittal images, followed by a Conditional Random Field (K. Hu et al., 2019). A. Myronenko won the 1st place in BraTS 2018 challenge with an encoder-decoder architecture, characterized by a large encoder to extract deep features; it was also added a Variational Autoencoder, used only during training to reconstruct input images and regularize the shared encoder (Myronenko, 2018). G. Wang et al. proposed a cascade of anisotropic CNNs, used to decompose the multi-class segmentation problem with three single-class segmentations in sequence, exploiting the hierarchical structure of tumoral regions (Wnet to segment WT, Tnet to segment TC, Enet to segment ETC) (G. Wang et al., 2017). S. Chen et al. introduced a new dual force training strategy that allows to learn high-quality hierarchical features, and applied it to UNet and MLDeepMedic (DeepMedic modified to extract multi-level features); they also implemented a Multi-Layer Perceptron-based post-processing approach that allows to refine segmentation results (S. Chen et al., 2019). W. Wang et al. implemented TransBTS, an encoder-decoder structure in which the encoder uses a 3D CNN to capture local information, while they exploited a Transformer between encoder and decoder to extract global (long-range) features, by feeding it with tokens obtained elaborating feature maps (W. Wang et al., 2021). Z. Zhou et al. developed a deeply-supervised encoder-decoder network, UNet++, with a series of nested,

dense skip pathways between encoder and decoder, to reduce the semantic gap and increase the combination of feature maps between encoder and decoder; it was applied in the segmentation of multiple medical images: nodule segmentation in the low-dose CT scans of chest, nuclei segmentation in the microscopy images, etc. (Z. Zhou et al., 2018). K. Kamnitsas et al. won the first position in BraTS 2017 with EMMA (Ensembles of Multiple Models and Architectures), which represents an ensemble of DeepMedic, three 3D FCNNs and two 3D UNets, trained following different pipelines to average away the variance, and combined at inference time by calculating for each voxel the average confidence of the individual models for that voxel, and assigning the class with the highest confidence (Kamnitsas, Bai, et al., 2017). M. Chen et al. introduced Deep Convolutional Symmetric Neural Network (DCSNN), which exploits the left-right asymmetry of tumor regions by passing to the network the original image and its left-right flipped version and by using Left-Right Similarity Masks (LRSMs) to extract features (H. Chen et al., 2020). D. Zhang et al. proposed a cross-modality deep feature learning framework, including two learning processes: the cross-modality feature transition (CMFT) process, which consists in a Generative Adversarial Network (GAN) that learns cross-modality features, and a cross-modality feature fusion (CMFF) process, which fuses features previously extracted to generate segmentation maps (D. Zhang et al., 2022). To solve class imbalance, one of the major problems for medical images segmentation, C. Zhou et al. developed a multi-task learning strategy based on incorporating the three brain tumor segmentation tasks into a single model: One-pass Multi-task Network (OM-Net). The three tasks are trained in an increasing order of difficulty and a Cross-talk Guided Attention (CGA) module is added to use the prediction results of preceding tasks to guide the following one (C. Zhou et al., 2019). Z. Jiang et al. designed a two-stage cascaded UNet which ranked 1st in 2019 BraTS challenge. The first stage consists of a UNet variant to obtain a coarse segmentation, while in the second stage the network is larger and with two decoders to increase performances; its inputs are the coarse segmentation maps produced by the previous model, and the raw images (Jiang et al., 2020). Brain SegNet, devised by X. Hu et al., can work efficiently for both brain tumor segmentation and stroke lesion outcome prediction. It consists in a 3D CNN with a Unet-like structure, based on an evolution of ResNet for segmentation, with a designed 3D refinement module to encode for both fine structures and high-level context information (X. Hu et al., 2020). M.T. Duong et al. implemented a 3D UNet to perform automated FLAIR lesion segmentation, following that most brain lesions are characterized by an hyperintense signal on FLAIR images. It was applied to perform segmentation on 19 different brain pathologies (Duong et al., 2019).

To conclude, all details about these cited methods, and many more, can be visualized in *Table 2.1a* and *2.1b,* which have been split for visualization purposes, but contain different information of the same articles.

Title	Authors	Citations	Publication	2D/3D	Aim	Sensitivity	Precision (PPV)	Dice score	HD95	Datasets
Extending nn-UNet for brain tumor segmentation	H.M. Luu et al.	0	2021	3D	BTS	/	/	88.36 (test set)	10.61 (test set)	BraTS 2021 Dataset
Redundancy Reduction in Semantic Segmentation of 3D Brain Tumor MRIs	M.M.R. Siddiquee et al.	0	2021	3D	BTS	/	/	89.11 (val. set)	6.16 (val. set)	BraTS 2021 Dataset
nnU-Net for Brain Tumor Segmentation	F. Isensee et al.	28	2020	3D	BTS	/	/	85.35 (test set)	14.55 (test set)	BraTS 2020 Dataset
Generalized Wasserstein Dice Loss, Test-time Augmentation, and Transformers for the BraTS 2021 challenge	L. Fidon et al.	0	2021	3D	BTS	/	/	89.4 (test set)	10 (test set)	BraTS 2021 Dataset
Optimized U-Net for Brain Tumor Segmentation	M. Futrega et al.	1	2021	3D	BTS	/	/	88.55 (val. set)	/	BraTS 2021 Dataset
HNF-Netv2 for Brain Tumor Segmentation using multi-modal MR Imaging	H. Jia et al.	0	2021	3D	BTS	/	/	89.21 (test set)	9.89 (test set)	BraTS 2021 Dataset
Swin UNETR:Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images	A. Hatamizadeh et al.	1	2021	3D	BTS	/	/	88.53 (test set)	12.12 (test set)	BraTS 2021 Dataset
DeepMedic for Brain Tumor Segmentation	K. Kamnitsas et al.	252	2016	3D	BTS	81.3 (train. Set)	80.83 (train. Set)	79.3 (train. set)	/	BraTS 2015/16 Dataset
Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation	K. Kamnitsas et al.	2367	2017	3D	BTS/SS	71.5, 63 (BraTS test, ISLES)	77.73, 77 (BraTS test, ISLES)	71.5, 66 (BraTS test, ISLES)	55.93 (ISLES)	BraTs 2015/ISLES SISS 2015
A deep learning model integrating FCNNs and CRFs for brain tumor segmentation	X. Zhao et al.	449	2018	2D	BTS	75 (test set)	/	73 (2015 test set)	/	BraTS 2013/2015/2016 Dataset
Brain Tumor Segmentation with Deep Neural Networks	M. Havaei et al.	2344	2017	2D	BTS	82 (test set)	78.67 (specificity)	80 (test set)	/	BraTS 2013 Dataset
Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images	S. Pereira et al.	1762	2016	2D	BTS	84.3 (2013 test set)	83 (2013 test set)	82.67 (2013 test set)	/	BraTS 2013/2015 Dataset
Brain Tumor Segmentation Using Multi-Cascaded Convolutional Neural Networks and Conditional Random Field	K. Hu et al.	73	2019	2D	BTS	85.3, 80.3, 84,6 (2013, 2015, 2018)	74.3, 82, 99.45 (2013, 2015, 2018)	77.67, 79.3, 78.28 (2013,2015, 2018)	9.30 (2018)	BraTS 2013/2015/2018 Dataset
Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM	N.B. Bahadure et al.	359	2017	/	BTS	97.72	94.2 (specificity)	0.82	/	DICOM Dataset, Brain Web Dataset
Deep learning based enhanced tumor segmentation approach for MR brain images	M. Mittal et al.	115	2019	/	BTS	/	98.81	/	/	BRAINIX medical images
3D MRI brain tumor segmentation using autoencoder regularization	A. Myronenko	496	2018	3D	BTS	/	/	82.2 (2018 test set)	4.83 (2018 test set)	BraTS 2018 Dataset (winner)
Automatic Brain Tumor Segmentation using Cascaded Anisotropic Convolutional Neural Networks	G. Wang et al.	376	2017	3D	BTS	/	/	81.18 (test set)	16.5 (test set)	BraTS 2017 Dataset
Dual-force convolutional neural networks for accurate brain tumor segmentation	S. Chen et al.	97	2019	3D	BTS	80.1, 79.78	82.08, 79.57 (2017	78.88, 79.20 (2017	/	BraTS 2015/2017 Dataset
3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation	O. Çiçek et al.	3641	2016	3D	/	/	/	/	/	Microscopic Dataset of Xenopus kidney
TransBTS: Multimodal Brain Tumor Segmentation Using Transformer	W. Wang et al.	36	2021	3D	BTS	/	/	83.62, 83.78 (2019, 2020)	5.14, 10.89 (2019, 202	BraTS 2019/2020 Dataset
No New-Net	F. Isensee et al.	323	2018	3D	BtTS	/	/	82.10 (test set)	4.67 (test set)	BraTS 2018 Dataset
UNet++: A Nested U-Net Architecture for Medical Image Segmentation	Z. Zhou et al.	1728	2018	3D	/	Improvement of Unet architecture evaluated using IoU				Cell nuclei/colon polyp/liver/lung nodule
Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation	K. Kamnitsas et al.	369	2017	3D	BTS	/	/	80 (test set)	21.37 (test set)	BraTS 2017 Dataset (winner)
Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge	F. Isensee et al.	376	2017	3D	BTS	78.67, 82.23 (2015 test, 2017 val)	75.33 (2015 test)	74.33, 76 (2015 test, 2017 test)	7 (2017 val)	BraTS 2015/2017 Dataset
A Fully Automated Deep Learning Network for Brain Tumor Segmentation	C.G.B. Yogananda et al.	18	2020	3D	BTS	82.67, 84.3, 80.67 (2017, 2018, Oslo)	/	82.67, 84, 80.67 (2017, 2018, Oslo)	6.9, 5.97, 5.25	BraTS 2017/2018, Oslo Datasets
Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images	M.A. Naser et al.	59	2020	2D	BTS	92	92 (specificity)	84	/	Cancer Imagig Archive (TCIA) and segmented
Brain tumor segmentation with deep convolutional symmetric neural network	M. Chen et al.	57	2020	3D	BTS	/	/	70.47 (test set)	/	BraTS 2015 Dataset
ERV-Net: An efficient 3D residual neural network for brain tumor segmentation	X. Zhou et al.	20	2021	3D	BTS	87.39 (test set)	/	86.57 (test set)	4.46 (test set)	BraTS 2018 Dataset
RescueNet: An unpaired GAN for brain tumor segmentation	S. Nema et al.	59	2020	2D	BTS	91.38, 91.05 (2015 test, 2017 test)	/	91.87, 91.26 (2015 test, 2017 test)	/	BraTS 2015/2017 Dataset
A novel end-to-end brain tumor segmentation method using improved fully convolutional networks	H. Li et al.	77	2019	2D	BTS	74.47 (2015 test set)	74.27 (2015 test set)	71.43, 76.03 (2015 test, 2017 test)	/	BraTS 2015/2017 Dataset
Cross-Modality Deep Feature Learning for Brain Tumor Segmentation	D. Zhang et al.	47	2021	3D	BTS	/	/	82.8, 84.3 (2017 val, 2018 val)	5.11, 5.12 (2017 val, 2	BraTS 2017/2018 Dataset
One-pass Multi-task Networks with Cross-task Guided Attention for Brain Tumor Segmentation	C. Zhou et al.	68	2020	3D	BTS	/	/	84.48, 85.88 (2017 val, 2018 val)	5.07, 4.9 (2017 val, 201	BraTS 2017/2018 Dataset
Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images	R. Ranjbarzadeh et al.	22	2021	2D	BTS	94.38 (10% of training set)	/	90.14	1.83	BraTS 2018 Dataset
Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy	M. Akil et al.	35	2020	2D	BTS	78.2, 79.03, 76	99.6, 99.6, 99.7 (specificity)	77.63, 78.47, 77.5	9.35, 8.72, 8.75	BraTS 2018 Dataset
Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation	G. Wang et al.	97	2019	2.5D	BTS	/	/	81.07, 80.7 (2017 test, 2018 test)	16.5, 5.61 (2017 test, 2	BraTS 2017/2018 Dataset
Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task	Z. Jiang et al.	128	2019	3D	BTS	/	/	85.25 (test set)	3.8 (test set)	BraTS 2019 Dataset (winner)
Bag of Tricks for 3D MRI Brain Tumor Segmentation	Y.X. Zhao et al.	55	2019	/	BTS	/	/	85.13 (test set)	3.82 (test set)	BraTS 2019 Dataset + Decathlon
Automatic Semantic Segmentation of Brain Gliomas from MRI Images Using a Deep Cascaded Neural Network	S. Cui et al.	104	2018	2D	BTS	82.3	85	84	/	BraTS 2015 Dataset
H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes	X. Li et al.	1026	2018	2D/3D	Liver TS	/	/	82.4, 93.7 (Lesion MICCAI, Tumor 3DIRCADb)	/	MICCAI Liver Tumor Segmentation 2017 + 3DIRCADb
Convolutional neural network with batch normalization for glioma and stroke lesion detection using MRI	J. Amin et al.	23	2020	/	BTS/SS	/	97.2, 1 (BraTS, ISLES)	97.54, 93.69 (BraTS, ISLES)	/	BraTS 2017 + ISLES 2015
A New Approach for Brain Tumor Segmentation and Classification Based on Score Level Fusion Using Transfer Learning	J. Amin et al.	47	2019	/	BTS/SS	/	99.94, 94.89 (BraTS, ISLES)	99.89, 94.66 (BraTS, ISLES)	/	BraTS 2017 + ISLES 2018
A Generative Probabilistic Model and Discriminative Extensions for Brain Lesion Segmentation— With Application to Tumor and Stroke	B.H. Menze et al.	77	2015	/	BTS/SS	/	/	78, 78 (BraTS, Zurich)	/	BraTS 2012 + Zurich Stroke Data
Brain SegNet: 3D local refinement network for brain lesion segmentation	X. Hu et al.	20	2020	3D	BTS/SS	73.7 (BraTS)	78.7 (specificity), 35 (BraTS, ISLES)	74, 30 (BraTS, ISLES)	/	BraTS 2015 + ISLES 2017
Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder	H.E. Atlason et al.	24	2019	3D	Lesion	/	75.7	76.6	/	AGES-Reykjavik Study
Convolutional Neural Network for Automated FLAIR Lesion Segmentation on Clinical Brain MR Imaging	M.T. Duong et al.	42	2019	3D	Lesion	76.7	76.9	78.9	/	Patients of Hospital of the University of Pennsylvania
Deep Active Lesion Segmentation	A. Hatamizadeh et al.	27	2019	2D	BTS	/	/	88.8 (Brain MR)	2.322 (Brain MR)	Multiorgan Lesion Segmentation (MLS)
MTANS: Multi-Scale Mean Teacher Combined Adversarial Network with Shape-Aware Embedding for Semi-Supervised Brain Lesion Segmentation	G. Chen et al.	0	2021	3D	BTS/SS	68.10, 75.4 (10% ISLES, 10% BraTS)	72.52 (10% ISLES)	61.66, 71.94 (10% ISLES, 10% BraTS)	37.39, 13.93 (10% ISLE	ISLES 2015 + BraTS 2018
Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT	V. Andrearczyk et al.	40	2020	/	Head and Neck Tumor	/	83.32	75.91	/	Data from 5 centers with H&N cancer
Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images	V. Andrearczyk et al.	29	2021	/	Head and Neck Tumor	/	/	77.85	3.08	Data from 6 centers with H&N cancer

Table 2.1a: Title, authors, number of citations, publication year, type of input images (2D/3D), aim and sensitivity, precision, Dice score, HD95 computed in the application to specific datasets, of articles studied during state-of-the-art analysis in brain tumor segmentation field.

Title	Method	Network	Loss Function	Optimization	Data augmentation	Normalization
Extending nn-UNet for brain tumor segmentation	MRI	nn-UNet modified	cross-entropy + bath Dice Loss	SGD with Nesterov momentum of 0.99	Random rotation and scaling, elastic deformation, additive brightness augmentation, gamma scaling	Group Normalization
Redundancy Reduction in Semantic Segmentation of 3D Brain Tumor MRIs	MRI	SegResNet from MONAI	Barlow Twins Loss + Dice Loss	Adam	Random axis mirror flip (for all 3 axes) with a probability 0.5	Zero mean and unit std for non-zero voxels
nnU-Net for Brain Tumor Segmentation	MRI	nn-UNet	cross-entropy + Dice Loss	SGD with Nesterov momentum of 0.99	Rotation, Scaling, Gaussian Noise and Blur, Brightness, Contrast, Low Resolution Simulation, Gamma augmentation, Mirroring	Instance Normalization
Generalized Wasserstein Dice Loss, Test-time Augmentation, and Transformers for the BraTS 2021 challenge	MRI	ensemble of 7 3D U-Nets with vision transformer	cross-entropy + generalized Wasserstein Dice loss	SGD, SGDP, ADAM, ASAM	Random zoom, rotation, Gaussian Noise, Gaussian Spatial Smoothing, Gamma augmentation, right/left flip	/
Optimized U-Net for Brain Tumor Segmentation	MRI	Unet modified	binary cross-entropy + Dice Loss for each region	Adam	Biased crop, Zoom, Flips, Gaussian Noise, Gaussian Blur, Brigthness, Contrast	Subtracting the mean and dividing by std of non-zero voxels
HNF-Netv2 for Brain Tumor Segmentation using multi-modal MR Imaging	MRI	HNF-Netv2	binary cross-entropy + generalized Dice Loss	Adam	/	Zero mean and unit std for non-zero voxels
Swin UNETR:Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images	MRI	Swin UNETR	Dice Loss	Not specified	Random axis mirror flip for all 3 axis, random intensity shift in range (-0.1,0.1) and scale in range (0.9, 1.1)	Zero mean and unit std for non-zero voxels
DeepMedic for Brain Tumor Segmentation	MRI	Deep Medic	not mentioned	Not mentioned	Reflection with respect to the mid-sagittal plane	Subtracting the mean and dividing by std of non-zero voxels
Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation	MRI	Deep Medic + CRF	CRF used for optimization	Stochastic gradient descent	Saggital reflections	Zero mean and unit std for non-zero voxels
A deep learning model integrating FCNNs and CRFs for brain tumor segmentation	MRI	FCNNs + CRF-RNN	CRF used for optimization	Not mentioned	/	N4ITK bias correction to T1 and T1c, subtracting image mode and normalizing std to be 1
Brain Tumor Segmentation with Deep Neural Networks	MRI	CNNs (different architectures, best: InputCascadeCNN)	Cross-entropy	Gradient descent with momentum	Flipping input images	Removing 1% highest and lowest intensities, N4ITK bias correction to T1 and T1c, subtracting the mean and dividing by std
Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images	MRI	CNNs (HGG and LGG) with small kernels	Categorical cross-entropy	SGD with Nesterov momentum	Rotating operations (90°)	N4ITK bias correction, intensity normalization method proposed by Nyúl et al., normalize to have zero mean and unit std
Brain Tumor Segmentation Using Multi-Cascaded Convolutional Neural Networks and Conditional Random Field	MRI	Multi-Cascade CNN (MCCNN) + CRF	CRF used for optimization	Stochastic gradient descent	/	Removing 1% highest and lowest intensities, N4ITK bias correction to T1c, subtracting the mean and dividing by std
Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM	MRI	Barkeley Wavelet Transform (BWT) + SVM	/	/	/	/
Deep learning based enhanced tumor segmentation approach for MR brain images	MRI	Random Forest + Growing CNN (GCNN)	/	/	/	Selected and estimated background (SEB) method
3D MRI brain tumor segmentation using autoencoder regularization	MRI	Encoder-Decoder (ResNet blocks) based CNN	Dice Loss + L2 Loss + KL Loss (VAE penalty term)	Adam	Random intensity and scale shift, axis mirror flip with probability 0.5	Zero mean and unit std for non-zero voxels
Automatic Brain Tumor Segmentation using Cascaded Anisotropic Convolutional Neural Networks	MRI	Cascaded Anisotropic CNNs (Wnet, Tnet, Enet)	Dice Loss	Adam	/	Subtracting the mean and dividing by std of non-zero voxels
Dual-force convolutional neural networks for accurate brain tumor segmentation	MRI	Dual-Force Training of MLDeepMedic + U-Net	(New) Label distribution-based loss function	Stochastic gradient descent	/	Zero mean and unit std for non-zero voxels
3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation	Microscope	3D U-Net	Weighted cross-entropy Loss	Stochastic gradient descent	Rotation, scaling and gray value augmentation, smooth dense deformation field	/
TransBTS: Multimodal Brain Tumor Segmentation Using Transformer	MRI	TransBTS (encoder (3D CNN) -Transformer - decoder)	Softmax Dice Loss	Adam	Random cropping, mirror flipping across saggital, coronal and axial planes, random intensity and scale shift	/
No New-Net	MRI	3D U-Net (small modifications)	Multiclass Dice Loss + negative log-likelihood	Adam	Random rotations, scaling, elastic deformations, gamma correction augmentation and mirroring	Subtracting the mean and dividing by std of non-zero voxels
UNet++: A Nested U-Net Architecture for Medical Image Segmentation	Microscopy/CT	Unet++	Binary cross-entropy + Dice Loss	Adam	/	/
Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation	MRI	Ensembles of Multiple Models and Architectures (EMMA)	Cross-entropy/IoU/Dice Loss	Adam	/	(1) Z-score, (2) Bias field correction followed by 1), (3) 2) followed by piece-wise linear normalization, followed by 1)
Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge	MRI	derived from Unet (modified)	Multiclass Dice Loss	Adam	Random rotations, scaling, elastic deformations, gamma correction augmentation and mirroring.	Subtracting the mean and dividing by std of non-zero voxels,clip the resulting images at [−5, 5] and rescale to [0, 1]
A Fully Automated Deep Learning Network for Brain Tumor Segmentation	MRI	ensemble of 3 3D-Dense-Unets	/	Adam	Horizontal flipping, vertical flipping, random rotations and translational rotations	Zero mean and unit std for non-zero voxels
Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images	MRI	Unet + transfer learning using a convolution-base of Vgg16 for classification	Negative value of Dice similarity coefficient (DSC)	Adam	/	Tissue regions pixel values are rescaled to be between [-1,1], non tissue regions are kept at -1
Brain tumor segmentation with deep convolutional symmetric neural network	MRI	Deep Symmetric Convolutional Neural Network (DSCNN)	Focal Loss	Minibatch SGD and Adam with momentum 0.9	/	Pixel values scaled to be between [0-1]
ERV-Net: An efficient 3D residual neural network for brain tumor segmentation	MRI	ERV-Net (encoder: ShufleNetV2, decoder: residual blocks)	Cross-entropy Loss + Dice Loss	Adam	Gamma correction, random rotations, Gaussian noise, elastic and scaling deformations, mirror and brightness transformation	Zero mean and unit std for non-zero voxels
RescueNet: An unpaired GAN for brain tumor segmentation	MRI	RescueWNet, RescueENet, RescueCNet	Aversial Loss + Cycle consistency Loss (in GANs)	Not mentioned	Unpaired training procedure	/
A novel end-to-end brain tumor segmentation method using improved fully convolutional networks	MRI	Cascaded of improved UNet models	Dice Loss	Adam	/	Zero mean and unit std for non-zero voxels
Cross-Modality Deep Feature Learning for Brain Tumor Segmentation	MRI	Cross-modality Feature Transition (CMFT) + Cross-modality Feature Fusion (CMFF)	Adversial Loss + Cycle consistency Loss	Adam	/	Zero mean and unit std for non-zero voxels
One-pass Multi-task Networks with Cross-task Guided Attention for Brain Tumor Segmentation	MRI	One-pass Multi-tak Network (OM-Net) con Cross-talk Guided Attention (CGA)	SoftMaxWithLoss	SGD with momentum of 0.99	/	Zero mean and unit std for non-zero voxels
Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images	MRI	Cascade CNN (C-CNN)	Cross-entropy Loss	Adam	/	Zero mean and unit std for non-zero voxels
Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy	MRI	SparseMultiOCM + InputSparseMultiOCM + DenseMultiOCM	Cross-entropy Loss with weights for the 4 classes	Minibatch SGD	Use of overlapping patches	Removing 1% lowest and highest intensities, zero mean and unit std for non-zero voxels, put background pixels to -9
Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation	MRI	Cascade of CNNs (Wnet, Tnet, Enet)	Dice Loss	Adam	Random rotation, flipping and scaling, intensity noise extracted using MonteCarlo simulations	Zero mean and unit std for non-zero voxels
Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task	MRI	Two Stage Cascaded Unet	Soft Dice Loss	Adam	Random intensity shift to [-0.1, 0.1] of std of each channel, intensity scaling between [0.9, 1.1], random crops and flips	Subtracting the mean and dividing by std of non-zero voxels
Bag of Tricks for 3D MRI Brain Tumor Segmentation	MRI	Self-ensemble Unet	Cross-entropy Loss + Dice Loss	SGD with momentum	Random axis mirror along the horizontal axis	Pixel values scaled to be between [0-1]
Automatic Semantic Segmentation of Brain Gliomas from MRI Images Using a Deep Cascaded Neural Network	MRI	Tumor Localization Network (TLN) + Intratumor Classification Network (ITCN)	Categorical Cross-entropy Loss	Minibatch SGD	/	Removing 1% lowest and highest intensities, zero mean and unit std for non-zero voxels
H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes	MRI	H-Dense Unet	Weighted cross-entropy Loss	SGD with momentum	Random mirror and scaling between 0.8 and 1.2	Truncate the image intensity values to the range of [-200,250] to remove the irrelevant details
Convolutional neural network with batch normalization for glioma and stroke lesion detection using MRI	MRI	CNN model	not specified	Not mentioned	/	Zero center normalization
A New Approach for Brain Tumor Segmentation and Classification Based on Score Level Fusion Using Transfer Learning	MRI	Threshold + Morphological Opening (Segmentation), AlexNet +GoogleNet (Classification)	not specified	/	/	Input images resized with 256
A Generative Probabilistic Model and Discriminative Extensions for Brain Lesion Segmentation— With Application to Tumor and Stroke	MRI	Generative Probabilistic model + Discriminative Probabilistic model	not specified	/	Augment the dataset with random samples from a Gaussian distribution with mean 0 and standard deviation of 1	Image intensities are scaled linearly
Brain SegNet: 3D local refinement network for brain lesion segmentation	MRI	Brain SegNet (3D refinement module)	Focal Loss	Adam	Random amplification of voxel intensities, rotations of slices, slices are rescaled, horizontal and vertical flips, random cropping	/
Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder	MRI	Segmentation Auto-Encoder (SegAE)	L = (Y p − Y` p) ^2	Adam	/	/
Convolutional Neural Network for Automated FLAIR Lesion Segmentation on Clinical Brain MR Imaging	MRI	Fine-tuned 3D Unet	Cross-entropy Loss	Adam	Random rotations, translations, scaling, and free-form deformations	Zero mean and unit std for non-zero voxels
Deep Active Lesion Segmentation	MRI/CT	Deep Active Lesion Segmentation (DALS)	Dice Loss	Adam	/	Zero mean and unit std for non-zero voxels
MTANS: Multi-Scale Mean Teacher Combined Adversarial Network with Shape-Aware Embedding for Semi-Supervised Brain Lesion Segmentation	MRI	Multi-Scale Mean Teacher Combined Adversarial Network (MTANS)	Segmentation + Consistency Loss (teacher) + multi-scale loss (discriminator)	AMSGrad	/	Zero mean and unit std for non-zero voxels
Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT	FDG-PET/CT	winner Unet with Squeeze and Excitation Normalization	Dice Loss + Focal Loss	Not mentioned	/	/
Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images	FDG-PET/CT	winner 3D nn-Unet with Squeeze and Excitation Normalization	Dice Loss + Focal Loss	Not mentioned	Rotation, scaling, mirroring, Gaussian noise and Gamma correction	Zero mean and unit std for non-zero voxels

Table 2.1b: Title, acquisition method, network name, loss function, optimization method, data augmentation and normalization pipelines performed, for the same articles as Table 1a.

2.2 Stroke Lesion Segmentation

Articles concerning stroke lesion segmentation were sought in literature at first by identifying winners and best performing methods in ISLES 2018, 2017 and 2016 challenges; then by researching in Google Scholar using keywords like “Stroke Lesion Segmentation”, “Automatic Stroke Lesion Segmentation” and “Stroke Lesion Segmentation Deep Learning”. Articles were then identified because cited by other papers, or analyzing reviews; in this scope, the only review studied was “Application of Deep Learning Method on Ischemic Stroke Lesion Segmentation” (Y. Zhang et al., 2022).

As for brain tumor segmentation, also in this case the cut-off year for articles selection was chosen to be 2015, in which, as stated before, UNet was first introduced, basically downgrading the majority of other methods used in medical images segmentation.

A brief description is provided for the most interesting methods, while details about architectures and respective performances can be visualized in *Table 2*. For my specific objective, investigation was mainly (but not only) focused on papers dealing with ischemic stroke.

To deal with the lesion variability between subacute and chronic stroke phases, and hemorrhagic and ischemic stroke lesions, Y. Xue et al. introduced a system defined as 2.5D, because composed by nine end-to-end UNets, each taking as input a 2D slice (sagittal, axial or coronal), along with three different normalizations. A 3D convolutional kernel is then used to merge 2D outputs from each path, showing that it gives a better accuracy than majority voting (Xue et al., 2020). Tao Sang proposed a 3D multi-scale U-shape network with ‘atrous convolution’, which allows to enlarge the field of view of filters to include a larger context without increasing the number of parameters (L.-C. Chen et al., 2016). This network also uses skip connections to combine low-level and high-level feature maps, and a multi-scale loss function, given by the combination of Jaccard index and binary cross entropy, to constraint the back-propagation of the network. It ranked 1st in ISLES 2018 competition (Song, n.d.). X. Hu et al. developed StrokeNet, a 3D segmentation network with an incorporated 3D refinement module, which aggregates rich fine-scale spatio-temporal features and allows to explore local and high-level context information. They also introduced a new training strategy that incorporates curriculum learning and Focal loss, allowing the network to naturally deal with data imbalance (X. Hu et al., n.d.). S. Wang et al. designed Consistent Perception GAN (CPGAN), composed by three networks: a segmentation network, with a UNet architecture provided with a similarity connection module (SCM) to capture long-range contextual

information, and fed with original and rotated image; a discriminator network that, with the help of an assistant network, learns to distinguish rotated images and so learns meaningful feature representations often lost during training stage (S. Wang et al., 2020). J. Donahue et al. developed Bidirectional GAN (BiGAN) for discrimination tasks: it consists basically in a GAN with the addiction of an inverse mapping, an encoder which maps data to their latent representations (opposite to the conventional generator), and the discriminator discriminates jointly in data and latent space. In this way, BiGAN encoder can learn useful feature representations (Donahue et al., 2016). To apply this structure in the medical segmentation field, C. Baur et al. modified BiGAN in AnoVAEGAN, where the deep generative model has a form of an encoder-decoder network, in which the encoder is a spatial Variational AutoEncoder (VAE), able to capture the "global" normal anatomical appearance, reconstructing the healthy tissues but avoiding anomalies, while the decoder is trained with the help of an adversarial network. Lesions are then segmented computing the distance between input and reconstructed images, and it was applied for multiple sclerosis lesions segmentation (Baur et al., 2018). J. Dolz et al. proposed a multi-path architecture, in which multiple modalities are processed in different paths to exploit their unique information, and layers of the same path and different paths are densely-connected. Single architectures are UNets, with two additional dilated convolutional blocks to learn larger context information (Dolz, Ayed, et al., 2018). Y. Zhou et al. introduced D-UNet, in which the encoder performs 3D and 2D feature extraction on a small number of consecutive slices and then those features are combined to achieve a small number of parameters and less computation time in comparison of 3D networks, while obtaining a better segmentation performance than 2D networks. They also proposed a new Enhance Mixing Loss (EML), which combines Dice loss and Focal loss and adds a weighted focal coefficient, that allows a faster convergence (Y. Zhou et al., 2019). To include unannotated data into the training of CNNs, W. Cui et al. designed a semi-supervised learning approach, adapting the mean teacher model (Tarvainen & Valpola, n.d.), developed for image classification. They built a teacher and a student model, sharing the same DeepMedic architecture, following a self-ensembling framework for training: the student model is updated at each step minimizing two losses computed using respectively annotated and unannotated data, while the teacher model is updated combining the current student model and the historical information of teacher models (Cui et al., 2019). R. Guerrero et al. implemented uResNet, a network able to segment and differentiate between white matter hyperintensities (WMH) and stroke lesions. WMHs are a characteristic of small vessel diseases, and the accurate assessment of their burden can be crucial for diagnosis and treatment of WMHs, and to determine their

associations with cognitive and clinical data; however, stroke lesions often appear hyperintense and can be confused with WMHs. uResNet is a U-shaped architecture characterized by an analysis path to capture context, and a symmetric synthesis path, to enable precise localization (Guerrero et al., 2018). For accurate acute and sub-acute ischemic lesion segmentation, A. Clèrigues et al. proposed a 3D asymmetric encoder-decoder network (75% of parameters in the encoder) based on the UNet architecture, with global and local residual connections. Class imbalance is addressed using small patches and a weighted loss function (Clèrigues et al., 2018).

For highly unbalanced segmentations (like ischemic stroke lesions), classic regional losses (Dice or cross-entropy losses) have limitations, because they assume identical importance for all classes and samples; for these reasons, H. Kervadec et al. devised a boundary loss, which uses integrals over the interface between regions, instead of over the regions. It is implemented as the sum of linear functions of the regional softmax probability outputs of the network, and can be combined with standard regional losses, improving their performances (Kervadec et al., 2019). N. Tomita et al. proposed a novel training strategy, called two-stage zoom-in&out strategy, based on training first using small volumes (this stage has a regularization effect and is computational inexpensive), and then finetuning the models on larger volumes. They applied this strategy on a modified UNet, obtained replacing each convolutional layer with a residual block and substituting batch normalization with group normalization (Tomita et al., 2020). To solve the problems of insufficient training data and high computational cost of 3D CNNs, H. Hui et al. introduced a partitioning-stacking prediction fusion (PSPF) method, which consists in three steps: first, slices are partitioned based on the acquisition plane, in subsets according to the similarity of brain's anatomical structures, and each subset is used to perform training and prediction separately (partitioning); then, the 2D slice results are stacked to form a 3D lesion map (stacking); and finally the three orthogonal planes 3D results are fused using soft voting (fusion). This method is applied to a UNet with an attention gate (AG), used to highlight salient features for a specific task (Hui et al., 2020). Finally, R. Zhang et al. proposed a 3D fully convolutional and densely connected convolutional network (3D FC-DenseNet) for segmentation of ischemic stroke lesions from DWI scans, where there exist many artifacts that mimic the intensity and shapes of stroke lesions. Basically, FC-DenseNet is obtained by extending DenseNets to 3D, by exploiting dense connectivity to boost information and gradients flow in the network (R. Zhang et al., 2018).

All details about architectures and performances of the cited methods, and many more, can be visualized in *Table 2.2.*

Title	Authors	Citations	Publication	Aim	2D/3D	Sensitivity	Precision (PPV)	Dice score	HD 95	Datasets	Method	Network	Loss Function	Opt-imization	Data augmentation	Normalization
A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images	Y. Xue et al.	121	2020	SS	2.5D	/	/	66 (ATLAS), 47 (FAMCW), 62 (CV)	/	MICCAI + ATLAS (train and test on different ways)	MRI	9 Unets + a 3D con-volutional kernel to combine them	Dice Loss	SGD with Nesterov momentum of 0.9	Flipped version as augmented synthetic modality	Three different normalizations for each of the axial, sagittal, and coronal planes
Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier	A. Subudhi et al.	74	2020	/	SS	92.5	95.8 (specificity)	92.2	851 (Jaccard)	MRI Dataset of DWI sequences	MRI	Expectation-Maximization (EM) algorithm + FODSPO approach	Log-likelihood function	/	/	/
3D Multi-scale U-Net with Atrous Convolution for Ischemic Stroke Lesion Segmentation	Tao Song	4	2018	3D	SS	/	/	55.86 (val. set)	/	ISLES 2018 (winner)	MRI	3D multi-scale network with atrous convolution and skip connections	Jaccard index (IoU) + binary cross-entropy	/	Random rotations, random crop and radial distortion	/
Ensembles of Modalities-Fused Model for Ischemic Stroke Lesion Segmentation	Y. Chen et al.	3	2018	2.5D	SS	/	/	/	/	ISLES 2018 (second place)	MRI	Ensemble of multiple backbone networks (e.g. Unet)	/	/	Several data augmentation approaches	/
2D Multi-Scale Res-Net for Stroke Segmentation	C. Lucas et al.	2	2017	2D	SS	/	/	/	/	ISLES 2017 (second place)	MRI	Fully convolutional architecture based on U-nets with residual connections	Dice Loss	Adam	/	/
StrokeNet: 3D Local Refinement Network for Ischemic Stroke Lesion Segmentation	X. Hu et al.	4	2018	3D	SS	/	58.84 (train. set)	52.26 (train. set)	/	ISLES 2018	MRI	StrokeNet	Focal Loss	/	/	/
Ischemic Stroke Lesion Segmentation with Convolutional Neural Networks for Small Data	Y. Choi et al.	3	2017	3D	SS	/	/	/	/	ISLES 2017 (winner)	MRI	Ensemble of CNNs (residual U-net and spatial pyramid pool-ing)	Generalized Dice overlap	/	/	/
Automated Segmentation of Chronic Stroke Lesions Using LINDA: Lesion Identification With Neighborhood Data Analysis	D. Pustina et al.	112	2016	/	SS	65.6 (val. set)	77 (val. set)	66.8 (test set)	24	Train/val. on a Dataset of 60 left-hemispheric stroke patients + test on separate dataset	MRI	Lesion Identification with Neighborhood Data Analysis (LINDA)	/	/	/	Images were corrected for signal inhomogeneity with the N4 algorithm
Brain Stroke Lesion Segmentation: Using Consistent Perception Generative Adversarial Network	S. Wang et al.	1	2022	2D	SS	55.6	70.15 (specificity)	61.7	58.11 (Jaccard)	Anatomical Tracings of Lesions After Stroke (ATLAS)	MRI	Consistent Perceptive Generative Adversarial Network (CPGAN)	Rotation Loss + Dice loss	Adam	Labelled and unlabelled images are rotated	Zero mean and unit std for non-zero voxels
Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images	C. Baur et al.	295	2018	2D	MISS	/	/	60.51	/	Dataset of healthy subjects and subjects with MS lesions	MRI	AnoVAEGAN	Reconstruction Loss + KL-divergence + Adversial Loss	Adam	/	Normalized into the range [0,1]
Adversarial Feature Learning	J. Donahue et al.	1.534	2016	2D	/	Method applied only for discrimination tasks, then adapted to segmentation in "Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images" (C. Baur et al.)				MNIST and ImageNet	MRI	Bidirectional generative Adversial Network (BiGAN)	10 autoencoder loss function	Adam	A 64 x 64 crop is randomly selected from image, the / crop is flipped horizontally with probability 0.5 and scaled to [-1,1]	
Dense Multi-path U-Net for Ischemic Stroke Lesion Segmentation in Multiple Image Modalities	J. Dolz et al.	68	2018	2D	SS	/	/	63.5	16.64	ISLES 2018	MRI	U-net architecture with N paths (for each modality)	not mentioned	Adam	/	Normalized into the range [0,1]
D-UNet: a dimension-fusion U shape network for chronic stroke lesion segmentation	Y. Zhou et al.	65	2019	2D/3D	SS	/	63.31	72.31	/	ATLAS	MRI	Dimension-fusion-U-net (D-Unet)	Enhance Mixing Loss (EML)	SGD	Translation, scaling and horizontal flipping	Scaling the input mean to 0
Ischemic stroke lesion segmentation using stacked sparse autoencoder	G.B. Praveen et al.	42	2018	/	SS	88.3 (specificity)	96.8	94.3	/	ISLES 2015	MRI	Unsupervised method based on Stacked Sparse Autoencoder (SSAE)	Mean Square Error (MSE) cost Function	Gradient descent	/	Zero mean and unit std for non-zero voxels
Segmentation of Ischemic Stroke Lesion in Brain MRI Based on Social Group Optimization and Fuzzy-Tsallis Entropy	V. Rajinikanth et al.	104	2018	/	SS	99.98	98.48	84.03	72.27 (Jaccard)	ISLES 2015 + hemorrhagic cerebral infarction database (Radiopaedia)	CT/MRI	SGO + FTE (thresholding) + GAC (CV/SRG/MV/DRLS)	/	Gradient descent	/	/
Semi-Supervised Brain Lesion Segmentation with an Adapted Mean Teacher Model	W. Cui et al.	74	2019	3D	SS	/	/	66.76	/	Ischemic Stroke Lesion Segmentation Dataset	MRI	Teacher-Student models with same architecture (DeepMedic)	Segmentation loss (unannotated data) + Segmentation Consistency Loss (annotated data)	RMSprop	Additive noise (from Gaussian distribution with 0 mean and 0.05 std) and multiplicative noise (mean of voxels 1 and std of 0.01)	Subtracting the mean and dividing by std of brain
The ENIGMA Stroke Recovery Working Group: Big data neuroimaging to study brain-behavior relationships after stroke	S.L. Liew et al.	120	2022	/	SS	Analysis of the best performing models on ATLAS, LINDA is identified as the best		ATLAS (developed by ENIGMA Stroke Recovery Group)		ATLAS	MRI	LINDA best performing	/	/	/	/
White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks	R. Guerrero et al.	148	2018	2D	SS	/	/	69.5, 40 (WMH, Stroke/ Lesion)	157 patients who presented to a hospital stroke service	ATLAS	MRI	uNetXmodules (analysis + synthesis path)	Classical Cross-entropy	Adam	Random shift Axxy of up to half the patch size applied in the axial plane before patch extraction	Zero mean and unit std for non-zero voxels
X-Net: Brain Stroke Lesion Segmentation Based on Depthwise Separable Convolution and Long-range Dependencies	K. Qi et al.	75	2019	2D	SS	/	60	48.67	/	ATLAS	MRI	X-Net (with DSC and FSM)	Dice Loss + Cross-entropy Loss	Adam	/	Standard normalization
Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI	L. Liu et al.	141	2020	2D	SS	/	/	74.20	2.33	SPES (acute stroke penumbra) Dataset of ISLES 2015 + Local Hospital Clinical Dataset (LHC)	MRI	Rec-CNN	Novel Loss Function (1TLV-P)	Not specified	Rotation of 90°, 180°, 270°, flip, mirror mapping, exclusion of redundant images	Intensity range standardization
Acute and sub-acute stroke lesion segmentation from multimodal MRI	A. Clèrigues et al.	123	2020	3D	SS	60, 89 (ISLES test, SPES test)	65, 82 (ISLES test, SPES test)	59, 84 (ISLES test, SPES test)	34.7, 20.7 (ISLES test, SPES test)	SPES and ISLES (sub-acute ischemic stroke) Datasets of ISLES 2015	MRI	cGeo2 (3D asymmetric encoder-decoder network based on Unet)	Focal Loss	Adadelta	Flip along the mid-sagittal axis, apply BSpline to register to original image	Normalized into the range [0,1]
Adaptive feature recombination and recalibration for semantic segmentation with Fully Convolutional Networks	S. Pereira et al.	138	2019	2D	87.5/SS	78.3, 55 (Bra-TS test, ISLES)	81.2, 26 (Bra-TS test, ISLES)	77.37, 24 (Bra-TS test, ISLES)	8.37 (Bra-TS test)	Bra-TS 2017 + SPES of ISLES 2015	MRI	2D FCN with feature recalibration (SegSE-UNet) and recombination	Dice Loss	Adam	Random sagittal flipping and random rotations of (0°,5°), (80°, 270°	Standardize the intensity histogram of each modality
Automatic Ischemic Stroke Lesion Segmentation from Computed Tomography Perfusion Images by Image Synthesis and Attention-Based Deep Neural Networks	G. Wang et al.	120	2020	2D	SS	/	SS (test set)	51 (test set)	21.26 (val. set)	ISLES 2018	CTP	Feature extraction (DWI) + DWI generator (Unet) + SUNet segmentation	Weighted Cross-entropy Loss + Hardness-aware Generalized Dice Loss	RMSprop	/	Normalized into the range [0,1]
Boundary loss for highly unbalanced segmentation	H. Kervadec et al.	196	2019	2D	SS	/	/	65.6, 74.8 (ISLES, WMH)	3.56/2.0/987 (ISLES, WMH)	ISLES 2017 + WMH (MICCAI 2017)	MRI	UNet	Boundary Loss + Generalized Dice Loss	Adam	/	Normalized into the range [0,1]
CSNet: A new DeepNet framework for ischemic stroke lesion segmentation	A. Kumar et al.	12	2020	2D	SS	/	61.41, 90.84 (2017 val., 2015 val.)	60.88, 89.9 (2017 val., 2015 val.)	/	ISLES SPES 2015 + ISLES 2017	MRI	Classification + Classifier-Segmenter Network (CSNet) (segmentation)	Binary Cross Entropy (BCE) Dice Loss	Adam	Different data augmentation techniques, not specified	Subtracting the mean and dividing by std of brain voxels
Attention convolutional neural network for accurate segmentation and quantification of lesions in ischemic stroke disease	L. Liu et al.	28	2020	2D	SS	/	/	76.39, 72.83 (Stroke, WMH)	3.19, 2.63 (Stroke, WMH)	ISSSS Dataset of ISLES 2015	MRI	Deep Residual Attention CNN (DRAANet)	Loss Function based on Dice Coefficient	Adam	Augmented but not specified	/
A deep supervised approach for ischemic lesion segmentation from multimodal MRI using Fully Convolutional Network	R. Karthik et al.	311	2019	2D	SS	/	/	70	/	ISLES 2015	MRI	Deep FCN (Unet architecture modified)	Dice Loss	Adam	Rotation, shifting, shearing, elastic distortion, zooming, flipping	Subtracting the mean and dividing by std of brain voxels
Automatic post-stroke lesion segmentation on MR images using 3D residual convolutional neural network	N. Tomita et al.	116	2020	3D	SS	/	62	64	20.4	ATLAS	MRI	3D-ResUNet (modified 3D Unet)	Binary Cross Entropy + Dice Loss	Adam optimizer and cosine annealing	Random crops	Standard normalization
A Partitioning-Stacking Prediction Fusion Network Based on an Improved Attention U-Net for Stroke Lesion Segmentation	H. Hu et al.	5	2020	2D/3D	87.5/SS	/	69.1, 88.7 (ATLAS, Bra-TS)	59.3, 87.7 (ATLAS, Bra-TS)	29.13, 38 (DWI&ADC, Bra-TS)	ATLAS + Bra-TS 2017	MRI	Partitioning-stacking prediction fusion (PSPF) method with improved attention Unet	Dice Loss	Adam	/	Intensity standardization normalization
Automatic Segmentation of Acute Ischemic Stroke From DWI Using 3-D Fully Convolutional DenseNets	R. Zhang et al.	120	2018	3D	SS	78.15 (DWI&ADC)	92.67, 60 (DWI&ADC, SSSI)	79.13, 58 (DWI&ADC, SSSI)	38.98 (SSSI)	ISLES 2015 + SSSI Dataset	DWI/MRI	3D FC-DenseNet	Cross Entropy Loss + Dice Loss (4 companion losses + 1 main loss)	SGD	Random crops and flips along the 3 axis	Normalized into the range [0,1]
C.C.-Net: Cross-Level Fusion and Context Inference Networks for Lesion Segmentation of Chronic Stroke	H. Yang et al.	36	2019	2D	SS	/	64.9	58.1	/	ATLAS	MRI	Cross-Level Fusion and Context Inference Network (CLCI-Net)	Dice loss	Adam	/	/
Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks	L. Chen et al.	189	2017	2D	SS	/	/	83	/	ISLES train + ISLES test	DWI	2 CNNS: FCN-Net (ensemble of 2 DeconvNets) + MU-Net-A	Weighted Cross-entropy Loss	SGD	Horizontal flips and random rotations, random patch extraction	Zero mean and unit std for non-zero voxels

Table 2.2: Title, authors, number of citations, publication year, type of input images (2D/3D), aim, sensitivity, precision, Dice score, HD95 computed in the application to specific datasets, acquisition method, network name, loss function, optimization method, data augmentation and normalization pipelines performed, for articles studied during state-of-the-art analysis in stroke lesion segmentation field

2.3 Relevant architectures: nnUNet ("no-new-UNet")

Among all the analyzed methods, nnUNet can now be considered the state-of-the-art model for brain tumors segmentation, and for this reason it was chosen for further examinations and tests within this thesis. It was first introduced by Isensee et al. in 2018 and it was applied, with limited modifications, including adaptation of postprocessing, region-based training, a more aggressive data augmentation, and small adjustments of nnUNet pipeline, to BraTS 2020 challenge, where it took the first place an won the competition. An extended version of nnUNet was also submitted to BraTS 2021 challenge by H.M. Luu et al., who proposed to use a larger encoder, to substitute batch normalization with group normalization, and introduced an axial attention in the decoder; this method won the competition too, ranking 1st.

The unique feature of nnUNet is that it automatically adapts to the specific dataset used for training, and so to the specific task, without needing to modify either the architecture, preprocessing or training processes adopted. It automatically configures segmentation pipelines for arbitrary biomedical datasets, without requiring expertise or extensive experimentation.

Moreover, the authors also stated that small modifications of the architecture are not superior to a properly tuned model, and often do not provide significant results. They demonstrated that some of the recently presented architectural modifications, such as residual connections, dense connections or attention mechanisms, are in part overfitted to the specific problem, or in part could suffer from imperfect validation that results from sub-optimal reimplementations of the state-of-the-art (Isensee et al., 2018). For this reason, they underlined the importance of hyperparameters tuning, instead of modifications of the architecture, pre-processing, training, inference or post-processing, which quite often cause the U-Net to underperform when used as a benchmark. It can be also demonstrated that these small architectural tweaks, which are intended to improve the performance of the model, work well only if the network is not yet fully optimized: regarding this aspect, they analyzed Kidney Tumor Segmentation (KiTS) 2016 challenge, showing that all top 15 methods are based on U-Net, but none of the architectural modifications used are an essential condition for performance improvement.

Much of the networks' performances can be gained or lost due to modifications in: preprocessing (e.g. resampling and normalization), training (e.g. loss, optimizer setting and data augmentation), inference (e.g. patch-based strategy and ensembling across test-time augmentations and models) and a possible post-processing (Isensee et al., 2018).

nnUNet was successfully applied to 19 datasets and 49 segmentation tasks, achieving a new state-of-the-art in 29 of them.

2.3.1 General description

nnUNet automatically generates three simple U-Net architectures ("no-new-UNet") and the best configuration is chosen through cross-validation after the training phase. The three networks are: a 2D UNet, a 3D UNet and a UNet cascade. The 2D UNet is intuitively suboptimal for 3D medical images segmentation, because it isn't able to aggregate information along the z-axis; however, it is possible to demonstrate that, in presence of anisotropic datasets (like the Prostate dataset of the Decathlon challenge), conventional 3D methods deteriorate in performance, and 2D methods could overcome. The 3D UNet trained on full resolution images is, in the majority of cases, the best method, as it represents the appropriate model of choice for 3D image data. As last, the UNet cascade consists in a first stage, where a 3D UNet is trained on downsampled images, while in the second stage, the results of the first UNet are upsampled to the original voxel spacing, and passed as additional input channels to a second 3D UNet, which is trained on patches at full resolution to refine the first segmentation maps. This last solution is ideal with datasets with large image sizes, where the 3D UNet patch size could be too small.

To automatically adapt to a specific task and generate specific methods for previously unseen datasets, nnUNet is able to generate and follow systematic rules. In particular, the pipeline optimization problem is condensed in a set of heuristic rules that robustly generate a high-quality pipeline fingerprint (key design and architectural choices for the specific network) from the corresponding dataset fingerprint (key features of the dataset) while considering the computational (hardware) constraints (*Figure 2.1,* from Isensee et al., 2019).

As a first step, nnUNet crops the training data to their nonzero region. This phase has no effect in most datasets, but it highly reduces image size of brain datasets, improving computational efficiency (Isensee et al., 2019).

Secondly, based on the cropped training images, nnUNet generates a dataset fingerprint, which represents a standard description of the dataset, including relevant parameters and properties such as images sizes before and after cropping, modalities, images spacings (i.e., the physical size of the voxels), number of classes for all images, dimension of training set, as well as mean, standard deviation and 0.5 and 99.5 percentiles of the intensity values in the foreground regions (voxels belonging to class labels).

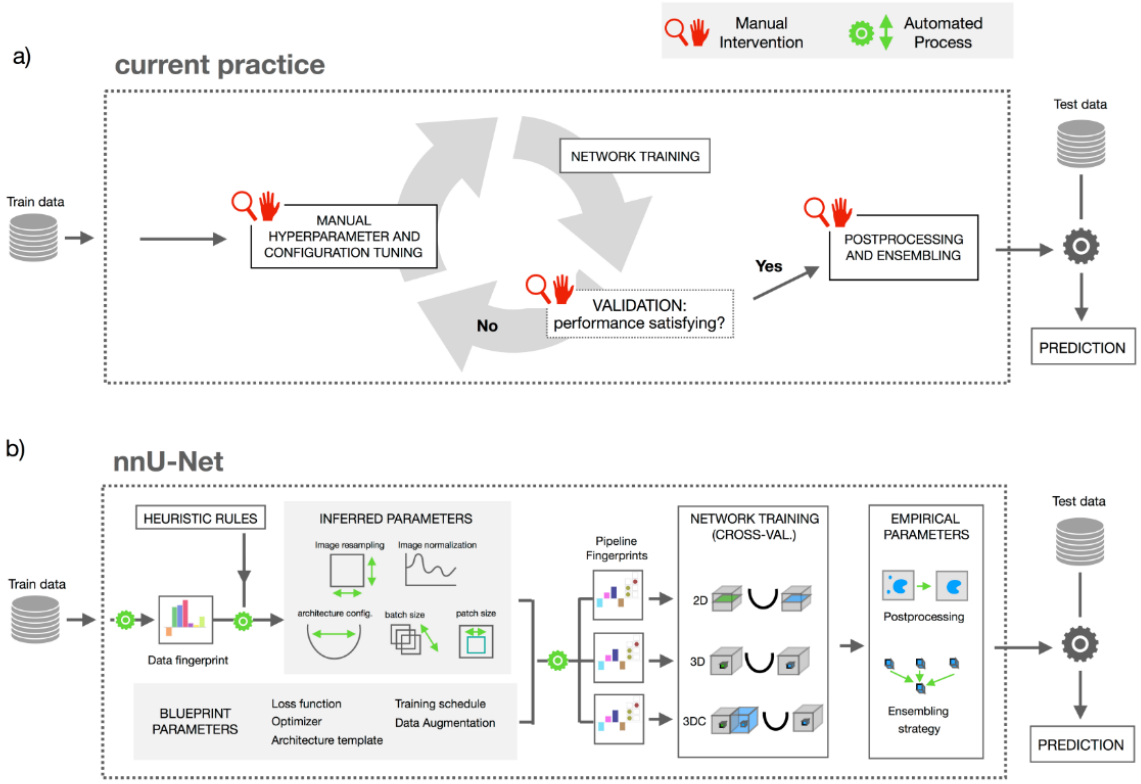


Figure 2.1: Comparison between current practice and nnUNet configuration identification. In current practice (a), the architecture is identified by an iteration between manual hyperparameters and configuration tuning and network training and validation. With nnUNet (b), starting from a dataset fingerprint, a pipeline fingerprint is generated (inferred parameters, blueprint parameters and empirical parameters) and three architectures are trained in a 5-fold cross validation. Finally, nnUNet automatically selects the optimal configuration.

After that, nnUNet automatically designs deep learning methods by generating a pipeline fingerprint, that contains all relevant architectural choices, and is composed by three types of parameters: inferred parameters, design choices inferred directly from the dataset fingerprint using a set of heuristic rules, considering the project-specific hardware constraints; blueprint parameters, data-independent; and empirical parameters, optimized during training.

2.3.1.1 Blueprint parameters

The blueprint parameters represent the baseline UNet template, from which all nnUNet configurations originate. This template follows the original UNet architecture and its 3D counterpart, with two blocks per resolution step in both encoder and decoder, in which each block consists in a convolution, followed by instance normalization and leaky ReLU; downsampling is then implemented as strided convolution, upsampling as convolution

transposed. The initial number of feature maps is 32, doubled (halved) at each downsampling (upsampling) operation. The original template has been slightly modified: it was exploited a small batch size (most of 3D configurations use a batch size of only 2), to enable for large patch sizes; batch normalization was substituted with instance normalization, because the former doesn't work well with small batch sizes. Furthermore, ReLU was replaced with Leaky ReLU, and the networks were trained with deep supervision: to ease training of all layers in the network and allow gradients to be injected deeper inside it, additional auxiliary losses are added in the decoder to all but the two lowest resolutions (Isensee et al., 2019). For what concerns the training schedule, the total number of epochs was set to 1000, where each epoch represents an iteration over 250 minibatches, and oversampling was implemented to handle class imbalance: training cases are chosen randomly and 66.7% of samples come from random locations inside the selected training case, while 33.3% surely contain one of the foreground classes. The learning rate follows the 'polyLR' learning rate decay, $(1 - \frac{epoch}{epoch_{max}})^{0.9}$, and the loss function was chosen to be the sum of cross-entropy and Dice loss. The optimization method chosen is Stochastic Gradient Descent (SGD) with Nesterov Momentum ($\mu = 0.99$).

Data augmentations techniques are applied directly on the fly during training; their parameters are randomly extracted from predefined ranges, which don't change across different datasets. The following augmentations are automatically applied by nnUNet:

1. *Rotation and Scaling*. They are applied together, with a probability of 0.2 each; if 3D patches are anisotropic (see Blueprint parameters) angles of rotation around the three axes (in degrees) are extracted from a uniform distribution between -30 and 30 (indicated as U(-30,30)). While if the patch is anisotropic or 2D, the angle of rotation is extracted from U(-180,180), and finally if the patch size is 2D and anisotropic, the angle is extracted from U(-15.15). Scaling is applied by multiplying coordinates for a scaling factor (<1 results in a zoom out, >1 in a zoom in) sampled from U(0.7,1.4);
2. *Gaussian Noise*. It is added to each voxel independently a zero centered Gaussian noise, with variance drawn from U(0,0.1). It's applied with a probability of 0.15;
3. *Gaussian Blur*. Applied with a probability of 0.2 for each sample. If is triggered in a sample, blurring is applied to each modality with a probability of 0.5, and the width (in voxels) of the Gaussian kernel used for blurring is extracted from U(0.5,1.5);
4. *Brightness*. Voxel intensities are multiplied for a value sampled from U(0.7,1.3) with probability 0.15;
5. *Contrast*. Voxel intensities are multiplied, as before, for a value extracted from U(0.65,1.3) with probability 0.15. After multiplication, values are clipped to their initial range;

6. *Simulation of low resolution*. Augmentation applied with a probability of 0.25 for sample and 0.5 for each associated modality. The considered modality is downsampled by a factor of U(1,2) using nearest neighbor interpolation, and then sampled back to its original dimensions using cubic interpolation;
7. *Gamma augmentation*. Patch intensities are scaled to a factor of [0,1] of their respective value range. Then, voxel intensities are raised to a value gamma drawn from U(0.7,1.5), and subsequently they are scaled back to their original range. With a probability of 0.15, voxel intensities are inverted before the augmentation, following the transformation $(1 - i_{new}) = (1 - i_{old})^{\gamma}$
8. *Mirroring*. All patches are mirrored with probability 0.5.

For the full resolution Unet of the Unet cascade, the following augmentations are applied to the segmentation map generated by the low resolution 3D Unet:

4.3 *Binary Operators.* A binary operator, randomly chosen from dilation, erosion, opening and closing, is applied to all labels in the predicted masks with probability 0.4. The structure element is a sphere with radius sampled from U(1,8);

4.4 *Removal of Connected Components.* Components smaller than 15% of the patch size are removed with probability 0.2.

To conclude, images are predicted with a sliding window approach, where the window size is equal to the patch size used during training, and test-time augmentation by mirroring is applied along all axes.

2.3.1.2 Inferred parameters

They represent modifications to the network topology directly derived from the dataset fingerprint and automatically computed by nnUNet. They basically regard:

- *Intensity normalization.* The default normalization scheme applied for all modalities except CT images is z-scoring, applied independently to each image during training and inference. CT images intensity values are quantitative and reflect physical properties of tissues; for this reason, during normalization this information is preserved by using a global normalization scheme: the mean and standard deviation used for images normalization are the average ones computed for all training cases, as well as the 0.5 and 99.5 percentiles used for clipping;

- *Resampling.* nnUNet resample all images to the same target spacing, because images, especially in the biomedical domain, can have a heterogeneous voxel spacing, while CNNs don't consider this aspect by operating on voxel grids. The default resampling method is third order spline, but for anisotropic images (maximum axis spacing / minimum axis spacing > 3) nearest neighbor method is applied along the low resolution axis (Z axis) to suppress resampling artifacts;
- *Target spacing.* Larger spacings result in too small images and a loss of information, while smaller spacings result in too large images limiting the network on learning conceptual information, so it represents a crucial parameter. For the 3D full-resolution UNet, the chosen default target spacing is the median value of the spacings found in the training cases, independently for each axis. For anisotropic datasets this method can result in a loss of information due to interpolation artifacts, so the target spacing of the lowest resolution axis is selected to be the 10^{th} percentile of the spacings computed in the training images. For the 2D UNet, nnUNet usually applies the same method to the two axes with the highest resolution;
- *Batch size and path size.* A big patch size allows for more contextual information to be aggregated, increasing segmentation results; however, a larger patch size decrease the batch size, resulting in noisier gradients during backpropagation. For these reasons, nnUNet fixes the patch size as large as possible, while allowing a minimum batch size of 2 and remaining inside a predefined GPU memory budget. The patch size is then initialized to the median image shape after resampling and when the patch size is configured, the optional available GPU memory headroom is used to increase the batch size. If the GPU is already fully utilized, the batch size is left to 2;
- *Architecture topology.* The default kernel size for convolution is 3x3x3 (for 3D UNets) and 3x3 (for 2d UNets). The number of downsampling operations along each axis depends on the patch size and voxel spacing. High resolution axes are downsampled separately until their resolution is within factor 2 of the low resolution axis, then all axes are downsampled simultaneously, and this process is stopped for each axis independently when continuing downsampling would produce a feature map smaller than 4 voxels, or the feature map spacings become anisotropic:
- *Adaptation of GPU memory budget.* When the patch size is set, it is initially too large to fit into the GPU in most cases (median image shape after resampling is really large). The patch size is therefore iteratively reduced, while updating network architecture in each step, until the memory budget is reached; the GPU memory consumption is

estimated by nnUNet based on the size of feature maps of the network, and using reference values of known memory consumption.

- *Configuration of 3D UNet cascade.* The 3D UNet cascade is designed so that the second, full resolution 3D UNet refines the segmentation maps of the first one, a low resolution network that uses maximal contextual information to create its output. When the 3D UNet cascade is triggered (only for datasets where the patch size of the 3D full-resolution UNet covers less than 12.5% of the median image shape), the target spacing for downsampled data and the architecture of the 3D low-resolution UNet are configured together iteratively. The target spacing is first set to the target spacing of the full resolution data, and is increased at each step by 1%, while updating the architecture accordingly, until the patch size of the network overcomes 25% of the current median image shape. The configuration of the second UNet is the same as the 3D full-resolution UNet.

2.3.1.3 Empirical parameters

They consist in the choice of the optimal UNet configuration, which is automatically determined by nnUNet based on the average foreground Dice coefficient estimated by cross-validation on the training set. The selected network can be a single model (2D UNet, 3D full-resolution UNet, 3D low-resolution UNet, 3D UNet cascade) or an ensemble of any two of these configurations, merged by averaging softmax probabilities.

To conclude, the last step for the architecture and pipeline definition is Postprocessing. nnUNet performs connected component-based postprocessing, which is commonly used in medical image segmentation because it consists in the removal of all but the largest connected component, and helps to remove spurious false positives. It is performed only once all 5 folds have been trained, and consists on considering all foreground classes as one component, and removing all but the largest component; if this doesn't lead to the reduction of the Dice coefficient for any of the classes, but actually improves it, then this postprocessing method is applied. After that, the same procedure is evaluated for individual classes, assessing if suppressing all but the largest component for each class, considering the difference of Dice coefficients before and after.

2.3.2 Implementation details

nnUNet is implemented using PyTorch (Paszke et al., 2019), and it needs a GPU, with at least 4 GB of VRAM. It already provides a lot of pretrained models for different tasks (e.g., Task001_BrainTumour, Task029_LiTS, Task082_BraTS2020).

It requires to create the following folders: *nnUNet_raw_data_base/nnUNet_raw_data*, where the initial raw data must be saved, *nnUNet_prprocessed*, where nnUNet will automatically save preprocessed data, and *nnUNet_trained_models*, where trained models will be saved, together with validation results and possible inference results.

Raw data must be saved following the Medical Segmentation Decathlon (MSD) dataset format: each segmentation dataset is described by a Task, associated to a specific ID (three-digit integer) and a task name (e.g., Task123_name); corresponding data are stored in a task-specific folder (marked by task ID and name) in nnUNet_raw_data. Each folder contains the following subfolders:

Task123_name/

- dataset.json
- imagesTr
- (imagesTs)
- labelsTr

Where imagesTr is the training set, containing all images used by nnUNet for networks training and architecture's tuning, labelsTr includes the ground truth associated with the training images, while images Ts is optional and represents the test set, that can be used during inference. Dataset.json holds the dataset metadata, and is automatically generated by nnUNet during preprocessing. All images must be 3D nifty files (*.nii.gz*) and follow the naming convention *case_identifier_XXX.nii.gz*, where case_identifier corresponds to a dataset specific name (like BRATS or FETS) followed by a number which is specific for the training case (001, 002, 003 etc.). Moreover, each image file can have multiple modalities, which are identified by a four-digit integer (XXXX) at the end of the filename; the association between this integer and the related modality can be found inside the dataset.json file. In the majority of brain tumor images, the four modalities, and corresponding identifiers, are: FLAIR (0000), T1 (0001), T1ce (0002), T2 (0003).

nnUnet requires also the ground truth segmentations to contain ordered labels, which is trivial for stroke lesions data, that only have two labels (0=background, 1=lesion), but not for the brain

tumors (-BraTS labels are 0,1,2,4). To correct and order those labels, it is possible to apply an nnUNet specific function, called *copy_BraTS_segmentation_and_convert_labels*. It must be specified that, for construction, this function substitutes label 4 with label 3, but also label 2 with label 1 and vice versa, inverting the first two labels.
Once the dataset has been organized in the correct way, preprocessing and subsequent training of the model can be achieved.
At the end of the model training, nnUNet makes automatically available the validation results (segmentation maps and corresponding metrics) for each fold, a visualization of the best performing network architecture, and also a plot of training and validation losses through the epochs.

2.3.3 Application to BraTS 2020

For the application to BraTS 2020 challenge, Isensee et al. incorporated BraTS-specific modifications to nnUNet, which can be considered just a baseline for method development, and so needed to be optimized for BraTS competition. The proposed modifications were (Isensee et al., 2020b):

- *Region-based training*. The provided BraTS labels in ground truth segmentation maps are 'edema', 'non-enhancing tumor and necrosis' and 'enhancing tumor', but (as will be discussed in the analysis of BraTS dataset in Chapter 3) the evaluation of performances is carried out on three overlapping regions: the whole tumor (WT, which corresponds to all three labels), the tumor core (TC, corresponding to the classes 'non-enhancing tumor and necrosis' and 'enhancing tumor') and enhancing tumor. It has been demonstrated that the optimization of these three regions leads to better performances on BraTS datasets, and for this reason the softmax non-linearity was replaced with a sigmoid, and the optimization target was moved to the three regions. Moreover, cross-entropy loss was substituted with binary cross-entropy loss, that works independently in each region;
- *Batch size*. To improve the model accuracy, batch size has been increased from 2 to 5. BraTS dataset has become, indeed, bigger and bigger each year, and with larger datasets a reduced batch size results in noisier gradients, which potentially reduce overfitting but also limit the model accuracy;

- *Data augmentation*. Even if nnUNet already uses a large amount of data augmentations, they proposed even more aggressive augmentations to increase the robustness of the model. More in detail: the probability to apply rotation and scaling was increased from 0.2 to 0.3, the scale range was increased from (0.85, 1.25) to (0.65, 1.6), the scaling factor was chosen to be selected independently for each axis, elastic deformation was applied with a probability of 0.3, brightness augmentation was added with probability of 0.3, the aggressiveness of Gamma augmentation was increased.
- *Batch normalization*. During their participation to other challenges, Isensee et al. noticed that a more aggressive data augmentation can help to reduce the domain gap between different datasets (especially the training and test ones) when used in conjunction with batch normalization. In BraTS the reduction of Dice scores of the test set, compared with training and validation sets, suggests a domain gap between these datasets, so they decided to replace instance normalization with batch normalization;
- *Batch Dice*. Sample Dice was replaced by Batch Dice. Sample Dice was the common Dice loss used, consisting in computing the loss for every sample in the minibatch independently, and then average the losses; but small errors in samples with few annotated voxels can cause large gradients. If these errors are caused by model imperfections, they should push the model to better predictions, but if they are caused by imperfect labels, these large gradients will be counterproductive during training. For this reason, they introduced Batch Dice, consisting in computing the dice loss over all samples in the minibatch, considering them just as a single large sample. In this way, samples with few annotated voxels are shadowed by the other samples in the batch;
- *Postprocessing*. BraTS challenge uses a 'rank than aggregate' approach for methods ranking, which means that each test case is ranked six times (one for each of the three evaluated regions, times two evaluation metrics, Dice score and Hausdorff distance HD95), and then the ranks are averaged across all cases and metrics. The final ranking is hence normalized by the number of participating algorithms, so that it ranges from 0 to 1. Knowing this, when a specific image doesn't contain 'enhancing tumor' region, BraTS identifies zero false positives, assigning a Dice score of 1 (it would be otherwise undefined due to division by 0), placing the corresponding model in the first rank for this test image. This aspect was exploited by introducing a postprocessing method which entirely removes 'enhancing tumor' region, if the predicted volume is less than a certain threshold, optimized during cross-validation, once by maximizing the mean Dice score, and once minimizing the ranking score. Removed enhancing tumor was replaced with

necrosis, guaranteeing that corresponding voxels continued to be considered part of the tumor core. In this way, they were able to obtain more perfect rankings, despite having some additional cases with a Dice score of 0 (worst rank), but the net gain out-weighed the losses.

These modifications were combined in different ways, leading to the training of various configurations. Models were ranked based on their performances on training and validation sets, leading to different top-performing networks; they chose to trust more the results on the validation set and selected the three best models: the one condensing all proposed modifications, the model with all changes but batch dice, and the network with all adjustments but without using more aggressive data augmentations. For each configuration, the five models from cross-validation, together with 10 model of the second network, each trained with a random 80:20 split of the training set, where ensembled, consisting in an ensemble of 5+5+5+10=25 models. The final ensemble achieved mean Dice scores of 88.95, 85.06 and 82.03 and HD95 of 8.498, 17.337 and 17.805 for whole tumor, tumor core and enhancing tumor, respectively on the test set, taking the first place in the competition (Isensee et al., 2020b).

2.3.4 Extension for the application to BraTS 2021

For the application to BraTS 2021 challenge, H.M. Luu et al. extended nnUNet by experimenting several modifications, with which they won the first place in the final ranking on unseen test data (Luu & Park, 2021).

The architecture used was the one including BraTS-specific modifications, with the inclusion of multiple architectural changes: first of all, the size of the network was asymmetrically increased by doubling the number of filters only in the encoder, to deal with the large BraTS dataset. BraTS 2021 dataset, indeed, includes four times the number of images with respect to BraTS 2020 dataset, so increasing the size of the network could help to model the larger variety of data and extract deep image features; the maximum number of filters was also increased, for this purpose, from 320 to 512.

The second proposed modification was to substitute batch normalization with group normalization. As previously stated, 3D convolutional networks require a high amount of GPU memory, which automatically limits the batch size that can be used during training. It has been

demonstrated that Group Normalization works better than batch normalization with low batch size, because its accuracy is stable for a large range of batch sizes, and its efficiency is independent from them(Wu & He, n.d.). The idea behind Group Normalization is to divide channels in groups (32 in this case), and normalize their feature maps by computing mean and standard deviation within each group, exploiting the consideration that there are groups of channels which share similar properties, and that can then be normalized together.

The final architectural change was the introduction of axial attention in the decoder. Self-attention mechanisms or transformers were originally introduced for Natural Language Processing (NLP), and basically consist in assigning a weight to each part of the input (each word for NLP) and updating that weight based on the dependency of each word from the context. In this way the input iteratively interacts with itself, learning where to give more importance, and allowing not to lose long-term dependencies. Its application to Computer Vision represents a breakthrough because, especially in the field of semantic segmentation, it addresses training towards the most relevant regions in the image. Self-attention layers in Computer Vision take feature maps as inputs and learn 'attention weights' between features; but the problem of its application in the segmentation field is that the computational cost of attention mechanisms scales quadratically with the input images size, becoming too large especially with 3D input data.

To overcome these problems on the application of attention mechanisms on multi-dimensional data, axial attention was introduced (Ho et al., 2019). Axial attention consists in the application of self-attention independently to each axis of the input, allowing the computation to scale only linearly with the image size, letting attention mechanisms to be applied also with 3D data. H.M. Luu et al. applied this method on the four lower resolutions of the network, finding that it was not possible to apply it to the highest resolution features (128x128x128). Axial attention consists in running the attention block on the output of the transposed convolution upsampling for each axis, and the results are then summed and added to the original input (*Figure 2.2*).

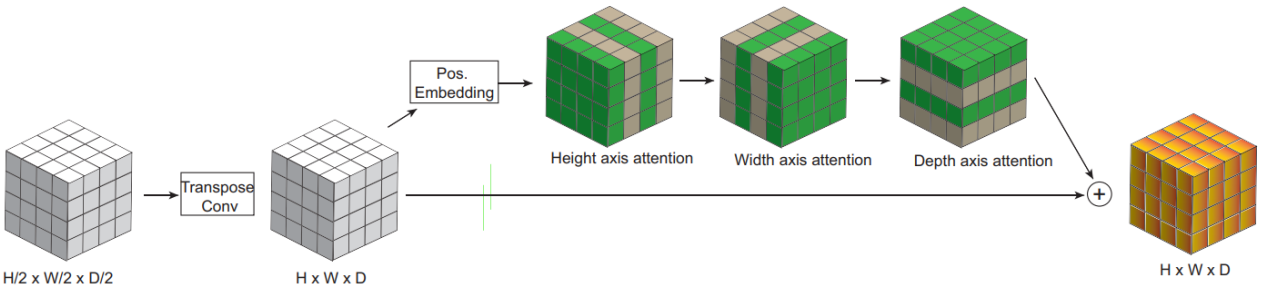


Figure 2.2: Explanation and visualization of the axial attention block. Self-attention is applied to each axis of the output from the transposed convolution in the decoder, results are then summed and added back to the original input, which is concatenated to the encoder.

The followed training procedure was the same as the original nnUNet, but the new, experimented modifications were tested with a batch size of two. By testing singular proposed changes, they observed that using group normalization instead of batch normalization decreased the dice metric; using the axial attention encoder did not improve the performance, but combining a larger encoder and group normalization increased slightly the results for the tumor core and the whole tumor, with the cost of a larger usage of GPU memory.

The final method was then obtained by ensembling the segmentations obtained by the tested configurations, leading to an improvement of all metrics when compared to the baseline nnUNet: average Dice score across all classes increased from 87.94 to 88.36, and HD95 decreased from 12.18 to 10.61.

2.4 Relevant architectures: IVD-Net: Dense Multi-path U-Net

For the purposes of this thesis, which are related on treating in a separate way images of different modalities, avoiding fusing at early stages information extracted from, in some cases, images with opposite intensities, another relevant and peculiar architecture is IVD-Net. The network of interest was published by Dolz et al. in 2018, and consisted in the development of a dense multi-path UNet for ischemic stroke lesion segmentation, but the corresponding code was not made publicly available.

By contacting directly the author (Jose Dolz), he highlighted the existence of a second architecture (IVD-Net), which was developed by the same working group, with an identical

structure, but for the purpose of segmentation of intervertebral disc (IVD) from multi-modal images (Dolz, Desrosiers, et al., 2018), whose code was publicly released.

The great advantage and improvement of this method is the development of a UNet architecture in which each image modality is processed by a different encoder to better exploit their unique information, allowing to extract relevant features from each modality independently. The majority of networks in fact, included nnUNet (where different input modalities are treated as different color channels), perform an early fusion strategy, based on merging multiple modalities from the original input space of low-level features (Dolz, Ayed, et al., 2018), assuming a simple relationship between modalities, which often doesn't correspond to reality, due to their different acquisition setups.

To keep images of different modalities separated, and allow the fusion of the extracted features only at late levels, the authors chose to disentangle input data, splitting the encoding path of the UNet architecture into N streams, one for each input modality (*Figure 2.3*); the only drawback of this structure is the incapability of extracting complex relationships between modalities, which have been demonstrated to be highly complex and could allow to extract relevant information for the segmentation task. For this reason, it was chosen to implement hyper-dense connections within the same and between multiple paths, to model better the existing and complex relationships between different modalities without merging them, to help improving the flow of information and gradients through the entire network. Dense connections also have a regularizing effect, reducing the risk of overfitting.

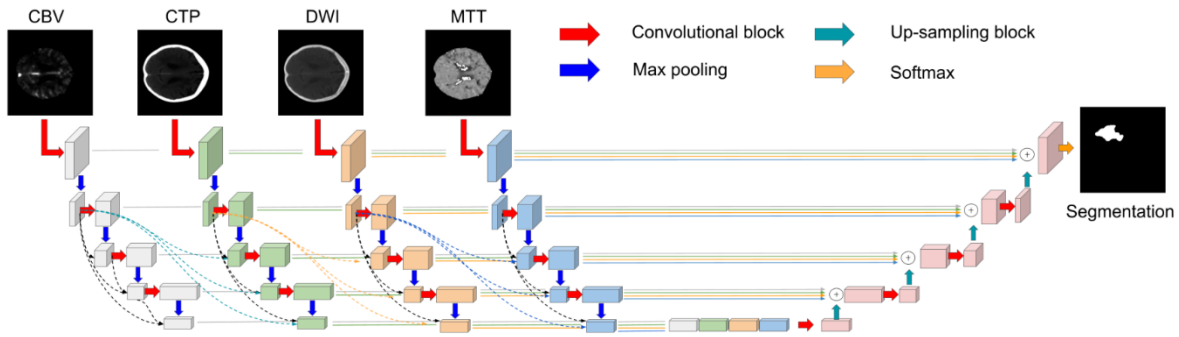


Figure 2.3: Representation of the multi-path dense UNet developed by Dolz et al. (taken from Dolz, Ayed, et al., 2018) for ischemic stroke lesion segmentation task. It can be noticed the presence of one different stream for each acquisition modality, whose feature maps are then concatenated creating a "bridge" before the decoding path. Dotted lines represent some of the dense connections adopted between and within paths.

While in standard CNNs, the output of each layer is obtained from the output of the previous layer usually by the application of a convolution layer followed by a non-linear activation; instead, in densely-connected networks all the outputs of previous layers are concatenated and

used as input of the current layer, not to lose past information. Moreover, hyper-dense connections allow to link also outputs from different paths, leading to a much more powerful representation than early or late fusion strategies, allowing the network to learn complex relationships between different modalities within and in-between all the levels of abstraction (Dolz, Ayed, et al., 2018).

The last addition introduced by the authors was the extension of the convolutional module of InceptionNet, facilitating learning of multiple context information. In ischemic stroke lesion segmentation, but also in brain tumor segmentation, the size of the area occupied by the lesion or by the tumoral subregions highly changes from one image to another, compromising the choice of the optimal kernel size. InceptionNet(Szegedy et al., 2016) includes convolutions with multiple kernel sizes operating on the same level, solving the problem of the choice of the kernel size, but increasing also the efficiency by factorizing n×n convolutions into a combination of 1×n and n×1 convolutions, which have been demonstrated to be more efficient. Dolz et al. further extended InceptionNet by adding two convolutional blocks with different dilation rates, which help the module to learn from multiple receptive fields and to increase the context.

Nothing changes in the architecture of IVD-Net, only the task. These methods were implemented to work with 2D images.

3. Publicly available datasets

First of all, a review about the most employed datasets in the fields of brain tumors and stroke lesions segmentation was performed, to allow a better understanding about the growing dimension of available data, and to point out the most advanced and avant-garde datasets.
Datasets were identified performing a state-of-the-art analysis about brain tumor and stroke lesion segmentation techniques, and searching information about the most cited datasets in the papers analyzed.

3.1 Brain Tumor datasets

Most of the algorithms developed in last ~5 years for brain tumor segmentation were trained and implemented using Brain Tumor Segmentation (BraTS) challenge datasets.
The first BraTS challenge was held in conjunction with the 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2012) on October 1st, 2012 in Nice, France (MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation Proceedings of MICCAI-BRATS 2012 October 1 St, Nice, France, n.d.).
At the time it was difficult to compare methods because of the large difference between validation datasets employed, type of lesions and state of the disease; for these reasons the aim of BraTS 2012 was to make available a large dataset of glioma brain tumor MR scans, in which tumor regions have been manually delineated by expert neuroradiologist. The introduction of such dataset allowed programmers to train algorithms on a dataset significantly large for that time, and that enabled comparison of different methods, because they were all evaluated on the same type of scans and lesions.
After 2012, BraTS challenge was proposed each year with a larger dataset, with new scans provided by new centers (BraTS 2012 consisted just of 80 gliomas from real patient cases) and with new challenges.
Because of the growing dimension of these datasets, I started analyzing from BraTS 2017, because previous editions are no more used. All the detailed information about these datasets is provided in *Table 3.1*.

BraTS 2017 consists of 477 (285 for training, 46 for validating, 146 for testing) pre-operative MRI scans provided by 16 distinct institutions worldwide. Different sequences of MRI scans are supplied: T2 and FLAIR images, which mostly highlight the whole tumor region (including infiltrative edema), and T1 and T1ce images, which give a better contrast for the tumor core region (G. Wang et al., 2019) (same modalities are provided every year). All images were preprocessed by co-registration to the same template, resampling to a uniform resolution ad skull-stripping; then they were segmented manually by 1-4 raters, and after annotations were approved by expert neuro-radiologists. The Challenge tasks were: 1) Segmentation of gliomas in sub-regions: the enhancing tumor (ET), the tumor core (TC) and the whole tumor (WT). The ET is characterized by areas that show hyper-intensity in T1ce when compared to T1; the TC represents what is typically resected, and it entails the ET, as well as the necrotic (NCR) and non-enhancing (NET) tumor core, which are typically hypo-intense in T1ce when compared to T1; finally, the WT represents the entire tumor, and it encloses the TC and the peritumoral edema (ED), which is typically hyper-intense in FLAIR. The labels in provided data were: 1 for NCR/NET, 2 for ED, 4 for ET, 0 for anything else, as specified in *Figure 3.1* (*MICCAI_BraTS_2017_proceedings_shortPapers*, n.d.). This task is shared by all the other BraTS challenges. 2) Prediction of patient overall survival, by extracting imaging features from the given data.

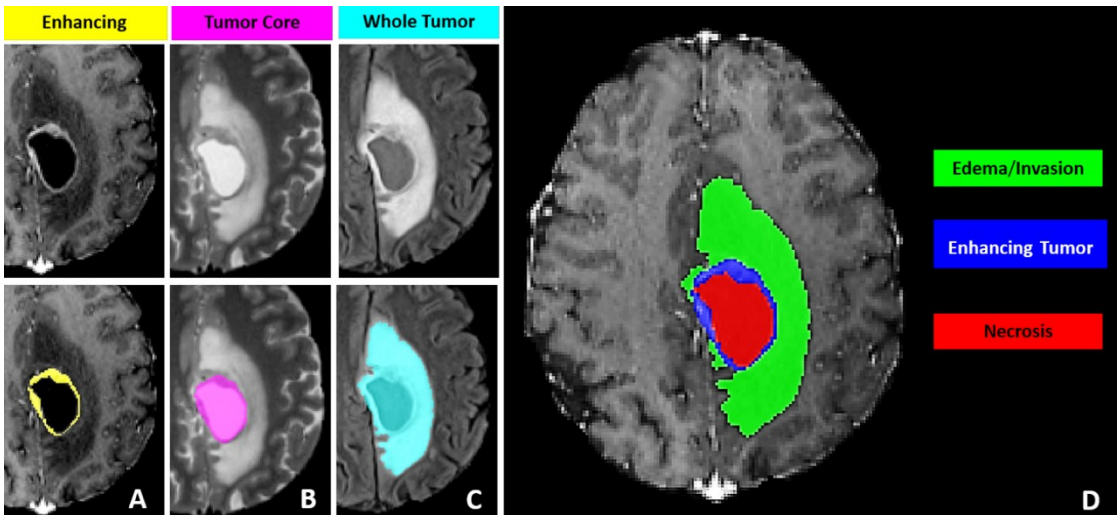


Figure 3.1: Glioma subregions considered in BraTS challenge: the provided labels are shown in panel D, and consist in Necrotic Tumor Core (label 1), Enhancing Tumor (label 2), and peritumoral edematous/invaded tissue (label 4), while the evaluation is performed based on Enhnacing tumor (panel A) which coincides with label 2, Tumor Core (panel B), which contains label 1 and 2, and Whole Tumor (panel C) that contains all labels.

BraTS 2018 dataset consists of 542 (285 for training, 66 for validating, 191 for testing) MRI scans provided by 19 different centers. Also in this case, images pre-processing and annotation

protocols were the same for all the BraTS challenges. The tasks were the same then BraTS 2017.

BraTS 2019 dataset consists of 646 (355 for training, 125 for validating and 166 for testing) MRI scans provided by 19 separate institutions. A third task was added for this challenge: quantification of uncertainty in segmentation, with the aim of rewarding participation methods which are confident when correct and uncertain when incorrect. An uncertainty map (associated with the traditional BraTS Dice metric) was provided for each segmented subregion.

BraTS 2020 dataset consists of 800 (420 for training, 150 for validating, 230 for testing) pre-operative and post-operative MRI scans provided by 21 different centers. Also in this case a third task was added, different from the one proposed in 2019: distinction of tumor recurrence from treatment related effects, with the aim of developing radiomics and machine learning solutions to reliably distinguish benign pseudo-progression (PsP), a benign side-effect of the chemoradiation therapy, from tumor recurrence.

BraTS 2021 dataset consists of 2000 (1251 for training, 219 for validating, 570 for testing) pre-operative and post-operative multi-parametric MRI (mpMRI) scans provided by 21 distinct institutions. It focuses only on glioblastoma and task 2 changed with respect to the previous editions: determination of MGMT (O[6]-methylguanine-DNA methyltransferase) promoter methylation status on pre-operative scans via integrative analysis of quantitative imaging features and machine learning algorithms. MGMT is a DNA repair enzyme, and the methylation of its promoter in diagnosed glioblastoma has been identified as a favorable prognostic factor and a predictor of chemotherapy response, and for this reason could influence decision making and treatment planning (Baid et al., 2021).

Another popular dataset is Medical Segmentation Decathlon (MSD) which is composed by 2633 3D images collected across multiple anatomies of interests, multiple modalities and multiple institutions, for various tasks. The most important for my purposes is the one associated with Task01_BrainTumor, which is just composed by a total of 750 (484 for training, 266 for testing) MRI scans coming from data used in 2016 and 2017 BraTS challenges.

Among the most advanced and newest datasets, Federated Tumor Segmentation (FeTS) 2021 and 2022 datasets must be cited. The performance of algorithms on “real-world” clinical data often remains unclear, as the data included in international challenges and used to train those algorithms are usually acquired in controlled settings by few institutions, with privacy and ownership hurdles. For these reasons, FeTS challenge was introduced: it includes scans from previous BraTS challenges and its goals are: 1) the training of models via federated learning from multiple institutions, while their data are kept within each institution, and 2) the evaluation

of the generalizability of brain tumor segmentation models on data different from training datasets, to assess their robustness (Pati et al., 2021).

Table 3.1: Most used datasets and related characteristics in the field of brain tumor segmentation

Dataset	Aim	Size	Centers providing Data	Acquisition Protocol	Modalities	Volume Size	Resolution	Preprocessing	Annotations	Normalization	N. Classes	Segmentation task
BraTS 2017	Brain Tumor Segmentation	477 (285 for training, 46 for validating, 146 for testing)	1) Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania (UPenn) (PA, USA), 2) University of Alabama at Birmingham (AL, USA), 3) Heidelberg University (Germany), 4) University Hospital of Bern (Switzerland), 5) University of Debrecen (Hungary), 6) Henry Ford Hospital (MI, USA), 7) University of California (CA, USA), 8) MD Anderson Cancer Center (TX, USA), 9) Emory University (GA, USA), 10) Mayo Clinic (MN, USA), 11) Thomas Jefferson University (PA, USA), 12) Duke University School of Medicine (NC, USA), 13) Saint Joseph Hospital and Medical Center (AZ, USA), 14) Case Western Reserve University (OH, USA), 15) University of North Carolina (NC, USA), 16) Fondazione IRCCS Instituto Neurologico C. Besta (Italy)	MRI	T1, T1Gd, T2, and T2-FLAIR	240x240x155	1 mm³	Co-registered to the same anatomical template, interpolated to the same resolution and skull-stripped	Segmented manually by 1-4 raters, following the same annotation protocol, and then approved by experienced neuro-radiologists	N4BiasCorrection and intensity normalization to zero-mean and unit variance	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
BraTS 2018	Brain Tumor Segmentation	542 (285 for training, 66 for validating, 191 for testing)	1) Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania (UPenn) (PA, USA), 2) University of Alabama at Birmingham (AL, USA), 3) Heidelberg University (Germany), 4) University Hospital of Bern (Switzerland), 5) University of Debrecen (Hungary), 6) Henry Ford Hospital (MI, USA), 7) University of California (CA, USA), 8) MD Anderson Cancer Center (TX, USA), 9) Emory University (GA, USA), 10) Mayo Clinic (MN, USA), 11) Thomas Jefferson University (PA, USA), 12) Duke University School of Medicine (NC, USA), 13) Saint Joseph Hospital and Medical Center (AZ, USA), 14) Case Western Reserve University (OH, USA), 15) University of North Carolina (NC, USA), 16) Fondazione IRCCS Instituto Neurologico C. Besta (Italy), 17) MD Anderson Cancer Center (TX, USA), 18) Washington University School of Medicine in St. Louis (MO, USA), 19) Tata Memorial Center, Mumbai (India)	MRI	T1, T1Gd, T2, and T2-FLAIR	240x240x155	1 mm³	Co-registered to the same anatomical template, interpolated to the same resolution and skull-stripped	Segmented manually by 1-4 raters, following the same annotation protocol, and then approved by experienced neuro-radiologists	N4BiasCorrection and intensity normalization to zero-mean and unit variance	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
BraTS 2019	Brain Tumor Segmentation	646 (355 for training, 125 for validating, 166 for testing)	1) Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania (UPenn) (PA, USA), 2) University of Alabama at Birmingham (AL, USA), 3) Heidelberg University (Germany), 4) University Hospital of Bern (Switzerland), 5) University of Debrecen (Hungary), 6) Henry Ford Hospital (MI, USA), 7) University of California (CA, USA), 8) MD Anderson Cancer Center (TX, USA), 9) Emory University (GA, USA), 10) Mayo Clinic (MN, USA), 11) Thomas Jefferson University (PA, USA), 12) Duke University School of Medicine (NC, USA), 13) Saint Joseph Hospital and Medical Center (AZ, USA), 14) Case Western Reserve University (OH, USA), 15) University of North Carolina (NC, USA), 16) Fondazione IRCCS Instituto Neurologico C. Besta (Italy), 17) MD Anderson Cancer Center (TX, USA), 18) Washington University School of Medicine in St. Louis (MO, USA), 19) Tata Memorial Center, Mumbai (India)	MRI	T1, T1Gd, T2 and T2-FLAIR	240x240x155	1 mm³	Co-registered to the same anatomical template, interpolated to the same resolution and skull-stripped	Segmented manually by 1-4 raters, following the same annotation protocol, and then approved by experienced neuro-radiologists	Intensity normalization performed across subjects based on the median intensity value of the cerebrospinal fluid label	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
BraTS 2020	Brain Tumor Segmentation	800 (420 for training, 150 for validating, 230 for testing)	1) University of Pennsylvania (PA, USA), 2) University of Alabama at Birmingham (AL, USA), 3) Heidelberg University (Germany), 4) University of Bern (Switzerland), 5) University of Debrecen (Hungary), 6) Henry Ford Hospital (MI, USA), 7) University of California (CA, USA), 8) MD Anderson Cancer Center (TX, USA), 9) Emory University (GA, USA), 10) Mayo Clinic (MN, USA), 11) Thomas Jefferson University (PA, USA), 12) Duke University School of Medicine (NC, USA), 13) Saint Joseph Hospital and Medical Center (AZ, USA), 14) Case Western Reserve University (OH, USA), 15) University of North Carolina (NC, USA), 16) Fondazione IRCCS Instituto Neurologico C. Besta, (Italy), 17) MD Anderson Cancer Center (TX, USA), 18) Washington University in St. Louis (MO, USA), 19) Tata Memorial Center (India), 20) University of Pittsburg Medical Center (PA, USA), 21) Cleveland Clinic (OH, USA), 22) Brigham and Women's Hospital (MA, USA)	MRI	T1, T1Gd, T2, and T2-FLAIR	240x240x155	1 mm³	Co-registered to the same anatomical template (SRI ATLAS), interpolated to the same resolution and skull-stripped	Segmented manually by 1-4 raters, following the same annotation protocol, and then approved by experienced neuro-radiologists	Intensity normalization performed across subjects based on the median intensity value of the cerebrospinal fluid label	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
BraTS 2021	Brain Tumor Segmentation	2000 (1251 with labels for training, 219 for validating, 570 for testing)	1) University of Pennsylvania (PA, USA), 2) University of Alabama at Birmingham (AL, USA), 3) Heidelberg University (Germany), 4) University of Bern (Switzerland), 5) University of Debrecen (Hungary), 6) Henry Ford Hospital (MI, USA), 7) University of California (CA, USA), 8) MD Anderson Cancer Center (TX, USA), 9) Emory University (GA, USA), 10) Mayo Clinic (MN, USA), 11) Thomas Jefferson University (PA, USA), 12) Duke University School of Medicine (NC, USA), 13) Saint Joseph Hospital and Medical Center (AZ, USA), 14) Case Western Reserve University (OH, USA), 15) University of North Carolina (NC, USA), 16) Fondazione IRCCS Instituto Neurologico C. Besta, (Italy), 17) MD Anderson Cancer Center (TX, USA), 18) Washington University in St. Louis (MO, USA), 19) Tata Memorial Center (India), 20) University of Pittsburg Medical Center (PA, USA), 21) University of California San Francisco (CA, USA).	MRI	T1, T1Gd, T2, and T2-FLAIR	240x240x155	1 mm³	Conversion of DICOM file into Nifti file formats, coregistration to the same template (SRI24), resampling to a uniform resolution, skull-stripping	Segmented manually by 1-4 raters, following the same annotation protocol, and then approved by experienced neuro-radiologists	By their mean and SD	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
Oslo Dataset	Brain Tumor Segmentation	271	52 preoperative subjects with LGG and HGG scanned from 2003 to 2012 from Oslo University Hospital (Oslo, Norway)	MRI	T2w, T2w-FLAIR, T1, and T1c/T1	12-14 slices 230x230	1 mm³	Coregistration to the T1 postcontrast, registering to the SRI24 template, resampling, skull stripping	Manually segmented by an in-house neuroradiologist	N4BiasCorrection and intensity normalization to zero-mean and unit variance	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
Medical Segmentation Decathlon (MSD)	Brain Tumor Segmentation	750 (484 for training, 266 for testing)	1) Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania (UPenn) (PA, USA), 2) University of Alabama at Birmingham (AL, USA), 3) Heidelberg University (Germany), 4) University Hospital of Bern (Switzerland), 5) University of Debrecen (Hungary), 6) Henry Ford Hospital (MI, USA), 7) University of California (CA, USA), 8) MD Anderson Cancer Center (TX, USA), 9) Emory University (GA, USA), 10) Mayo Clinic (MN, USA), 11) Thomas Jefferson University (PA, USA), 12) Duke University School of Medicine (NC, USA), 13) Saint Joseph Hospital and Medical Center (AZ, USA), 14) Case Western Reserve University (OH, USA), 15) University of North Carolina (NC, USA), 16) Fondazione IRCCS Instituto Neurologico C. Besta (Italy), 17) Washington University School of Medicine in St. Louis (MO, USA), 18) Tata Memorial Center, Mumbai (India)	MRI	T1, T1Gd, T2, and T2-FLAIR	240x240x155	1 mm³	Co-registered to a reference atlas space using the SRI24 brain structure template, resampled to the same voxel resolution and skull-stripped	Gold standard annotations for all tumour subregions in all scans were approved by expert board-certified neuroradiologists	/	3	Whole tumor (WT), Tumor core (TC), Enhancing tumor (ET)
Federated Tumor Segmentation (FeTS) 2021	Brain Tumor Segmentation	not mentioned	FeTS 2021 training data represent just a subset of BraTS 2020 data, but with the addition of non-imaging data including informations about their partitioning based on the originating institution (*partitioning_1.csv*) and extra partitioning based on the median tumor size was made in the 5 largest institutions (*partitioning_2.csv*). The test data are a subset of BraTS 2020 test data and data offered by independent geographically distinct institutions that participated in the FeTS federation									
Federated Tumor Segmentation (FeTS) 2022	Brain Tumor Segmentation	1666 for training (each with 4 modalities)	FeTS 2022 data are FeTS 2021 data with many more routine clinically-acquired mpMRI scans									

3.2 Stroke Lesion datasets

In the field of stroke lesions segmentation, the most used datasets are Anatomical Tracings of Lesions After Stroke (ATLAS) and Ischemic Stroke Lesion Segmentation (ISLES) challenge datasets.

The first version of ATLAS (v1.2) was released in 2018 and it comprised 304 T1w MRIs (T1 is a commonly used modality in stroke rehabilitation research, as it provides excellent spatial resolution at subacute or chronic stage, when rehabilitation usually starts) and manually segmented lesion masks. However, many methods developed using this dataset reported low accuracy and were improperly validated, limiting their application. For these reasons, in 2021 it was introduced ATLAS v2.0, a large dataset of 955 T1w MRI scans, composed by 655 public training images and 300 hidden test images coming from 11 different cohorts worldwide and harmonized by the Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) stroke recovery working group (Liew et al., n.d.). ATLAS was also released with an additional hidden test set composed by 135 MRI scans coming from completely different cohorts, available only on segmentation challenges for the evaluation of generalizability of algorithms on unseen data. Images were first manually segmented by team members, who received lesion-tracing training and followed a standard protocol for lesions identification to ensure consistency among tracers. After that, lesion masks were checked by two separate raters. The raw dataset was released together with a preprocessed dataset, following the same pipeline of ATLAS v1.2: intensity normalization, registration to a standard template and remotion of non-brain data.

Current challenges using ATLAS data are Rapid Analytics and Model Prototyping (RAMP) Challenges and Ischemic Stroke Lesion Segmentation (ISLES) Challenge (Liew et al., n.d.).

The ISLES challenge is one of the best-known stroke lesion segmentation challenges; it was firstly introduced in 2015 at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in Munich, Germany, with two sub-challenges: Sub-Acute Stroke Lesion Segmentation (SISS), dealing with the later stroke analysis, and Stroke Perfusion Estimation (SPES), dealing with the study of the acute phase of stroke. The reasons for its introduction were multiple: first of all, stroke lesions vary significantly their appearance over time, especially starting from the sub-acute phase, at the beginning of which lesions usually are hyperintense in DWI sequence and moderately hyperintense in FLAIR; while, towards the second week, the hyperintensity in FLAIR sequence increases while the DWI becomes isointense, underlying the necessity of acquiring multiple modalities to have the whole picture about stroke lesion progression. Second, lesions can show in any location in the

brain and with any shape, not always manifesting as homogeneous regions. To make ischemic stroke lesion segmentation even more complicated, their appearance can be really similar to other pathologies, like chronic stroke lesions or white matter hyperintensities (WMHs) (Maier et al., 2017).

SISS dataset is composed by 64 sub-acute ischemic stroke cases (28 for training, 36 for testing) provided by two institutions, and the modalities included were T1, T2, DWI and FLAIR. MRI sequences were skull-stripped, resampled to an isotropic spacing and co-registered to the FLAIR sequence; annotations were created on the FLAIR sequence (lower inter-rater differences) by experienced raters.

SPES dataset is composed by 50 MRI scans (30 for training, 20 for testing) provided by a single center; sequences acquired were T1ce, T2, DWI, cerebral blood flow (CBF), cerebral blood volume (CBV), time-to-peak (TTP) and time-to-max (Tmax). All images were skull-stripped, resampled to an isotropic spacing and registered to the T1ce sequence, and they were segmented semi-manually by a medical doctor with Slicer 3D Version 4.3.1.

ISLES 2016 dataset consists of 54 MRI scans (35 for training, 19 for testing) acquired from patients treated for acute ischemic stroke at the University Hospital of Bern. All MRI sequences were initially skull-stripped and anonymized, and they were annotated by a certified neuroradiologist using again 3D Slicer v4.5.0-1 and based on a 90-day follow up T2 image acquired only for this purpose and not released. Test cases were annotated by two raters independently, and segmentation maps were merged via the STAPLE algorithm. The associated challenge consisted in two tasks: I) lesion outcome prediction (segmentation map generation) and II) clinical outcome prediction, through the mRM score, denoting the degree of disability (Winzeck et al., 2018).

ISLES 2017 dataset consists of 75 MRI scans (43 for training, 32 for testing) acquired, preprocessed and annotated in the same way than ISLES 2016 dataset, but in this case, the additional test cases were annotated by a single rater.

ISLES 2018 dataset consists of 156 MRI/CTP scans (94 for training, 62 for testing) from patients presenting acute ischemic stroke from 3 US centers and 1 Australian center; CT scans were acquired within 8 hours from the stroke onset. CT scans were motion corrected and registered to a standard template, while DWI scans were coregistered to the corresponding CTP acquisitions; annotations were manually delineated by expert neuroradiologists, and then subjected to group review, on additional MR DWI images not included in the challenge dataset,

where the infarct zone is seen more clearly and which were acquired within 3 hours of the initial CT scan (Clèrigues et al., 2019).

ISLES 2022 dataset consists of 400 pre- and post- interventional MRI scans (250 for training, 150 for testing) provided by 3 institutions. Data were anonymized, brain-extracted, FLAIR images were coregistered to corresponding DWI, and DWI and ADC maps were skull-stripped and resliced. For labeling purposes, a 3D Unet was trained on DWI data that were pre-annotated for other research projects by using a single MRI modality. Segmentations were then checked by medical students with special stroke lesion segmentation training, and their annotations were revised by a neuroradiology resident. After a new pre-segmentation algorithm was trained on correct annotations, and the process was repeated (Petzsche et al., 2022). The aim of this year's challenge is different from previous editions because it targets I) the delineation of not only large infarct lesions but also multiple embolic and/or cortical infarcts from DWI, ADC and FLAIR images; and II) single-channel T1w lesion segmentation in acute, sub-acute and chronic stroke (ATLAS challenge).

All these information can be seen in detail in *Table 3.2.*

Dataset	Aim	Size	Centers providing Data	Acquisition Protocol	Modalities	VolumeSize	Resolution	Preprocessing	Annotations	N. Classes	Segmentation task
ATLAS (Anatomical Tracings of Lesions After Stroke) v2.0 (previous v1.2 released in 2018)	Stroke Lesions Segmentation	955 (304 in v1.2) (655 for training, 300 for testing) + 135 for testing from completely different cohorts	33 research cohorts across 20 institutions worldwide, including: 1) University of Southern California (Los Angeles, California 90089, USA), 2) University of California Irvine, Irvine, California 92697, USA), 3) Tianjin Medical University General Hospital (Tianjin 30052, China), 4) University of Tübingen (Tübingen 72076, Germany), 5) Sirmass Rehabilitation Hospital HT (Nieuwkuijen 1453, Norway), 6) NORMENT and KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital (Oslo 0372, Norway), 7) Department of Psychology, University of Oslo (Oslo 0315, Norway), 8) Child Mind Institute (New York, New York 10022, USA), 9) Nathan S. Kline Institute for Psychiatric Research (Orangeburg, New York 10962, USA), 10) University of Texas Medical Branch (Galveston, Texas 77555, USA), 11) University of Michigan (Ann Arbor, Michigan 48104, USA)	MRI	T1	197 × 233 × 189	1 mm^3	Raw and pre-processed (defaced and normalized to the MNI-152 template) datasets released	Images were first manually segmented using ITK-SNAP by tracers, who received lesion-tracing training and followed a standard protocol for lesions identification to ensure consistency among tracers. After that, lesion masks were checked by two separate trained team members.	Variable	Varying lesions sizes and locations (cortical, subcortical or other lesion locations)
SPES (Stroke Perfusion Estimation) dataset provided by ISLES (Ischemic Stroke Lesion Segmentation) 2015	Stroke Lesions Segmentation	50 (30 for training, 20 for testing)	Patients treated at the University Hospital of Bern between 2005 and 2013	MRI	T1c, T2, CBF, CBV, DWI, Tmax, and TTP	24 slices 256x256 (DWI), 19 slices 230x230 (PWI)	2 mm^3	Skull stripped using BET2, resampled to an isotropic spacing and co-registered to the T1w sequence	Segmented semi-manually with Slicer 3D Version 4.3.1 by a medical doctor with a predjusted threshold for Tmax of 6 seconds applied to regions of interest, followed by a manual correction step consisting in removing subcortical, non-stroke pathologies and previous infarcts	3	Core(dead), penumbra(dead, intergeodesic) (only penumbra(dead used in ISLES challenge)
SSIS (Sub-acute Ischemic Stroke Lesion Segmentation) dataset provided by ISLES (Ischemic Stroke Lesion Segmentation) 2015	Stroke Lesions Segmentation	64 (28 for training, 36 for testing)	1) 35 cases supplied by the University Medical Center Schleswig-Holstein (Lübeck, Germany), 2) 6 cases added to the test set, scanned at the Department of Neuroradiology at the Klinikum rechts der Isar (Munich, Germany)	MRI	FLAIR, T2, T1, DWI	Not mentioned	1 mm^3	Skull stripped, resampled to an isotropic spacing and co-registered to the FLAIR sequence	Segmentation performed on FLAIR sequence and 2 ground truth sets (GT01 and GT02) were created	1	Sub-acute infarct
ISLES (Ischemic Stroke Lesion Segmentation) 2016	Lesion and Clinical Outcome prediction of stroke lesions	54 (35 for training, 19 for testing)	Patients treated for acute ischemic stroke at the University Hospital of Bern or at the UMC Freiburg between 2005 and 2015	MRI	ADC, rBF, rBV, MTT, TMAX, TTP, and Raw PWI	24 slices 256x256 (DWI), 19 slices 230x230 (PWI)	1 mm^3	Skull-stripped and anonymization was performed	Lesion outcome status was manually segmented by a board certified neuroradiologist using 3D Slicer v4.5.0-1, and based on the 90-day follow-up T2 image; lesion labels for the test data were generated by two raters independently, and merged via the STAPLE algorithms	1, different clinical parameters	Lesion outcome prediction
ISLES (Ischemic Stroke Lesion Segmentation) 2017	Lesion and Clinical Outcome prediction of stroke lesions	75 (43 for training and 32 for testing)	Patients treated for acute ischemic stroke at the University Hospital of Bern or at the UMC Freiburg between 2005 and 2015	MRI	ADC, rBF, rBV, MTT, TMAX, TTP, Raw PWI and T2-FLAIR	from 128x128 to 256x256 x19 to 30)	from 0.9x0.9 to 1.8x1.8) x 5 (to 6.5) mm	Skull stripped, anonymized and co-registered for each subject individually	Lesion outcome status was manually segmented by a board certified neuroradiologist using 3D Slicer v4.5.0-1, and based on the 90-day follow-up T2 image; only core ground truth is labeled for test data	1, different clinical parameters	Lesion outcome prediction
ISLES (Ischemic Stroke Lesion Segmentation) 2018	Stroke Lesions Segmentation	156 (94 for training, 62 for testing)	103 patients presenting with acute large artery occlusion anterior circulation ischemic stroke from 3 US centers and 1 Australian center	MRI/CTP scans	CT, CT-PWI and CBF, CBV, MTT, Tmax, DWI	from 2 to 22 axial slices 256x256	from 0.8x0.80 to 1.04x1.04) x4 to 12) mm	CTP motion corrected, resampled to a standard temporal resolution; DWI was coregistered to the CTP acquisition by aligning both to the Montreal Neurological Institute atlas. DWI provided only for training data	The provided gold standard was manually delineated by a a stroke neurologist with >10 years of experience, and then subjected to group review, on additional MR DWI trace images not included in the testing set, where the infarct core is seen more clearly, taken within 3 h of the initial CT scan	Variable	Predict the infarction core based on CTP imaging from acute ischemic stroke patients, using DWI-based manual segmentation from MRI acquired shortly after
ISLES (Ischemic Stroke Lesion Segmentation) 2022	Stroke Lesions Segmentation	400 (250 for training, 150 for testing)	1) University Hospital of the Technical University Munich, Munich, Germany; 2) University Hospital of Bern, Bern, Switzerland; 3) University Medical Center Hamburg-Eppendorf, Hamburg, Germany.	MRI	multimodal MRI data (DWI, ADC, FLAIR)	Not mentioned	FLAIR: (0.2x0.23-1x1 mm, DWI: 0.88x0.88-2x2 mm	Data were converted in NIfTI formats and brain-extraction was performed mask-based using HD-BET algorithm; FLAIR images were coregistered to the corrisponding DWI, and DWI and their corrisponding ADC maps were skull-stripped and resliced to an axial isotropic voxel size of 2x2 mm^3.	A 3D Unet was trained on DWI data that was pre-annotated for other research projects by using a single-MRI modality. This algorithm was trained with labeled data from University Hospital of Munich. Segmentations were then checked by medical students with special stroke lesion segmentation training, and their annotations were revised by a neuroradiology resident. After a new pre-segmentation algorithm was trained on correct annotations, and the process was repeated.	Variable	It targets not only delineation of large infarct lesions but also of multiple embolic and/or cortical infarcts; and evaluates both pre- and post-interventional MRI images.

Table 3.2: Most used datasets and related characteristics in the field of stroke lesion segmentation

3.3 Other datasets

Before going into details about the datasets used during this study, it is worth analyzing also the largest and most used datasets in the segmentation of other types of medical images. These datasets were met when analyzing certain types of deep learning algorithms really efficient in other fields, and consequently they're reported to get to know the most advanced datasets for these scopes. Detailed information is reported in *Table 3.3.*

Liver Tumor Segmentation (LiTS) challenge was organized in 2017 in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI) and International Conference On Medical Image Computing & Computer Assisted Intervention (MICCAI). The challenge's task is the segmentation of patients' liver lesions (primary and secondary tumors, metastases).

LiTS dataset consists of 201 (131 for training, 70 for testing) CT volumes of patients with hepatocellular carcinoma (HCC) provided by seven hospitals and research institutions worldwide. CT scans are used to study livers' anomalies, which can be important biomarkers for diagnosis of primary and secondary hepatic tumors, and they show different types of tumor contrast levels (hyper- /hypo-intense), shape and size, making it very difficult for intensity-based methods to generalize well on unseen test data (Bilic et al., 2019). Data have been manually annotated by trained radiologists and oncologists, and verified by three experienced radiologists.

The first HEad and neCK (H&N) TumOR (HECKTOR) challenge was organized as a satellite event of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2020. The task of the competition was the automatic segmentation of head and neck primary Gross Tumor Volumes in the oropharynx region, from a dataset of 254 (201 for training, 53 for testing) FDG-PET/CT images (which include complementary information about cancerous lesions) acquired by five institutions. The task is challenging due to the variation in image acquisition across centers and the presence of lymph nodes with high metabolic response in PET images (Andrearczyk et al., n.d.). Images weren't preprocessed, but pieces of code to load, crop, resample the data and train a baseline CNN were provided. Data annotations were generated by expert radiation oncologists and later modified by a VOI (Volume Of Interest) quality control, supervised by an expert who is both radiologist and nuclear medicine physician.

HECKTOR 2021 dataset consists of 325 (224 for training, 101 for testing) FDG-PET/CT scans provided by six centers worldwide. This year's challenge was composed by 3 tasks: 1) the automated segmentation of H&N primary Gross Tumor Volumes; 2) the automated prediction

of Progression Free Survival (PFS); and 3) the same as task 2, but with ground truth annotations provided to participants. Initial data annotations were made by expert oncologists and later re-annotated; while data added with respect to 2020 challenge (from CHUP center) were segmented with a Fuzzy Locally Adaptive Bayesian (FLAB) segmentation and corrected by an expert oncologist, while re-annotations were drawn by three experts with the MIM software (Andrearczyk et al., 2022).

Dataset	Aim	Size	Centers providing Data	Acquisition Protocol	Modalities	Volume Size	Resolution	Preprocessing	Annotations	Heterogeneity	N. Classes	Segmentation task
MICCAI Liver Tumor Segmentation (LITS)	Liver Tumor Segmentation	201 (131 for training, 70 for testing)	1) Ludwig Maximilian University of Munich (Munich MC), 2) Radboud University Medical Center of Nijmegen (Netherlands NL), 3) Polytechnique & CHUM Research Center Montréal (Canada CA), 4) Tel Aviv University (Israel ISR), 5) Sheba Medical Center (Israel ISR), 6) IRCAD Institute Strasbourg (France FR), 7) Hebrew University of Jerusalem (Isreale, ISR)	CT scans	Different CT scanners and acquisition protocols	Number of slices in 2 ranges from 42 to 1026	varying in-plane resolution from 0.55 mm to 1.0 mm	/	Manually annotated with two labels (liver and lesion) by a trained radiologists or oncologists; the quality of the test segmentations was verified by three experienced radiologists in an blinded review	/	1	primary tumors, secondary tumors and metastasis
MICCAI HEad and neCK TumOR (HECKTOR) challenge 2020	Head and Neck Tumor Segmentation	254 (201 for training, 53 for testing)	1) HGJ: Hôpital Général Juif, Montréal (Canada CA), 2) CHUS: Centre Hospitalier Universitaire de Sherbooke, Sherbrooke (Canada CA), 3) HMR: Hôpital Maisonneuve-Rosemont, Montréal (Canada CA), 4) CHUM: Centre Hospitalier de l'Université de Montréal, Montréal (Canada CA), 5) CHUV: Centre Hospitalier Universitaire Vaudois (Switzerland CH)	CT image, PET image	For the PET portion of the FDG-PET/CT scan, image acquisition was performed using multiple bed positions. Attenuation corrected images were reconstructed using OSEM (or LOR-RAMLA, depending on the center) iterative algorithm. For the CT portion, an energy of 120/140 kVp with an exposure of 11/12/210/350 mAs was used (depending on the center)	Not mentioned	Median in-plane resolution was 3.52 x 3.52 mm^2 or 4x4 mm^2 (depending on the center) (PET), 0.98x0.98 mm^2 (CT)	No Preprocessing applied but provided various pieces of code to load, crop, resample the data	Initial annotations, i.e. 3D contours of the GTVt, were made by expert radiation oncologists and were later modified by a VOI quality control; supervised by an expert who is both radiologist and nuclear medicine physician, while two non experts made an initial cleaning in order to facilitate the expert's work.	Variation in image acquisition and quality across centers and presence of lymph nodes with high metabolic responses in the PET images	1	Primary Gross Tumor Volume (GTVt) segmentation of H&N tumor in the oropharynx region
MICCAI HEad and neCK TumOR (HECKTOR) challenge 2021	Head and Neck Tumor Segmentation	325 (224 for training, 101 for testing)	1) HGJ: Hôpital Général Juif, Montréal (Canada CA), 2) CHUS: Centre Hospitalier Universitaire de Sherbooke, Sherbrooke (Canada CA), 3) HMR: Hôpital Maisonneuve-Rosemont, Montréal (Canada CA), 4) CHUM: Centre Hospitalier de l'Université de Montréal, Montréal (Canada CA), 5) CHUV: Centre Hospitalier Universitaire Vaudois (Switzerland CH), 6) CHUP: Centre Hospitalier Universitaire Poitiers (France FR)	CT image, PET image	For the PET portion of the FDG-PET/CT scan, image acquisition was performed using multiple bed positions. Attenuation corrected images were reconstructed using OSEM (or LOR-RAMLA, depending on the center) iterative algorithm. For the CT portion, an energy of 120/140 kVp with an exposure of 11/12/210/350 mAs was used (depending on the center)	Not mentioned	Median in-plane resolution was 3.52 x 3.52 mm^2 or 4x4 mm^2 (depending on the center) (PET), 0.98x0.98 mm^2 or 1.17x1.17 mm^2 or 0.853x0.853 mm^2 (depending on the center) (CT)	No Preprocessing applied but provided various pieces of code to load, crop, resample the data	Initial annotations, i.e. 3D contours of the GTVt, were made by expert radiation oncologists and were later re-annotated. The delineation from the CHUP center were obtained semi-automatically with a Fuzzy Locally Adaptive Bayesian (FLAB) segmentation and corrected by an expert radiologist. Two non-experts made an initial cleaning in order to facilitate the expert's work. The expert either validated or edited the VOIs. For the CHUP center the re-annotation were performed by three experts: one nuclear medicine physician, one radiation oncologist and one who is both radiologist and nuclear medicine physician	Variation in image acquisition and quality across centers and presence of lymph nodes with high metabolic responses in the PET images	1	Primary Gross Tumor Volume (GTVt) segmentation of H&N tumor in the oropharynx region

Table 3.3: Most used and advanced datasets and related characteristics in the segmentation of different biomedical images

3.4 FeTS 2022

Among all the existing datasets used for brain tumor segmentation, FeTS 2022 represents the most advanced and complete one, not only because it contains BraTS 2021 data, which represents nowadays the most complete and largest dataset about gliomas, but also because it is based on federated learning.

3.4.1 Description of FeTS Dataset

In order to analyze properly BraTS data, it was necessary to study the characteristics of data provided by the donating cohorts, to understand possible differences between them. After contacting directly BraTS 2020 challenge official email, it was possible to discover that this information is not released by BraTS challenge, but only by FeTS, a challenge created expressly for this purpose.

FeTS 2021 was the first challenge introduced based on federated learning, a machine learning branch that carries out the training of an algorithm through a multitude of decentralized servers with local data, without merging those data but keeping them locally. For this reason, it allows to create a strong and effective machine learning model without sharing data, and so solving problems of privacy, data security, data access rights and so on. Moreover, it also allows to study properties of data coming from different centers, without pooling their data together.

Specifically, FeTS 2022 data are multi-parametric MRI (mpMRI) provided by a total of 23 institutions worldwide: some of them come from BraTS 2021 challenge, while others from various remote independent institutions included in the collaborative network of a real-world federation (The Federated Tumor Segmentation (FeTS) Challenge 2022: Structured Description of the Challenge Design Mission, n.d.). The dataset is provided with the addition of non-imaging data including information about their partitioning based on the originating institute (*partitioning_1.csv*) and also it was made an extra partitioning based on the median tumor size in the 5 largest institutions (*partitioning_2.csv*).

It was possible to access FeTS 2022 data by enrolling for the challenge, without uploading any method but just downloading the dataset for analysis' purpose.

Only training data were used; they were realized in the 8th April 2022 and downloaded in the 31st May 2022.

3.4.2 Intensities distribution analysis

To better understand the characteristics and peculiarities of FeTS 2022 dataset, it was performed an intensity distribution analysis on images provided by the different centers. As a first step, based on the information provided inside *partitioning_1.csv,* it was realized an histogram on the number of images supplied (*Figure 3.2*). Unfortunately, the partitioning ID is just a number identifying the providing cohort, therefore it was not possible to obtain any detail about the name or location of those institutions, but only to understand that there was a total of 23 different centers.

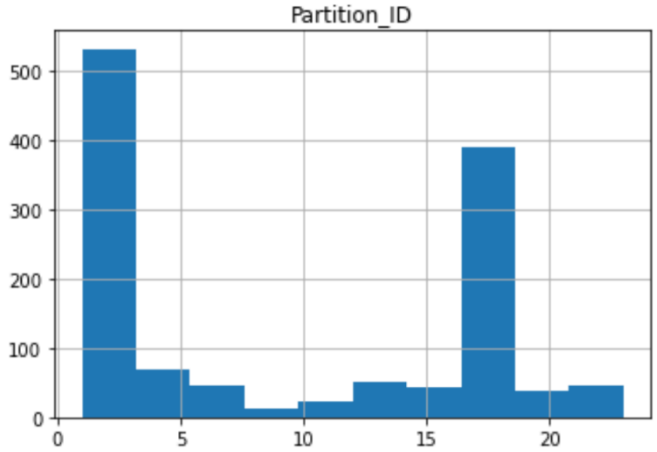
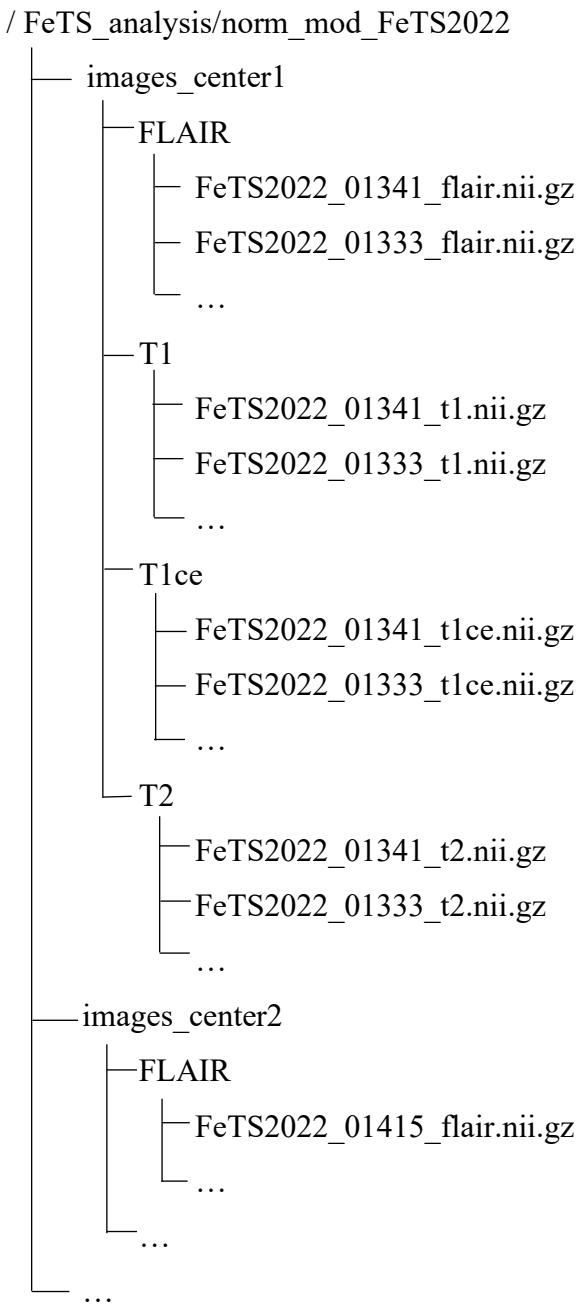


Figure 3.2: Histogram describing the number of images provided by the 23 different institutions contributing to FeTS 2022 dataset

From *Figure 3.2* it is evident that center 1 and 18 are the ones providing the largest number of images (511 the former, 382 the latter, out of a total of 1666 images), with the first institution providing more or less one third of the total number of images, and thus being the most representative one. Before proceeding, it must be specified that each "image", as stated before, is in reality a folder containing 4+1 scans, one for each MR modality provided (T1, T1ce, T2, FLAIR) and a segmented image containing the ground truth labels, all associated to the same subject and so, obviously, coregistered. For this reason, excluding the annotated images, the real number of scans inside FeTS 2022 dataset is 6664.

After that, different folders were created and images were divided based on the providing center, and inside it based on the different modalities, as described by the underlying schema.



Before proceeding with the study, some random images and the corresponding intensities were analyzed, observing that the range of values assumed by images provided by different centers changes a lot. To make results and histograms more comparable, it was decided to normalize all images' values between 0 and 1, creating new folders which follow the schema previously introduced.

As stated before, the intensity distribution analysis consists in the realization of four histograms per center, one for each modality, in which, as a matter of fact, each histogram represents the distribution of intensities (values) of images of that specific center. Four images were then created (one for each modality), each comparing the distribution of images' values of all

centers, for the same modality. Given that the number of images provided is highly imbalanced, to select the number of images to be used for the realization of such histograms it was chosen to use thresholds based on the specific number of images supplied by each center: if the images are more than 100, it was decided that it was enough to select 10% of them; while if they are less than 100, then the threshold used was 50%. At this point, the selected percentage of images taken randomly were concatenated for each modality, and the corresponding histograms were calculated.

As last step, values equal to zero were discarded from the histograms, as they are associated with the background and so they represent the majority while being non informative.

For histograms visualization, it is chosen to set *density=True* so that the area under each histogram would sum up to 1, allowing histograms' normalization and so a better comparison between them. Moreover, to increase the differentiation between histograms, it is used the '*rainbow*' Matplotlib palette, extracting 23 random colors (one for each histogram) for each modality (*Figure 3.3, Figure 3.4, Figure 3.5, Figure 3.6*).

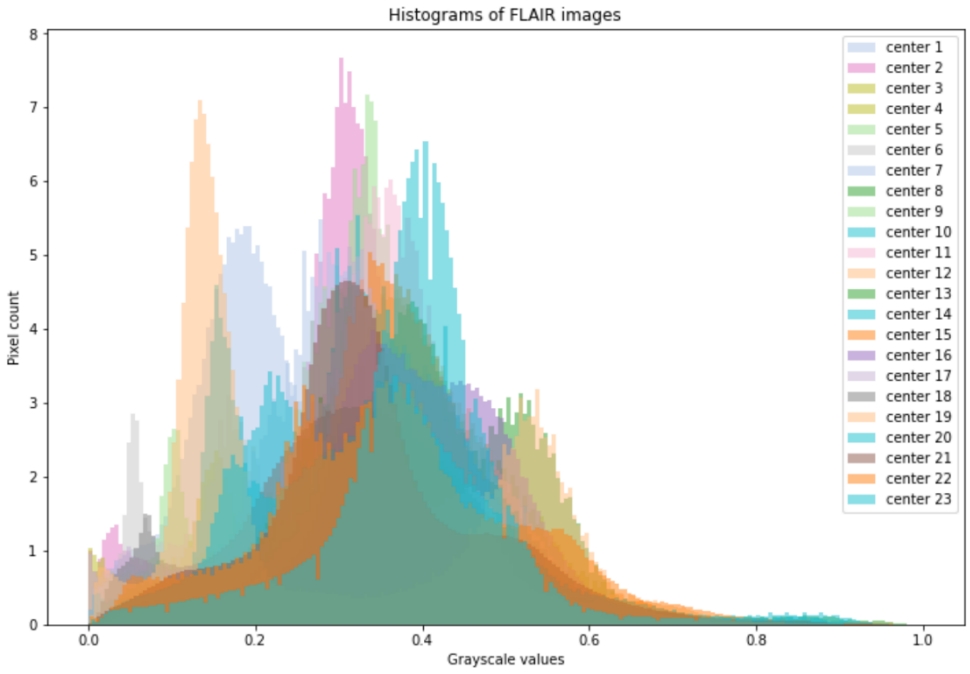


Figure 3.3: Average histograms of intensities distribution of FLAIR images, for the 23 providing centers

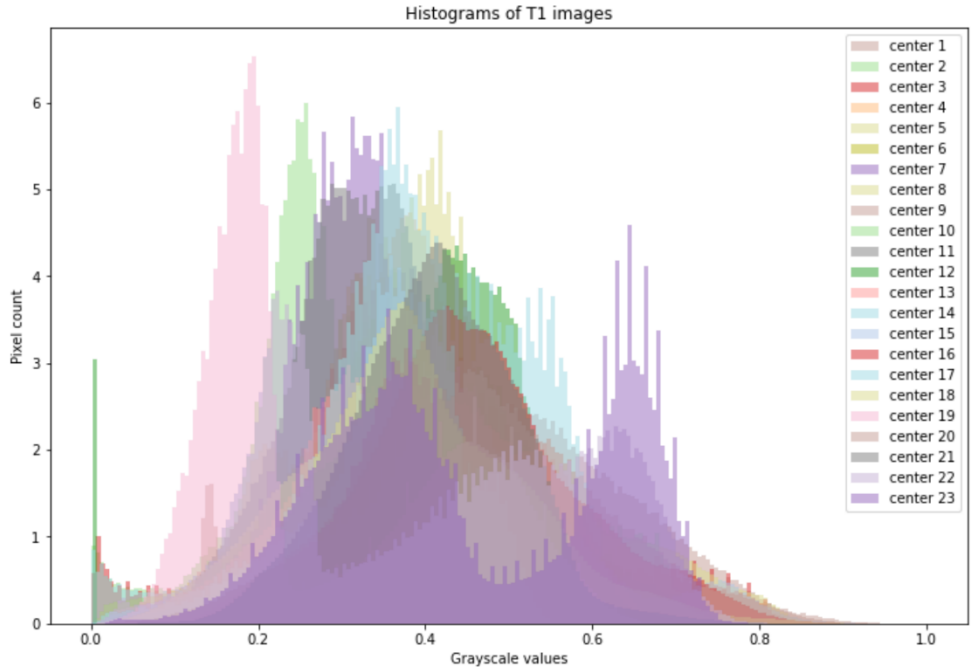


Figure 3.4: Average histograms of intensities distribution of T1 images, for the 23 providing centers

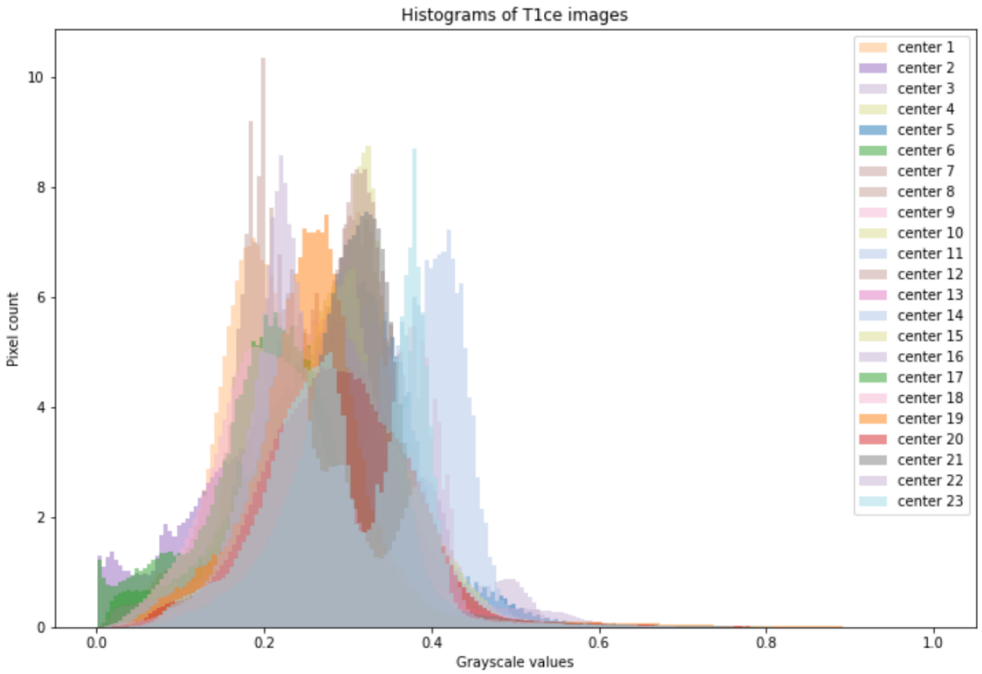


Figure 3.5: Average histograms of intensities distribution of T1ce images, for the 23 providing centers

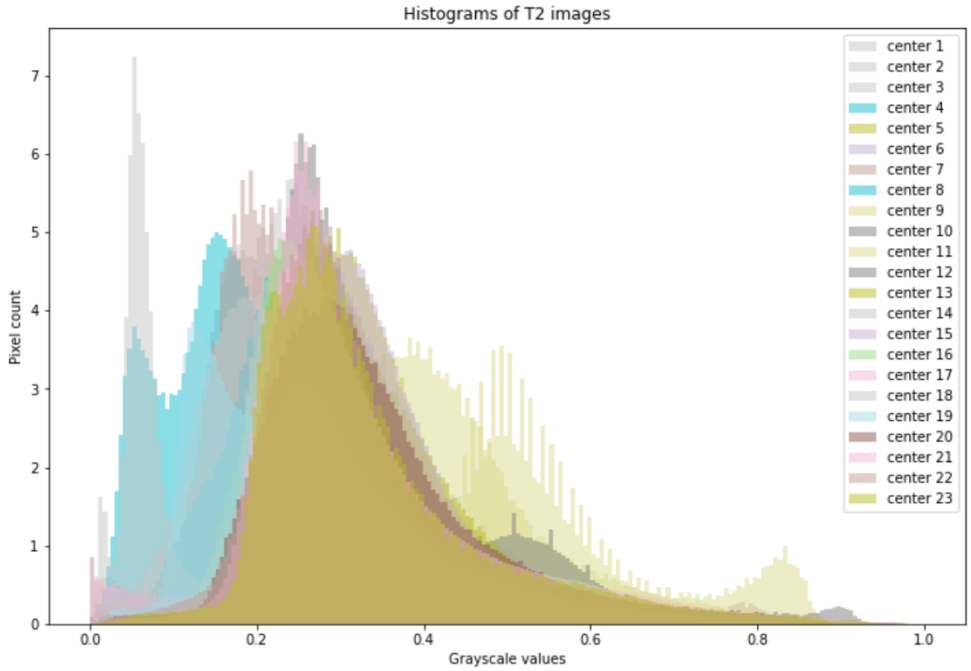


Figure 3.6: Average histograms of intensities distribution of T2 images, for the 23 providing centers

4. Methods

This study was performed exploiting the computing power of a machine in the Padova Neuroscience Center (PNC), which has access to two powerful NVIDIA Tesla V100-PCIE-16GB. All the analysis was done working on Jupyter Lab, with Python 3.9 and PyTorch 1.11.0.

4.1 Explorative analysis of the used datasets

Before proceeding with the primary analysis of this thesis, a preliminary examination of the utilized datasets was performed, to understand the diversity of the exploited images.
The descriptive statistics technique that was chosen to be used for this research is the Box Plot, which is a method for visualizing the locality, spread and dispersion of data through quartiles, providing information on the variability of the considered dataset, indicating how the values of images are spread out. The term "box-plot" derives from its structure: it is composed by a rectangle (box), in which the top represents the third quartile (Q_3 or 75th percentile, which is the minimum value below which it falls the 75% of values of the dataset), a horizontal line inside the box represents the median (Q_2) and the bottom of the rectangle indicates the first quartile (Q_1 or 25th percentile, which represents the minimum value below which it falls the 25% of values). The difference between the 25th and 75th percentiles, which consists also in the height of the box, is called Interquartile Range (IQR). Two vertical lines, called "whiskers", extend then from the top and the bottom of the rectangle, indicating the maximum and minimum values of the dataset: the maximum value is obtained by adding the interquartile range to the third quartile, while the minimum is computed by subtracting the interquartile range from the first quartile. In a normally distributed dataset, 99.3% of data stand inside these whiskers, while for all the other cases, in which the assumption of Gaussian distribution can't be done (like in this analysis), it can be concluded anyway that the majority of data stay between those vertical lines. All the other values, which are not included inside the box or whiskers, are data points which are really far away from most of the others; they consist in outliers and are represented as small circles.
The Box Plots have been derived for each of the analyzed datasets (BraTS 2020, FeTS 2022, ISLES 2022), considering a defined number of representative images. In particular, for each

image it was computed a mask for all the labels present in the segmentation map (1,2,4 for brain tumors images, 1 for stroke lesion) and multiplied for the corresponding images of the different modalities. In this way, from a single picture it was possible to obtain one image for each class, containing only the voxels belonging to the corresponding brain tumor subregion or to the stroke lesion. After that, Box Plots were computed for images of the same class and of the same modalities, and were compared between each other.

4.1.1 BraTS 2020 dataset

In the analysis of the BraTS 2020 dataset, 5 subjects were randomly extracted from the training set. For each subject, the previously explained procedure was repeated for all the present modalities (FLAIR, T1, T1ce, T2): a mask for each class (ED, NCR, ET) was obtained and multiplied for the corresponding image. For each tumoral subregion, from the 5 masks of the same modality 1D arrays was obtained by selecting only values different from 0: much of the images were indeed composed by zeros, so that the resulting box plots were highly unbalanced. For these reasons it was chosen to discard those values which weren't informative for the study. After that, noting that images of different subjects had different ranges of values, to make the box plots of different subjects, for each modality, more comparable, they were also normalized to the maximum value reached by the considered images, and the corresponding box plots were obtained.

Given that no information was included inside the BraTS dataset about which centers provided the different images, it was not possible to compare Box Plots of images of the same modalities between different centers. The box plots of the 5 images, and their corresponding normalized box plots for the first class (ED), and for the different available modalities (FLAIR, T1, T1ce, T2), are showed respectively in *Figures 4.1, 4.2, 4.3, 4.4, 4.5, 4.6*, *4.7* and *4.8,* while the related ones for the second class (NCR) are showed in *Figures 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15* and *4.16;* and finally for last class (ET) in *Figures 4.17, 4.18, 4.19, 4.20, 4.21, 4.22, 4.23* and *4.24.*

It can be observed that there are some subjects in which the distribution of values is more similar to others, while, in other cases, the dispersion of values of images of different subjects acquired with the same modality is completely different, having a variable number of outliers, a variable size of the box and of the whiskers, and a different median value.

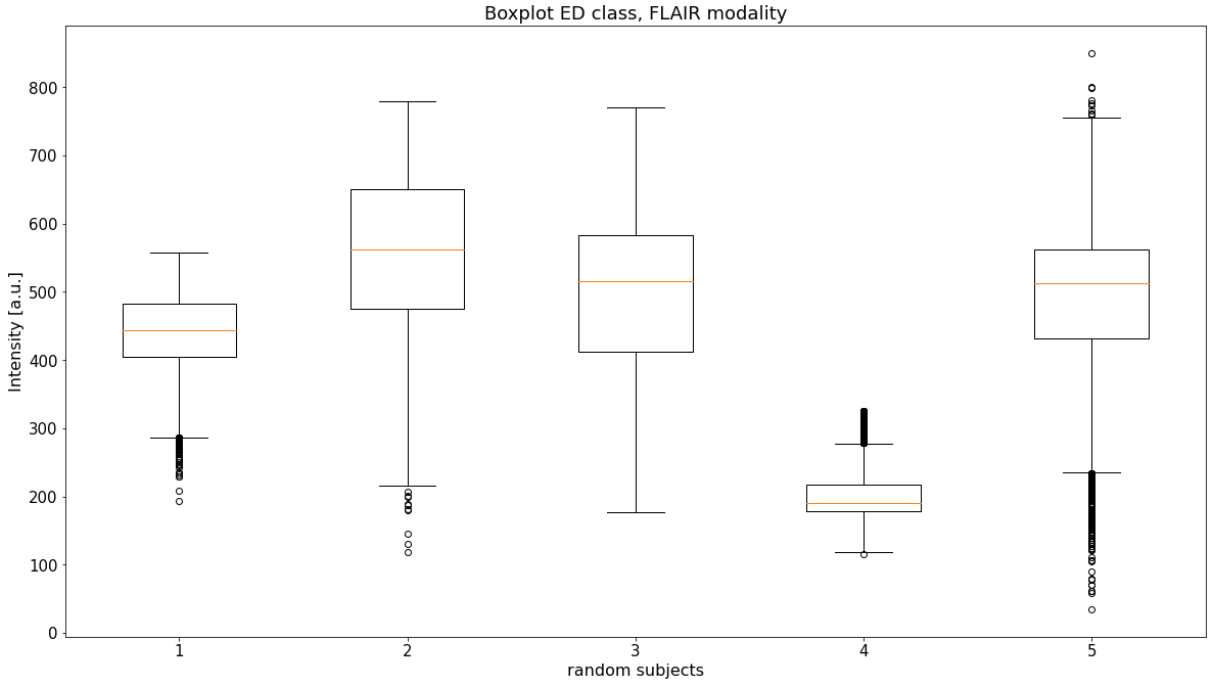


Figure 4.1: Box Plots of FLAIR images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

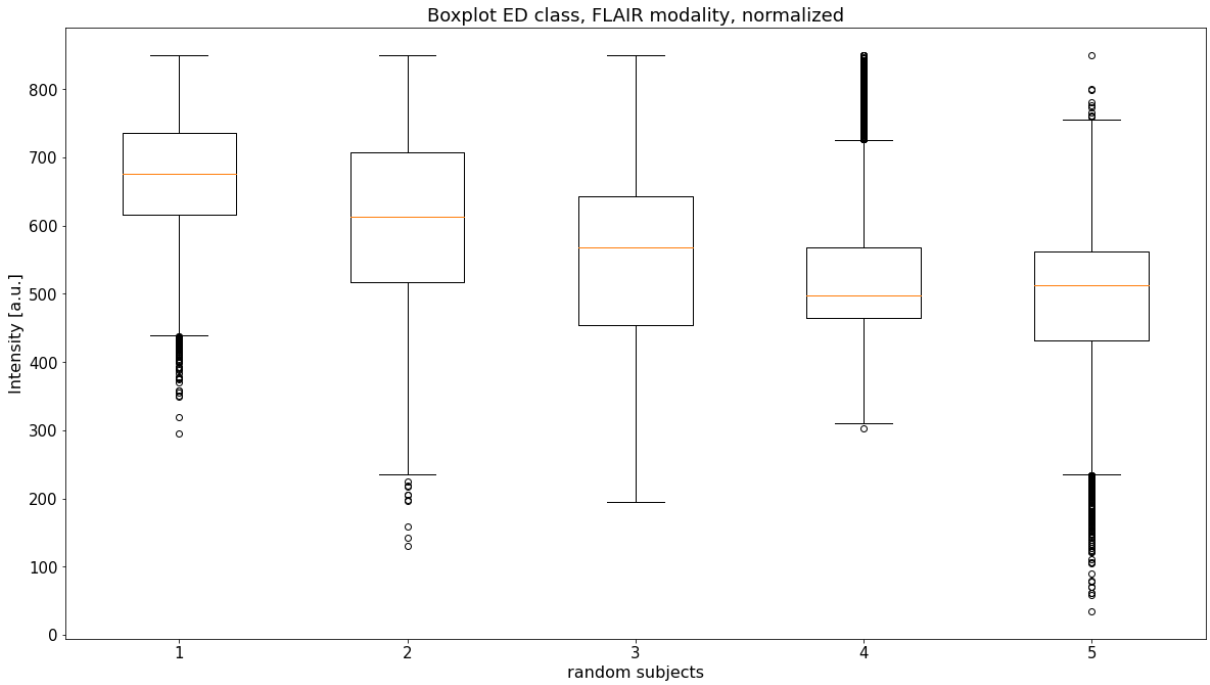


Figure 4.2: Box Plots of FLAIR images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

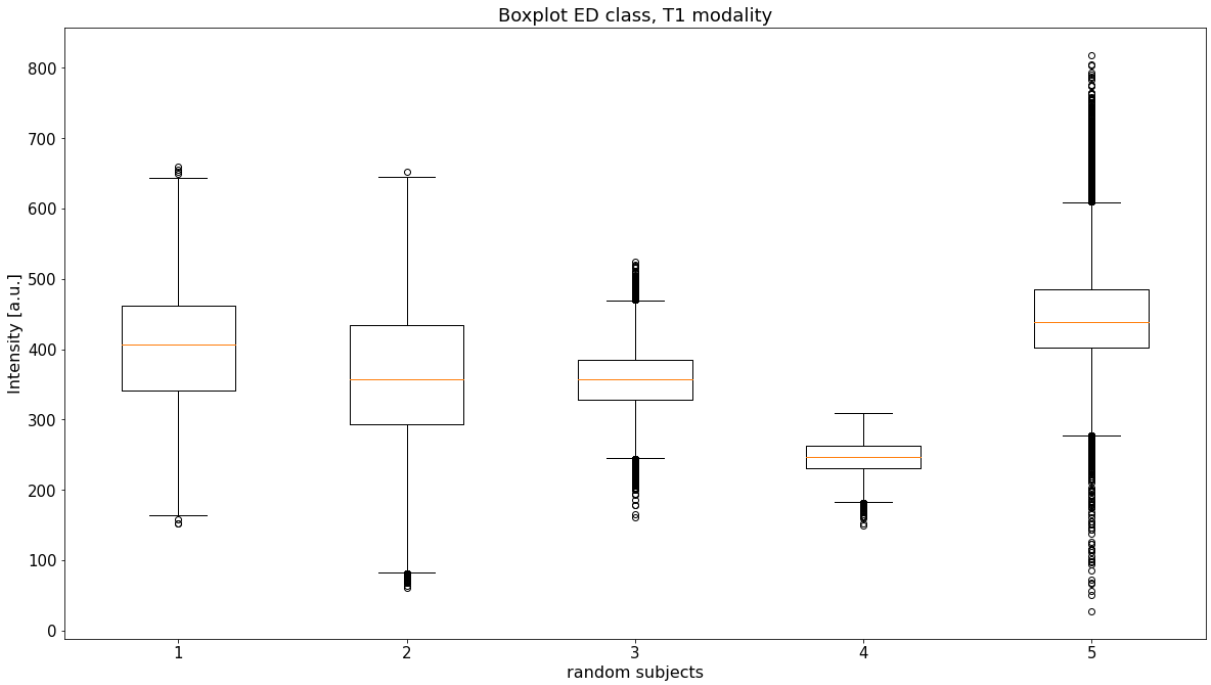


Figure 4.3: Box Plots of T1 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

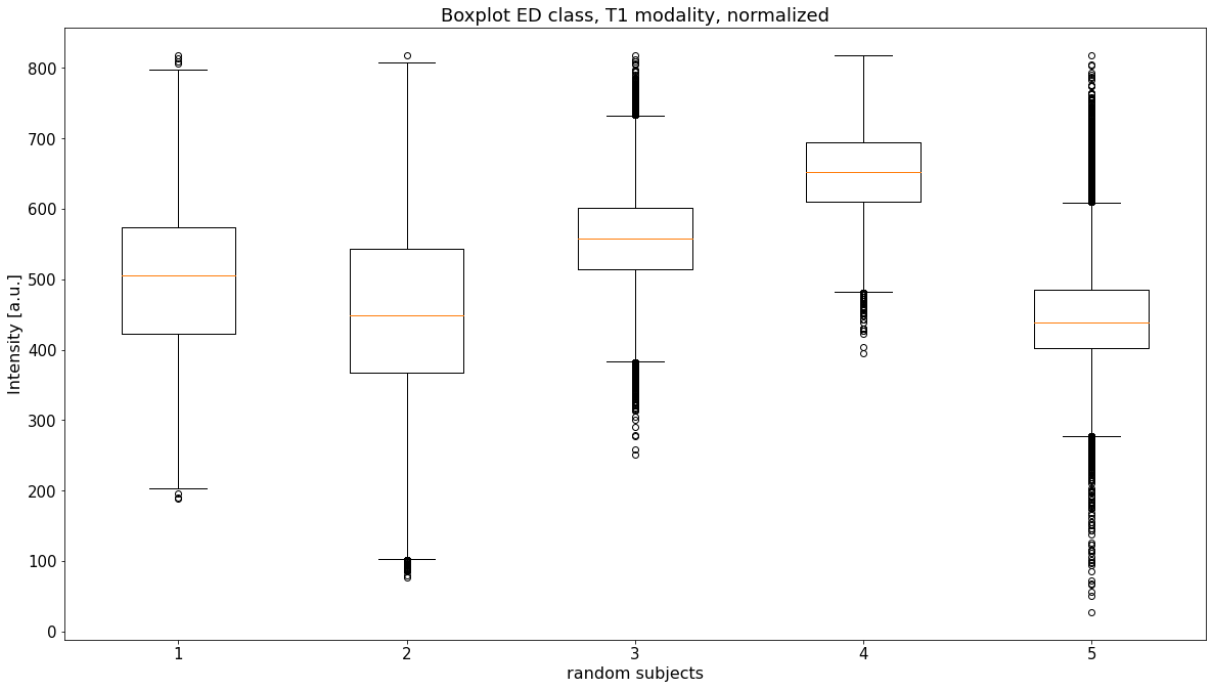


Figure 4.4: Box Plots of T1 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

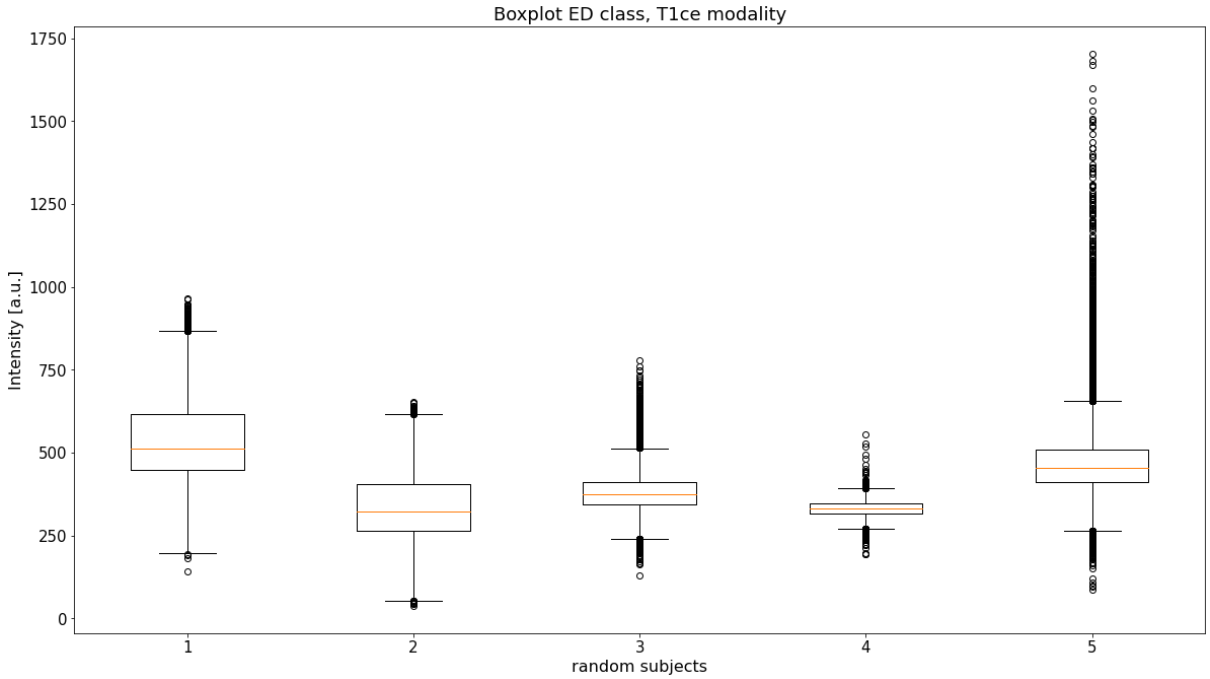


Figure 4.5: Box Plots of T1ce images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

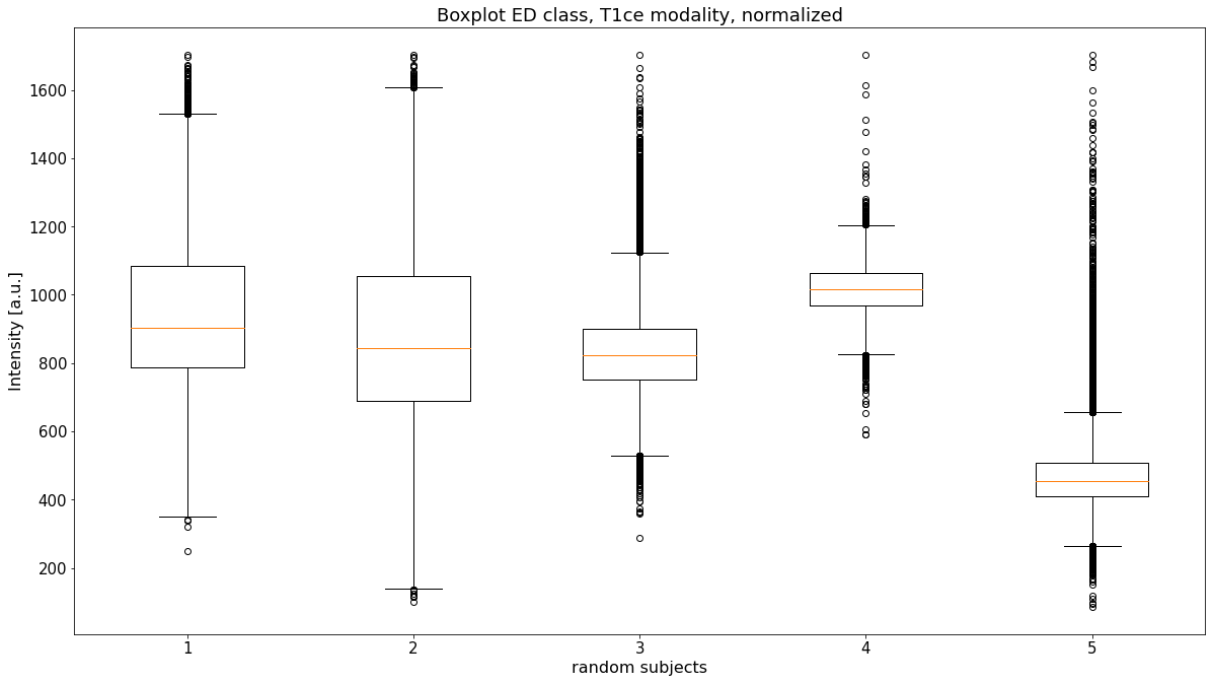


Figure 4.6: Box Plots of T1ce images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

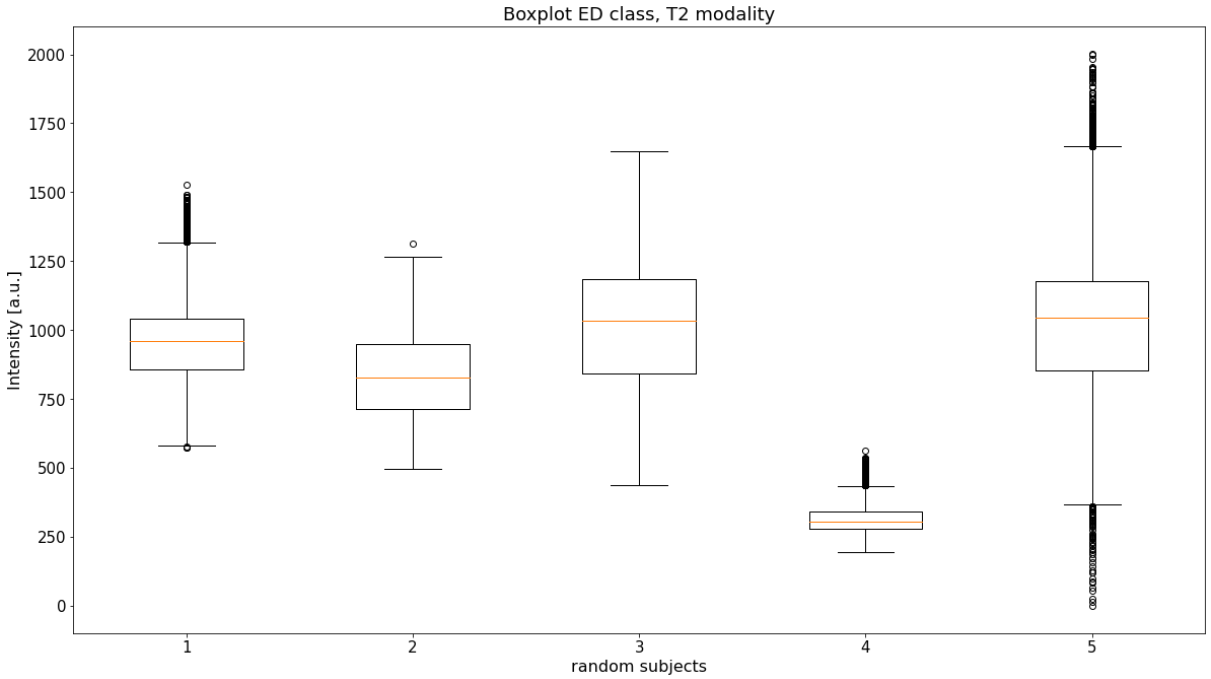


Figure 4.7: Box Plots of T2 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

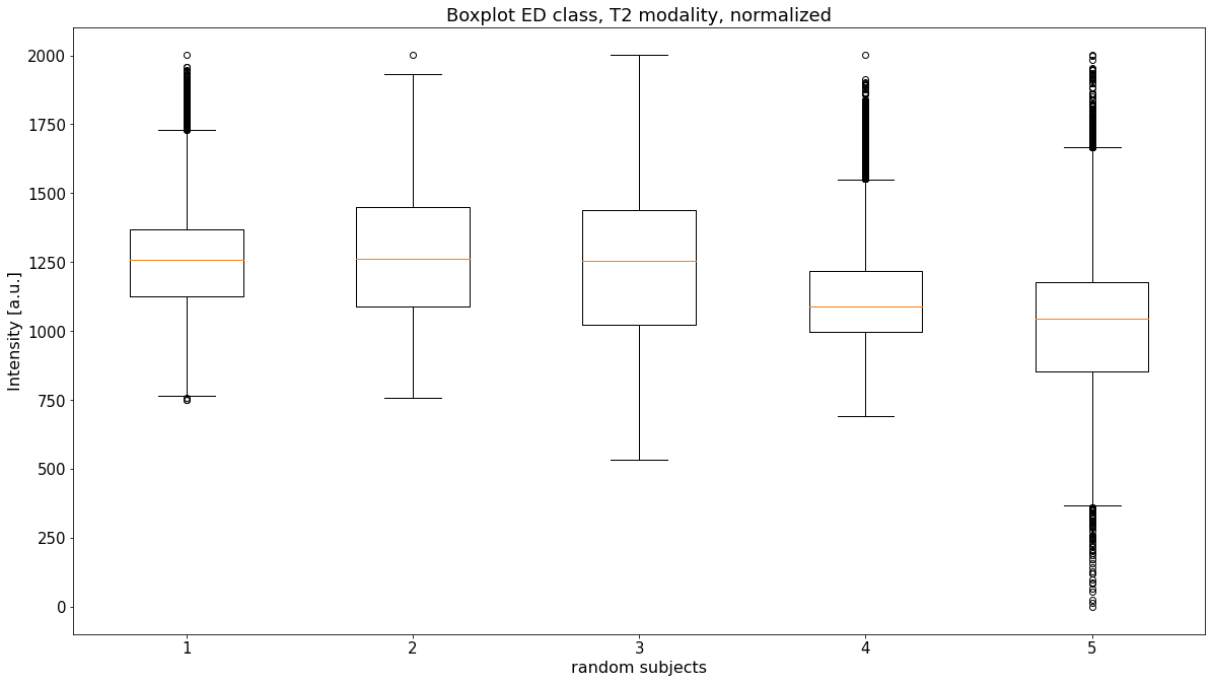


Figure 4.8: Box Plots of T2 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

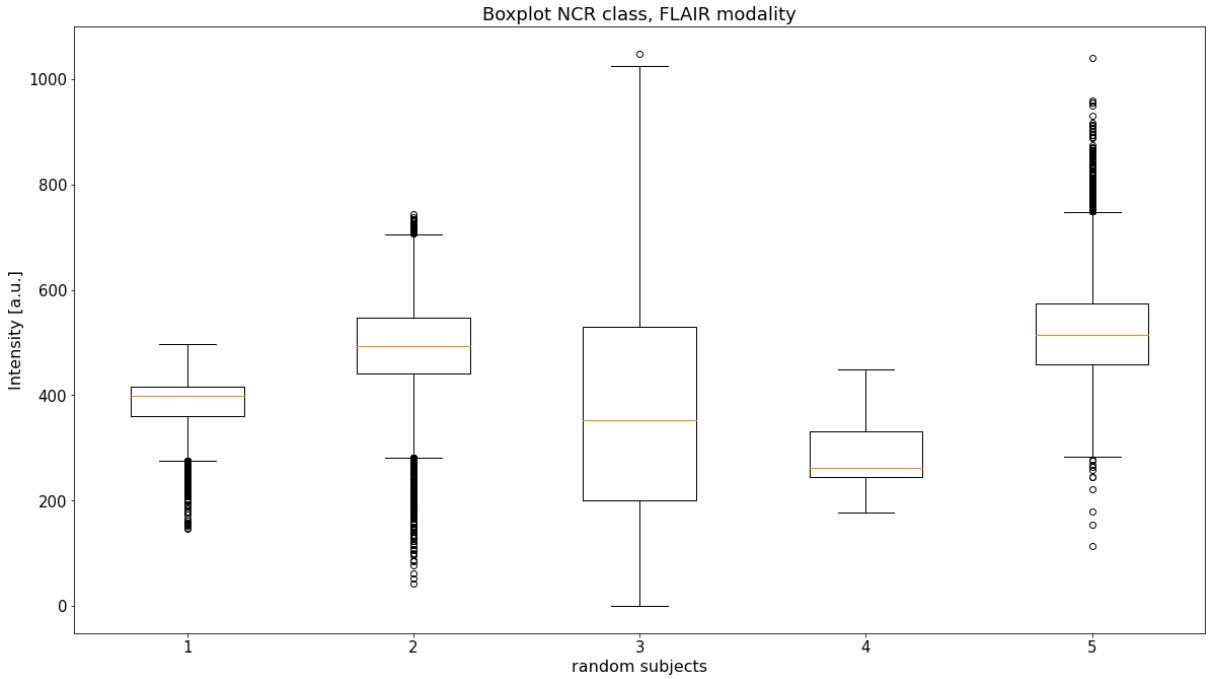


Figure 4.9: Box Plots of FLAIR images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

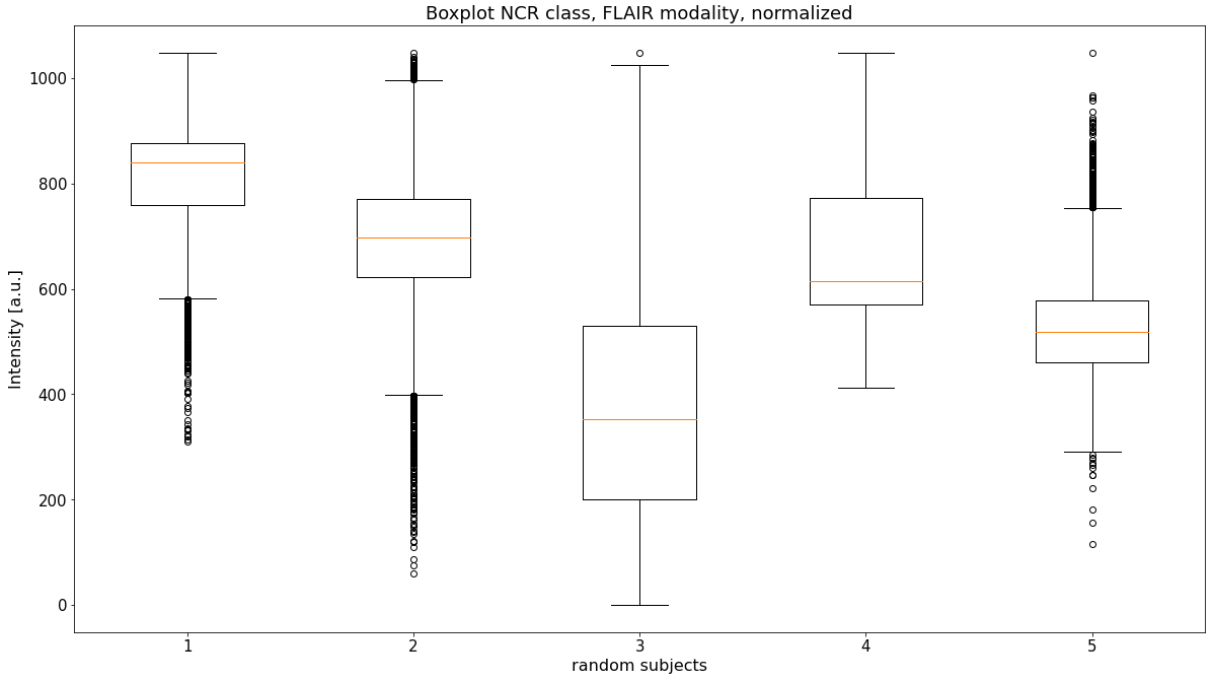


Figure 4.10: Box Plots of FLAIR images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 3).

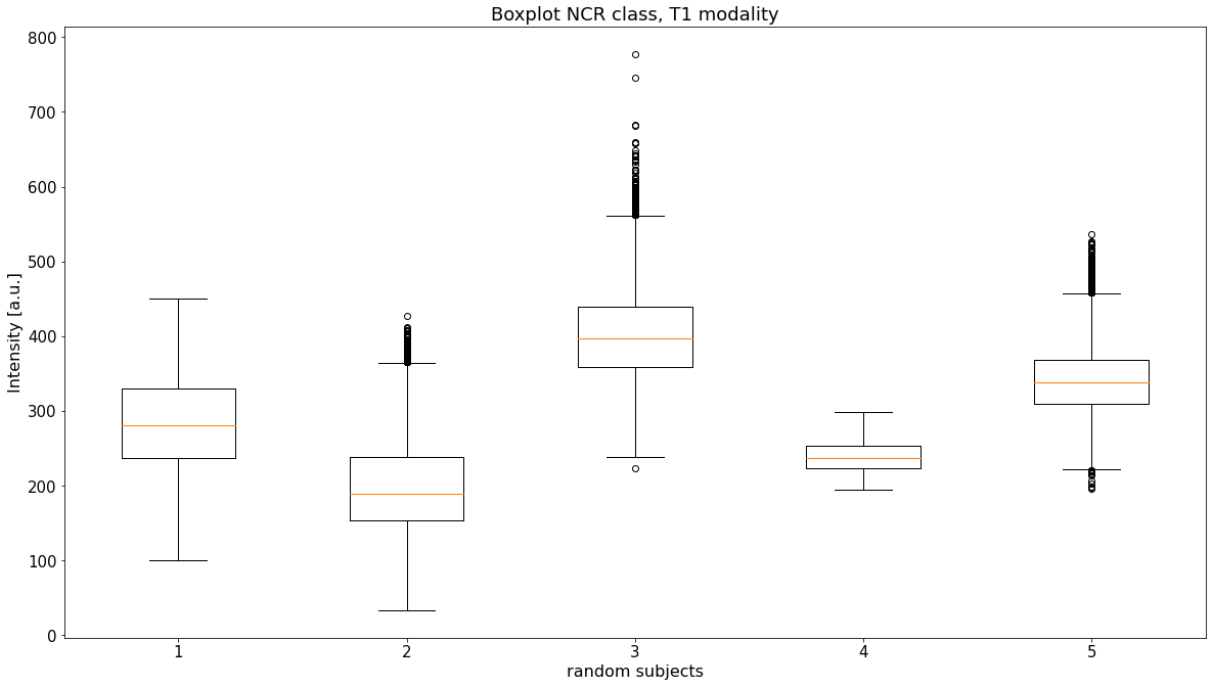


Figure 4.11: Box Plots of T1 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

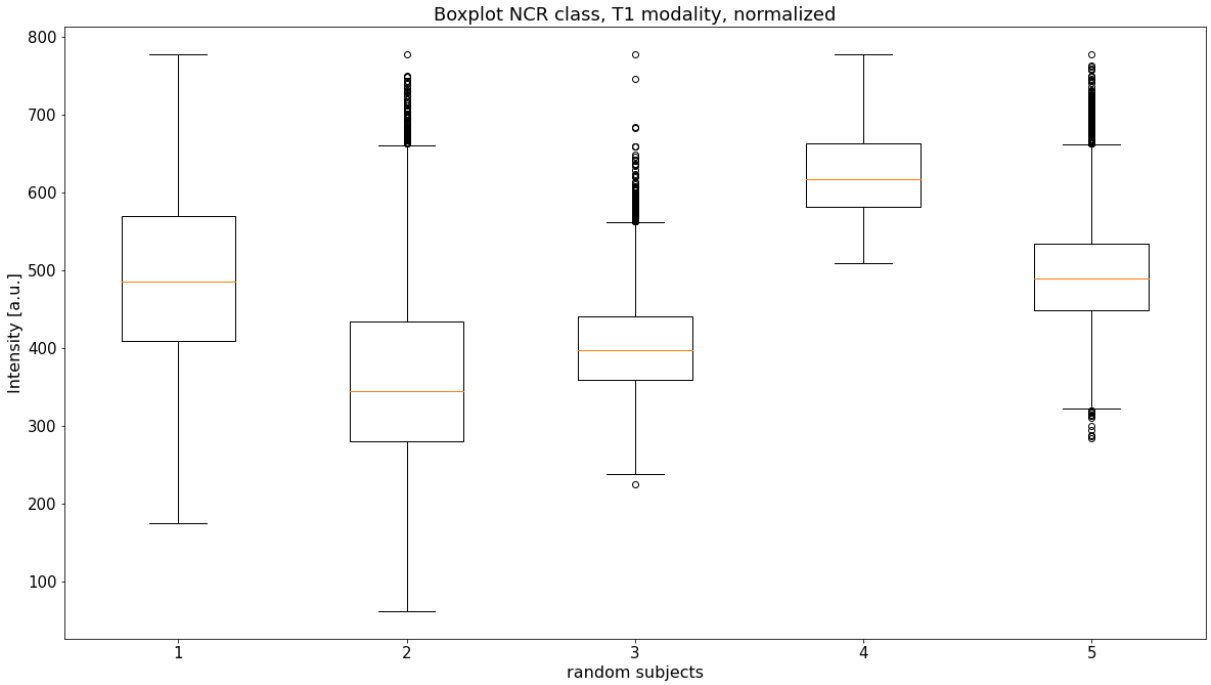


Figure 4.12: Box Plots of T1 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 3).

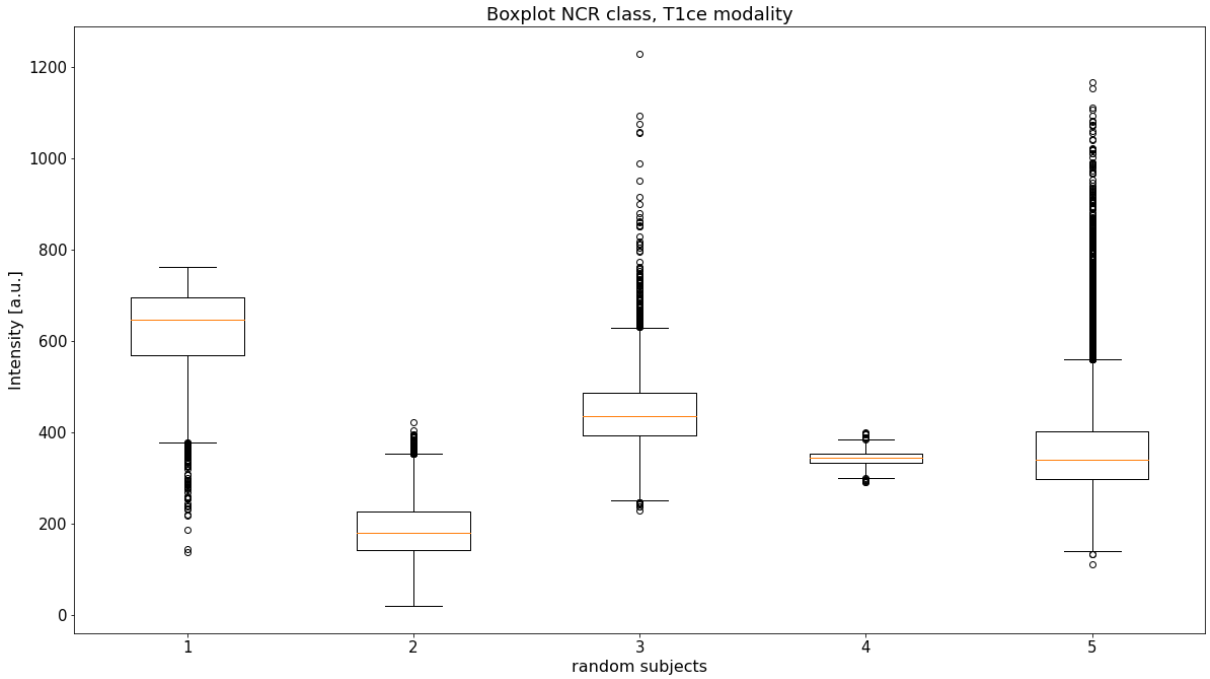


Figure 4.13: Box Plots of T1ce images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

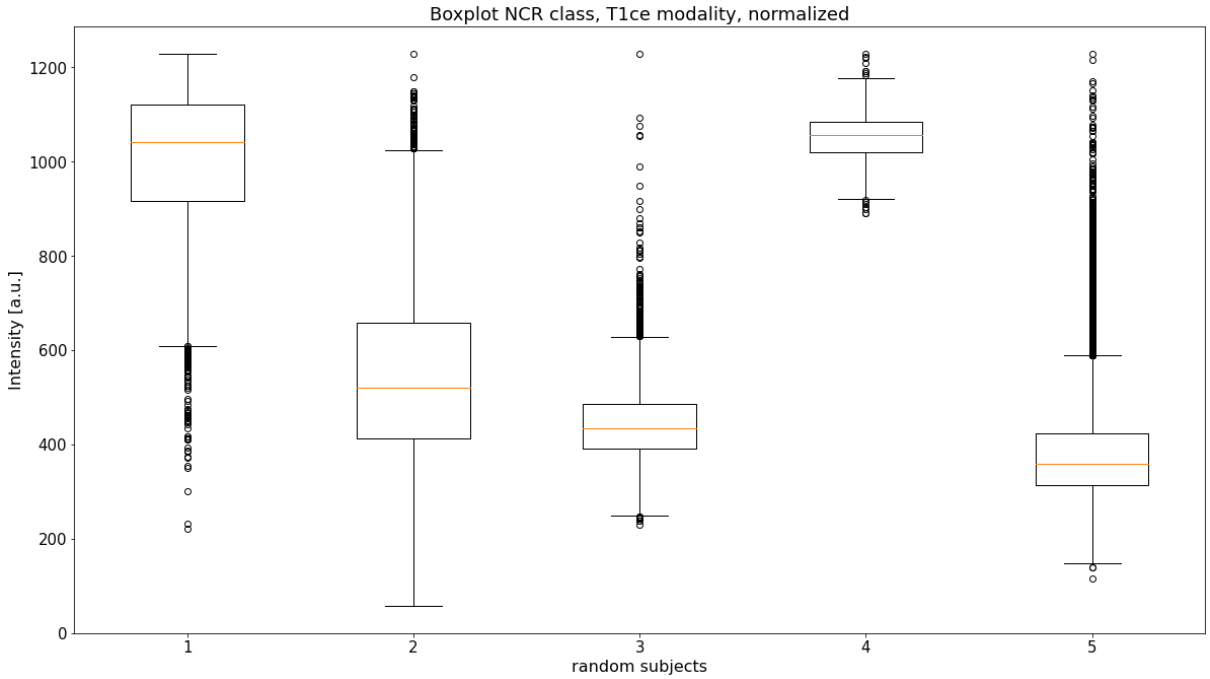


Figure 4.14: Box Plots of T1ce images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 3).

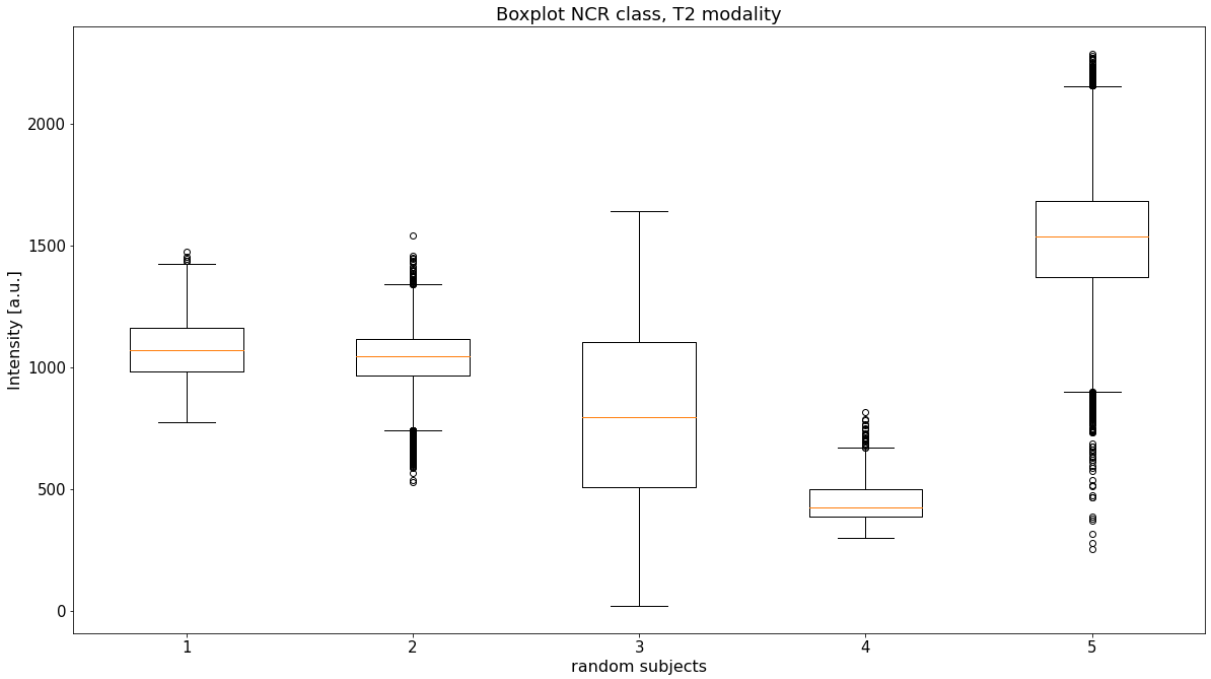


Figure 4.15: Box Plots of T2 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

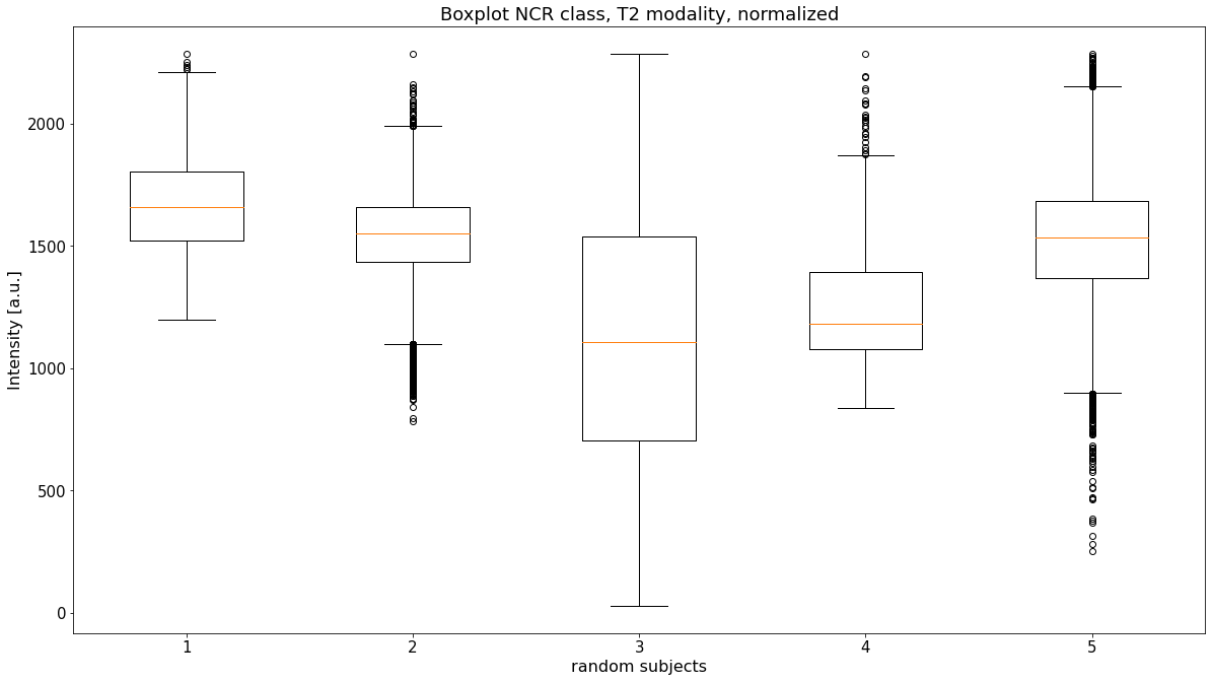


Figure 4.16: Box Plots of T2 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

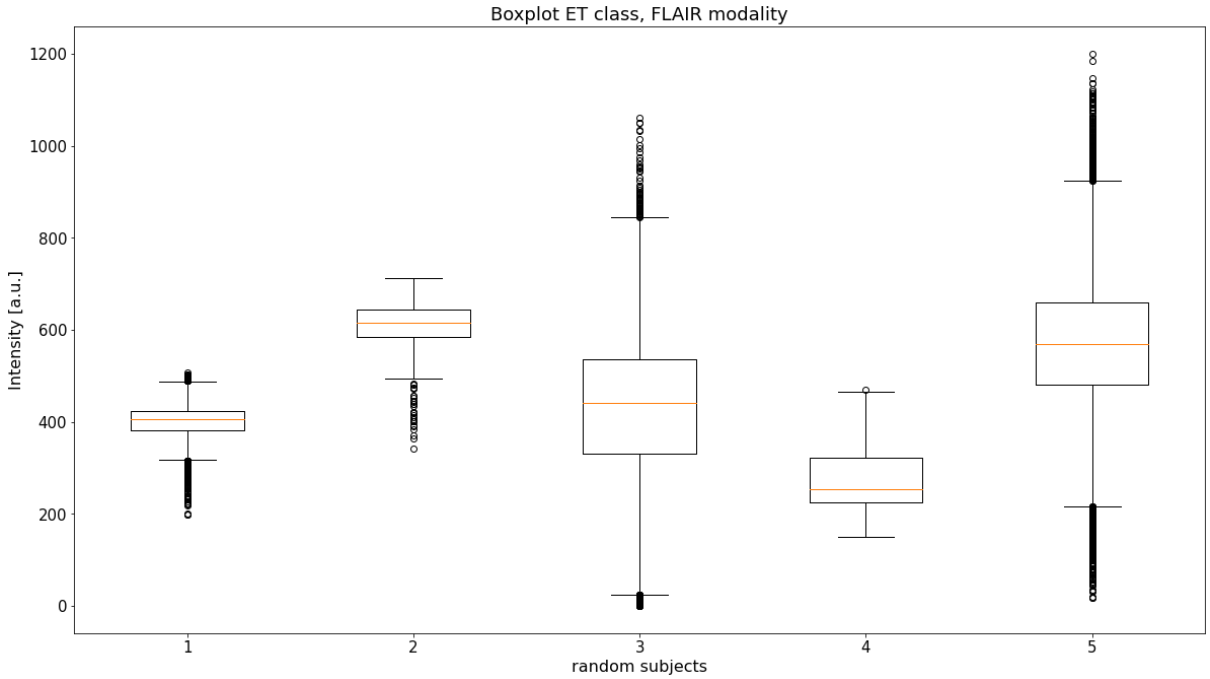


Figure 4.17: Box Plots of FLAIR images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the third class (ET: enhancing tumor) and having removed zero values.

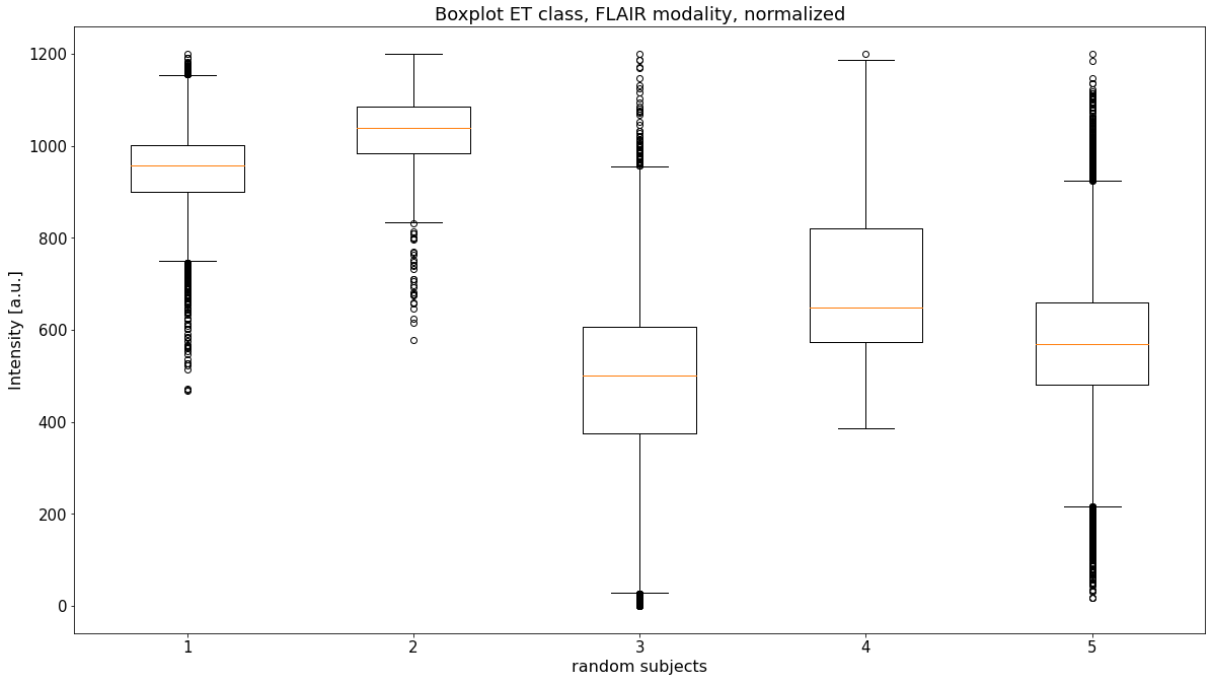


Figure 4.18: Box Plots of FLAIR images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

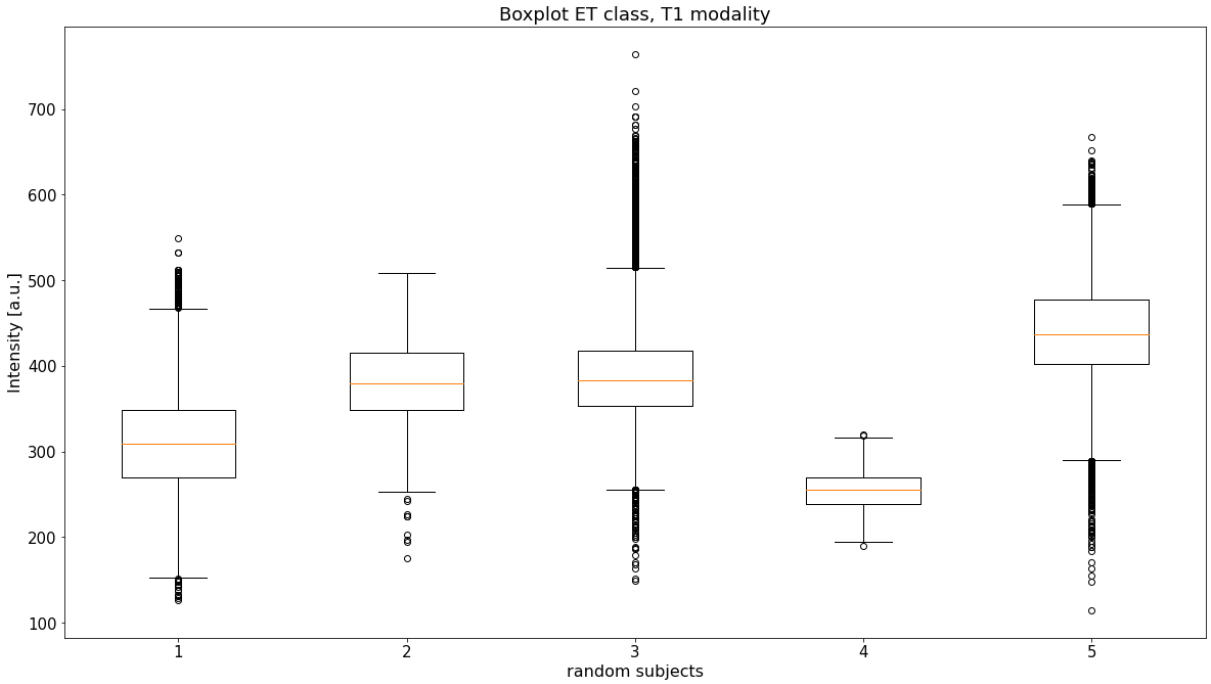


Figure 4.19: Box Plots of T1 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the third class (ET: enhancing tumor) and having removed zero values.

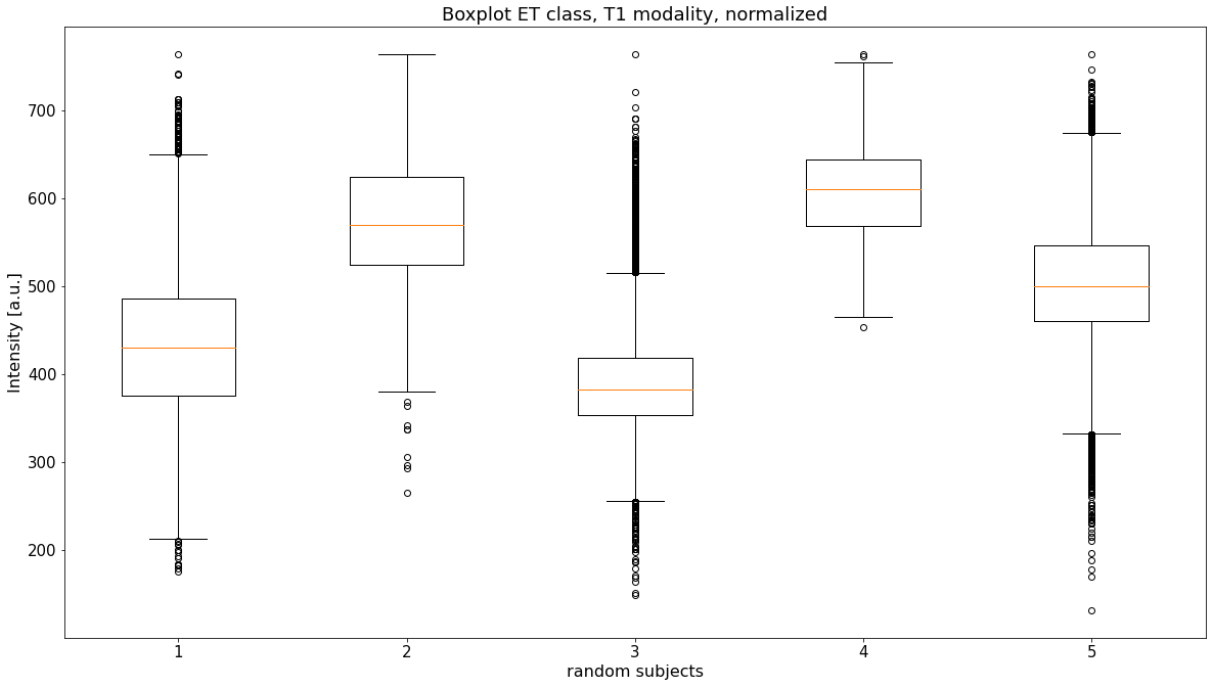


Figure 4.20: Box Plots of T1 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 3).

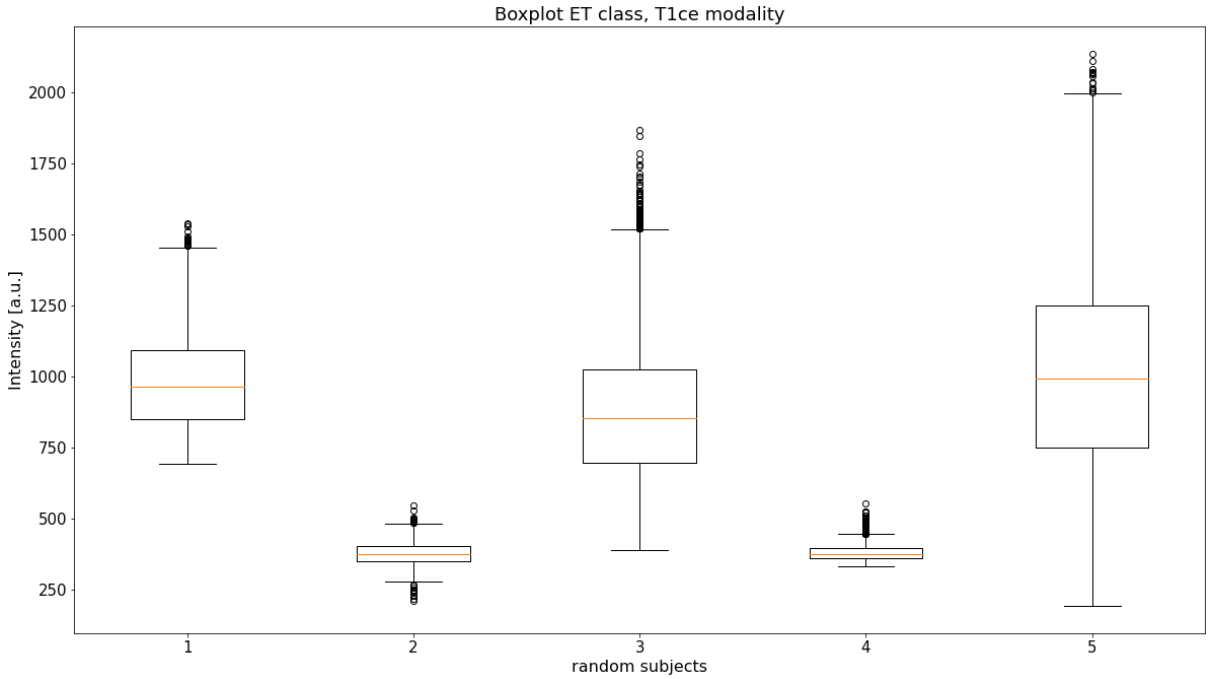


Figure 4.21: Box Plots of T1ce images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the third class (ET: enhancing tumor) and having removed zero values.

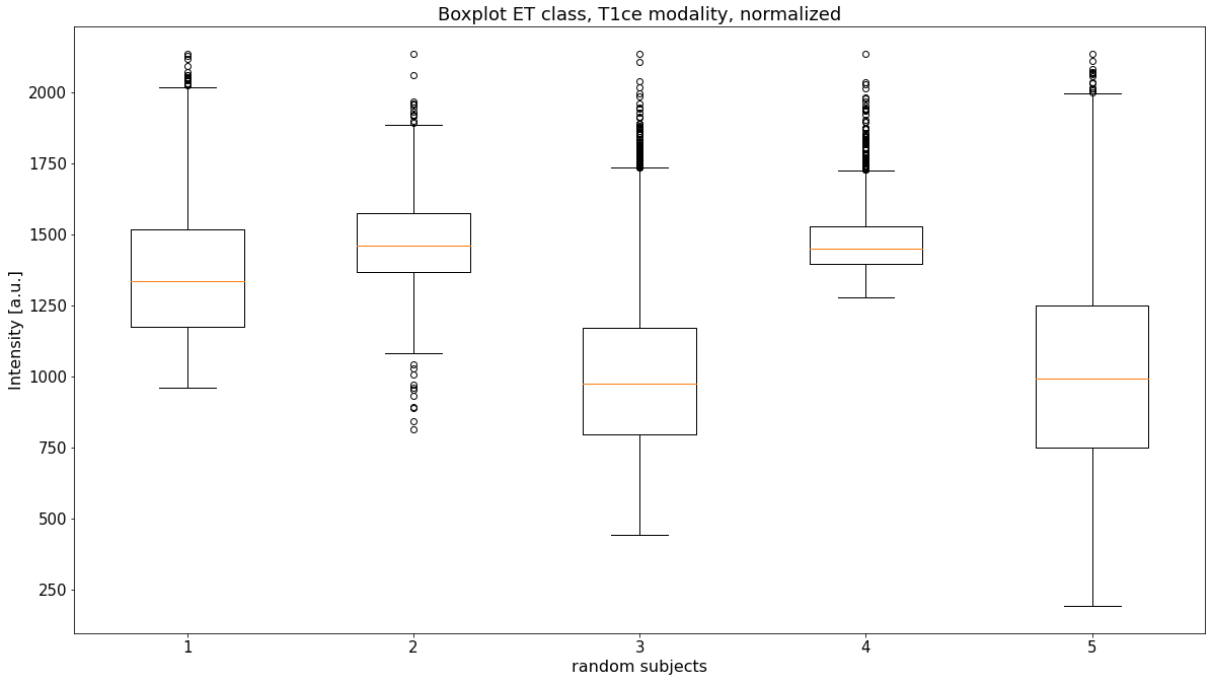


Figure 4.22: Box Plots of T1ce images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

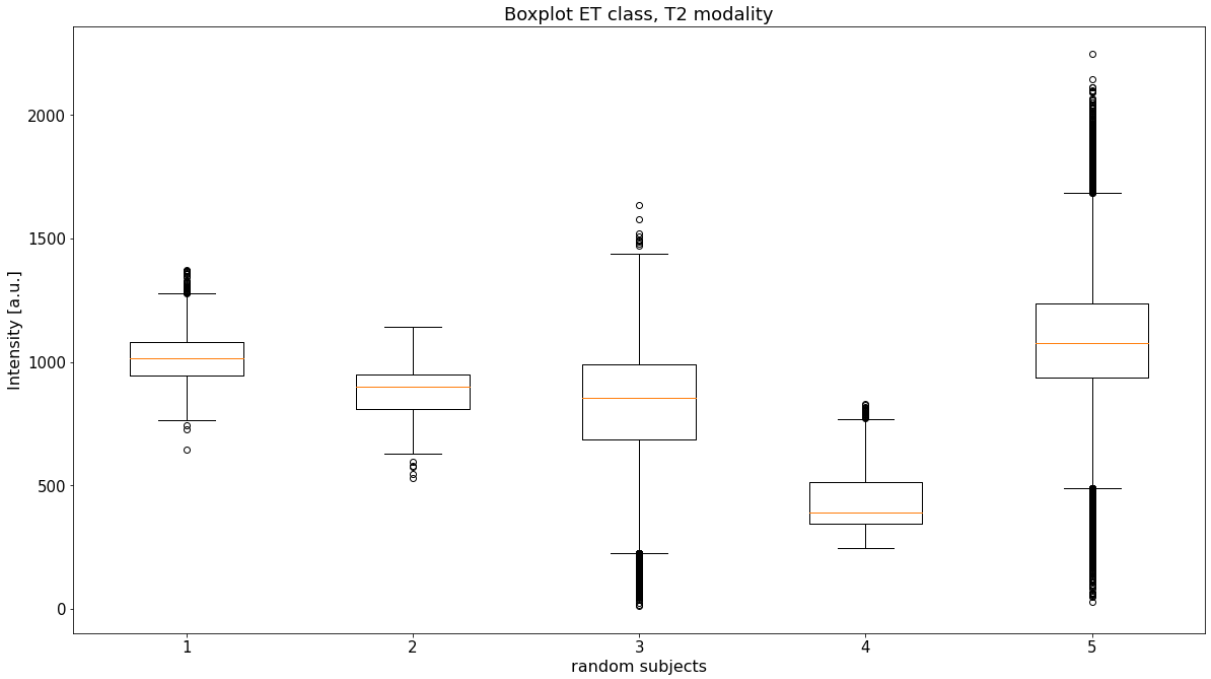


Figure 4.23: Box Plots of T2 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the third class (ET: enhancing tumor) and having removed zero values.

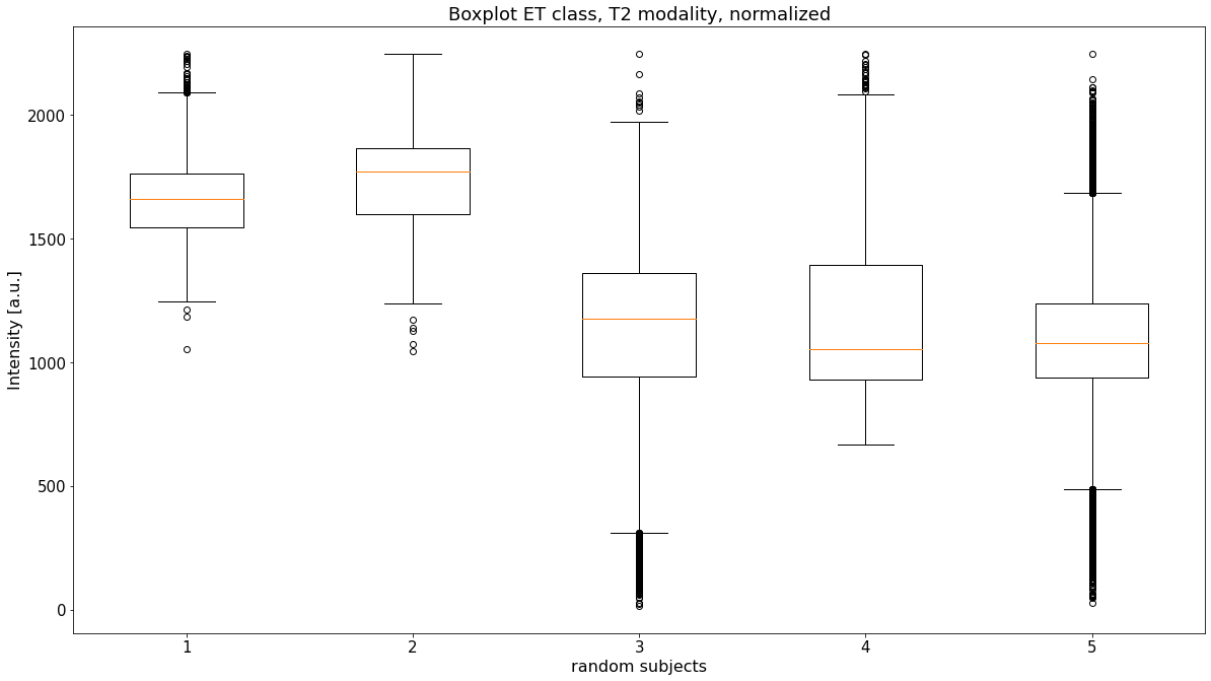


Figure 4.24: Box Plots of T2 images of 5 randomly extracted subjects, after having multiplied them for the corresponding segmentation masks for the first class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

4.1.2 FeTS 2022 dataset

Luckily, information about centers providing data is available for FeTS 2022 dataset, allowing a better comparison between values of images coming from different cohorts.

Not to make the analysis too heavy or tedious, only 5 centers from the 23 existing ones have been randomly extracted and compared between each other; the random extracted centers for this study were centers 4, 5, 13, 20, 21. To limit the magnitude of the analysis, it was chosen to randomly sample just one subject for each institution, after that a mask was created for each label (ED, NCR, ET) and multiplied for the images of the different modalities (FLAIR, T1, T1ce, T2). From the masked images, for each modality and for each class, Box Plots were obtained, always after removing zero values which were not informative

In this way, it was possible to compare box plots obtained from images of the same modalities for different centers, for each class, to actually examine the difference between images provided by the various institutions.

It was also chosen to normalize each group of visualized box plots (for each class and for each modality) for the maximum value reached by the visualized images, to effectively show the differences between images provided by dissimilar centers.

The box plots of the 5 images extracted from random centers, and their corresponding normalized box plots for the first class (ED), and for the different available modalities (FLAIR, T1, T1ce, T2), are showed respectively in *Figures 4.25, 4.26, 4.27, 4.28, 4.29, 4.30*, *4.31* and *4.32,* while the related ones for the second class (NCR) are showed in *Figures 4.33, 4.34, 4.35, 4.36, 4.37, 4.38, 4.39* and *4.40;* and finally for last class (ET) in *Figures 4.41, 4.42, 4.43, 4.44, 4.45, 4.46, 4.47* and *4.48.*

From all these images it is possible to appreciate the different dispersion of values of images acquired with the same acquisition setup, but from different centers.

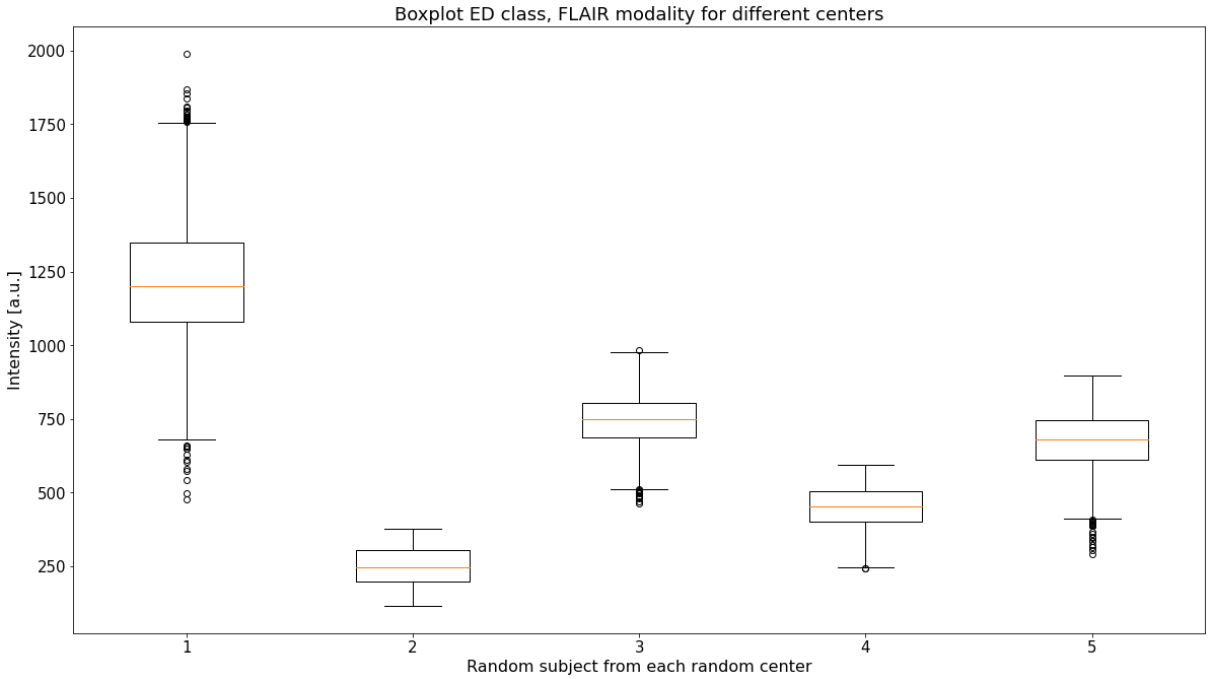


Figure 4.25: Box Plots of FLAIR images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

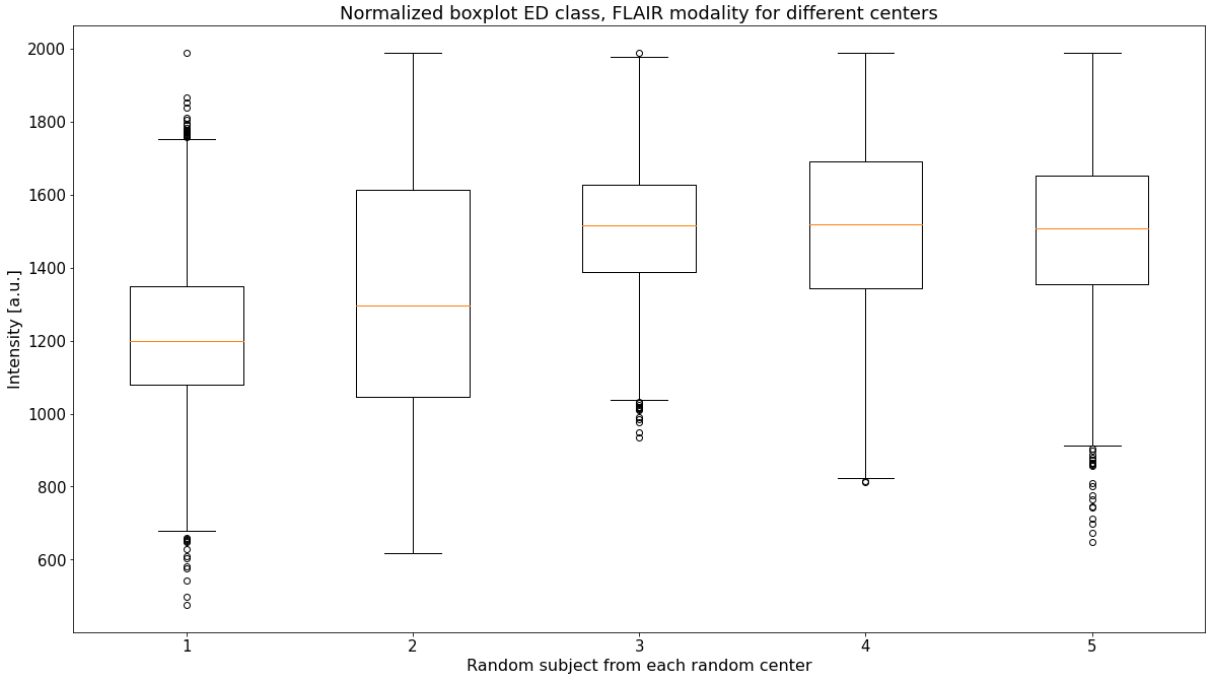


Figure 4.26: Box Plots of FLAIR images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

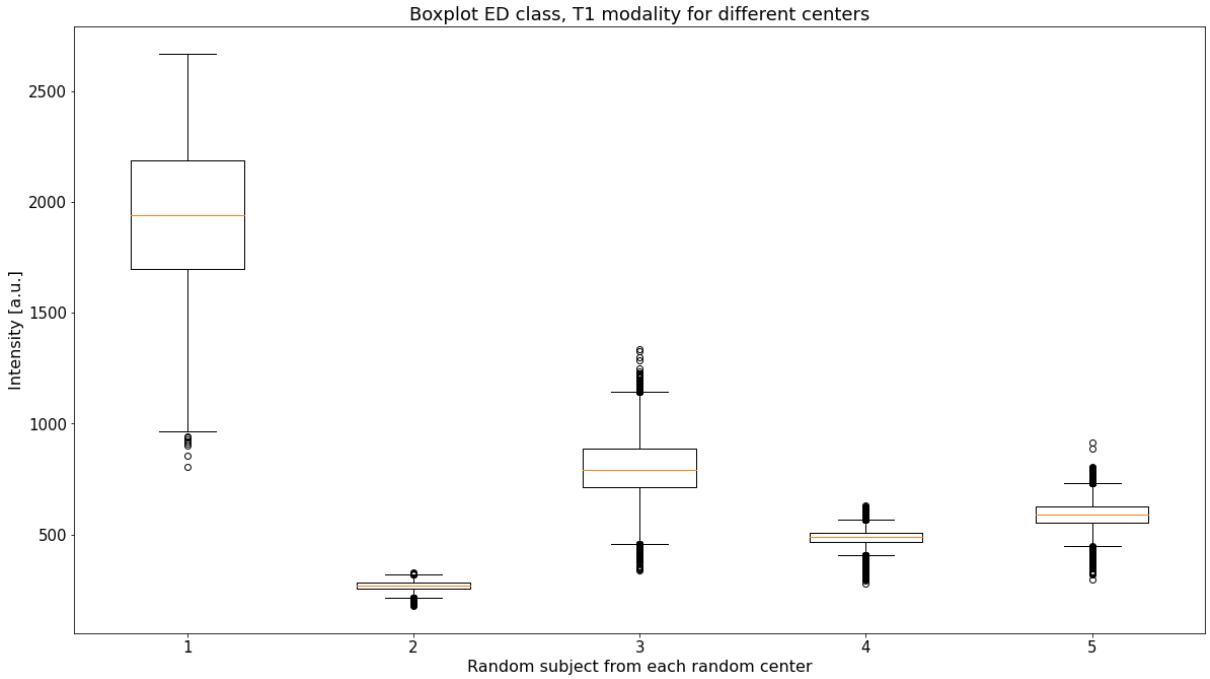


Figure 4.27: Box Plots of T1 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

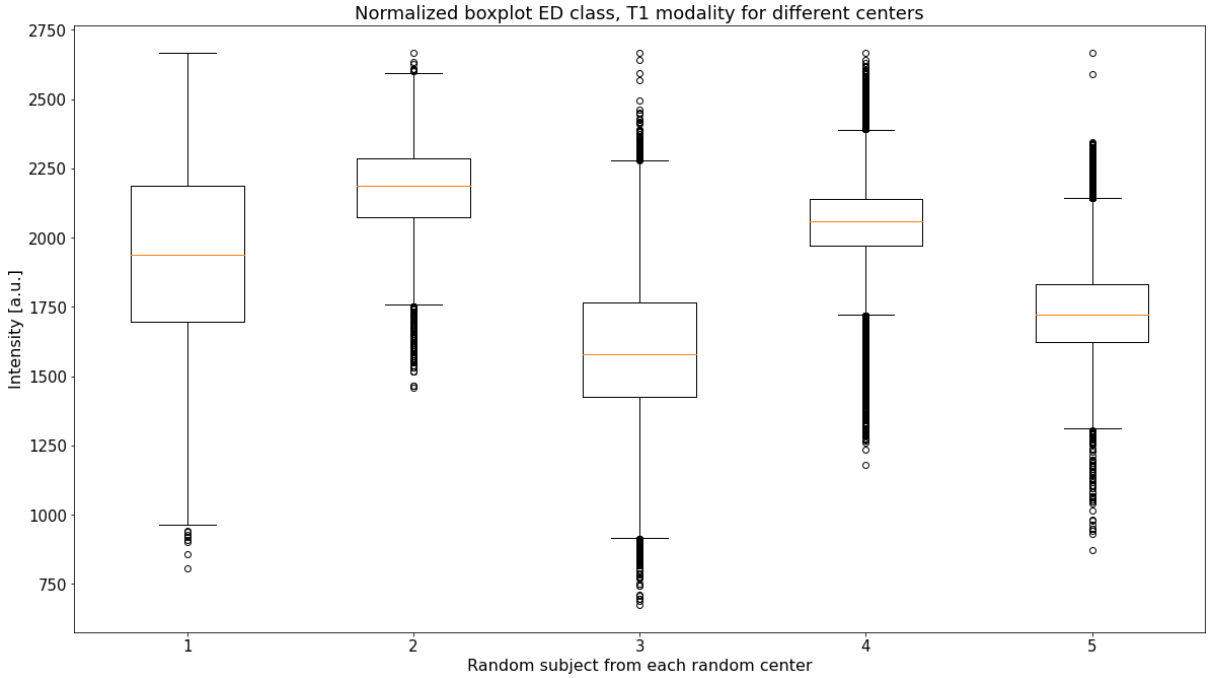


Figure 4.28: Box Plots of T1 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

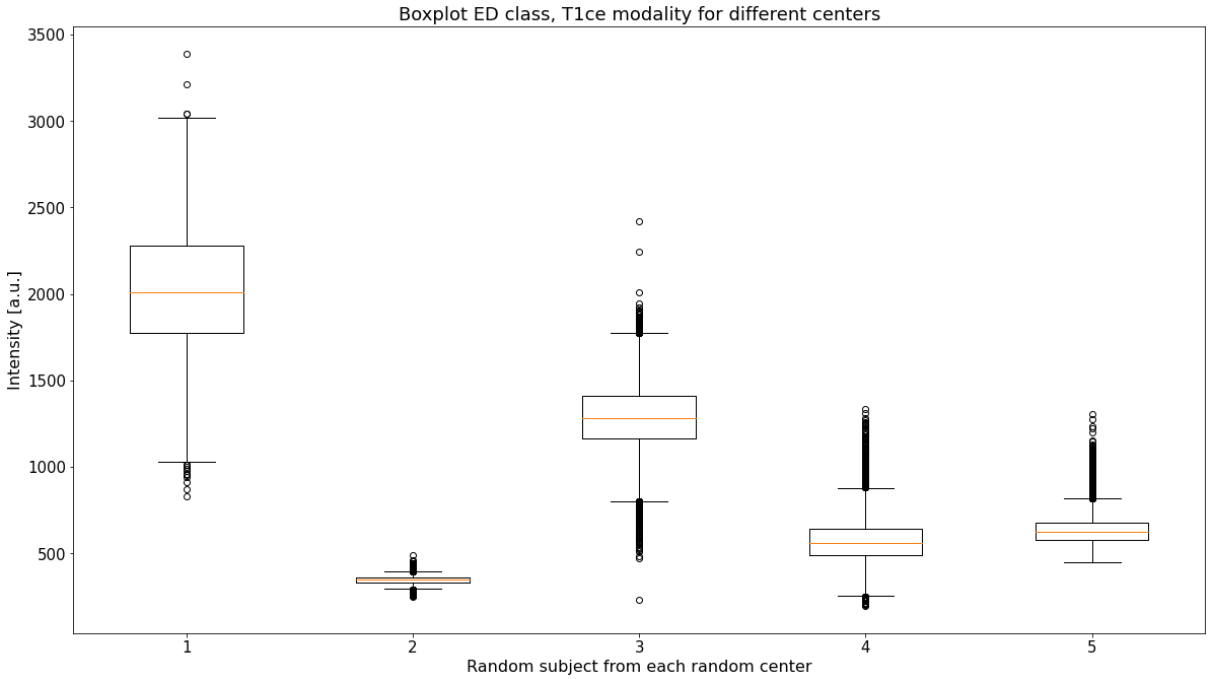


Figure 4.29: Box Plots of T1ce images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

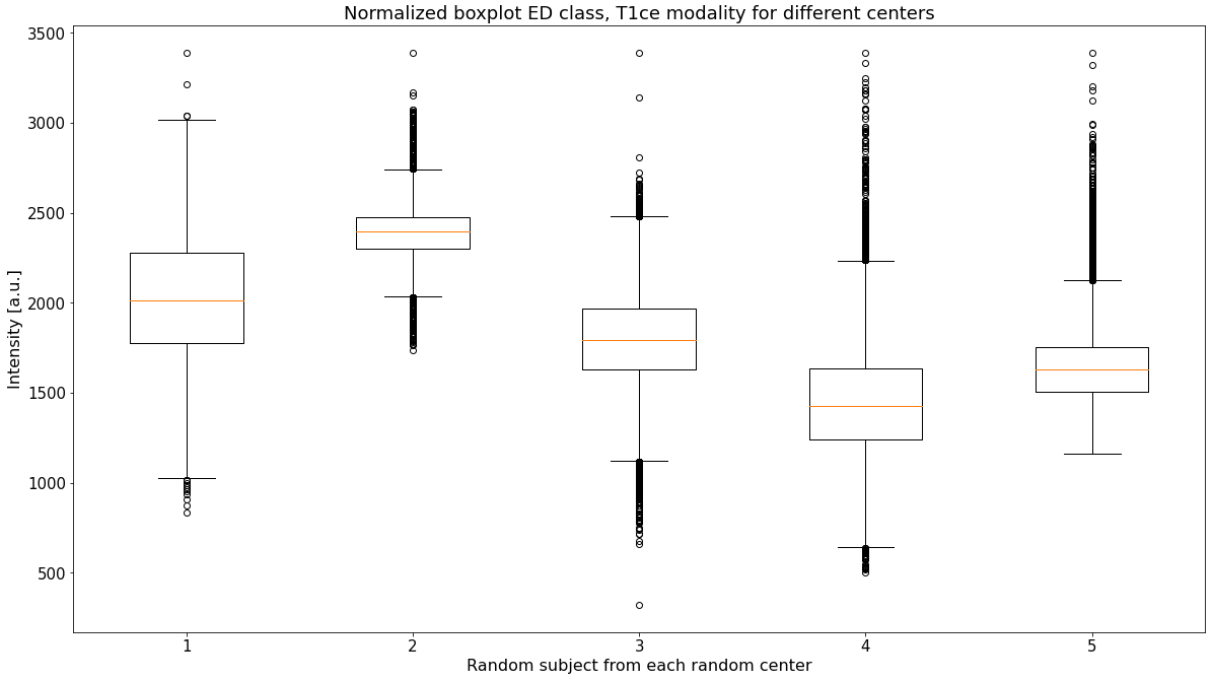


Figure 4.30: Box Plots of T1ce images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

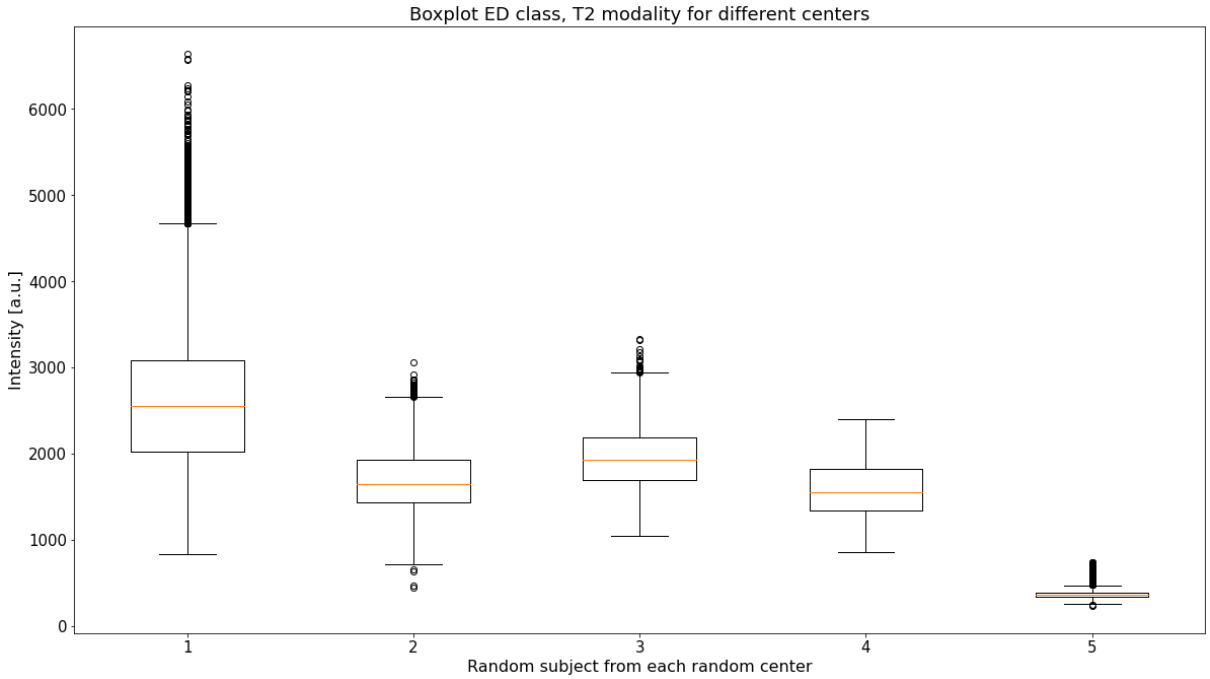


Figure 4.31: Box Plots of T2 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue) and having removed zero values.

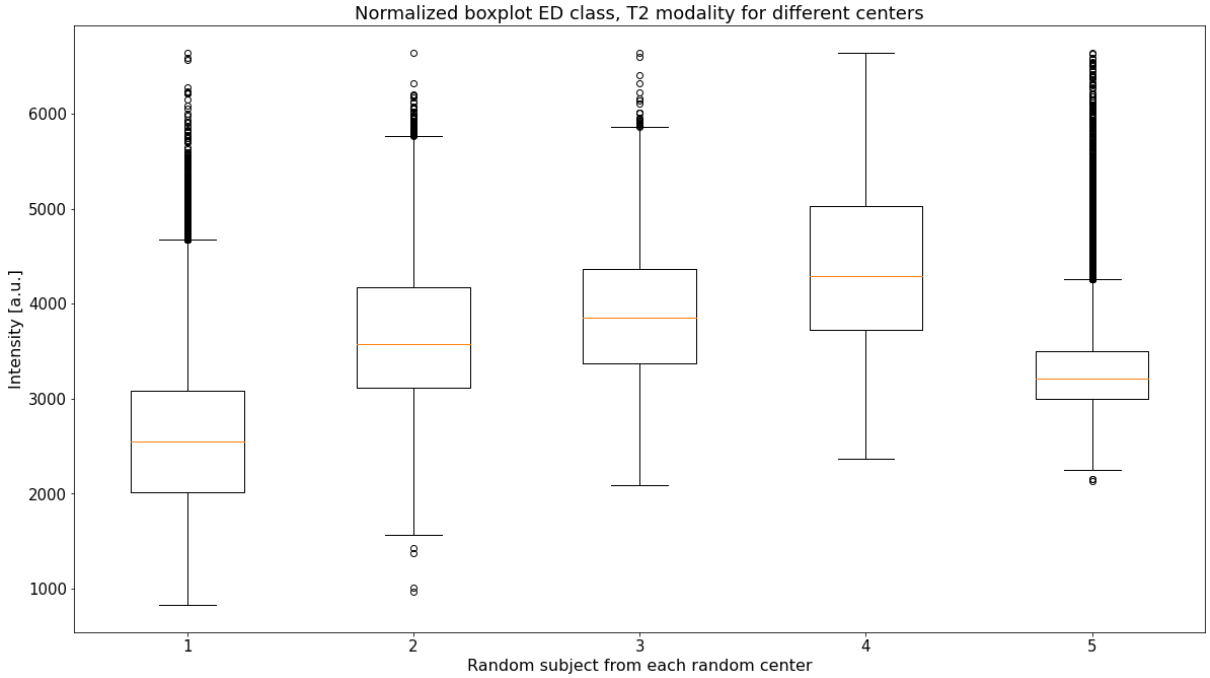


Figure 4.32: Box Plots of T2 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the first class (ED: peritumoral edematous/invaded tissue), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

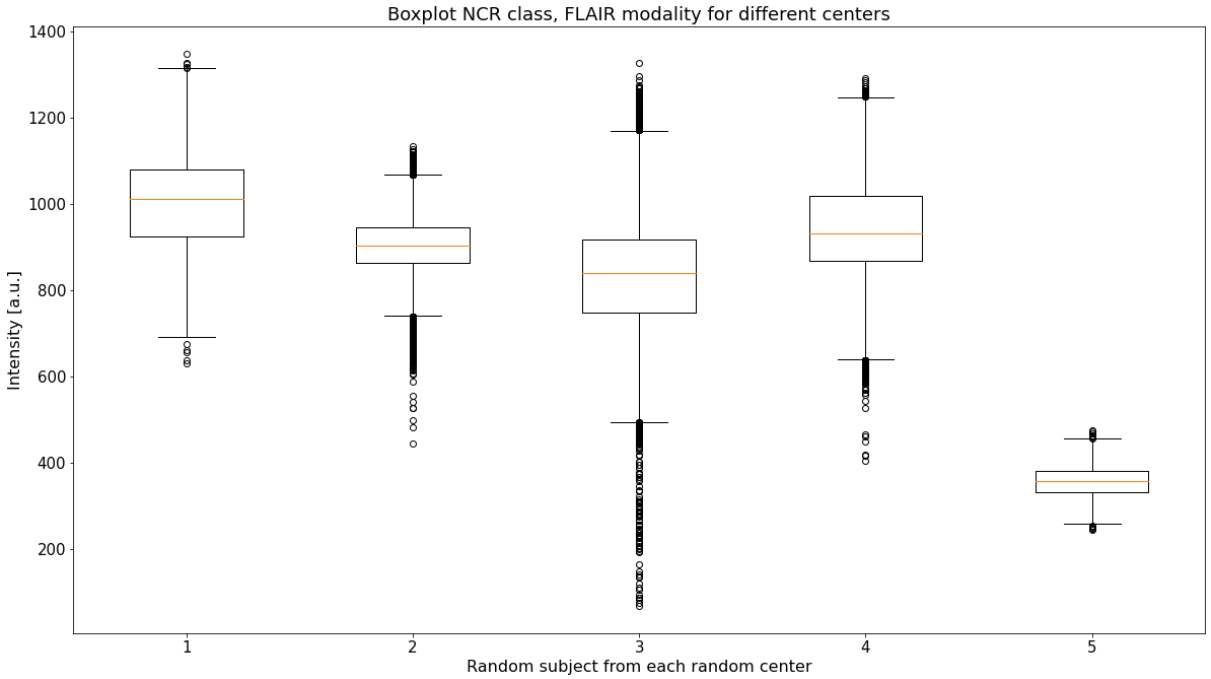


Figure 4.33: Box Plots of FLAIR images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

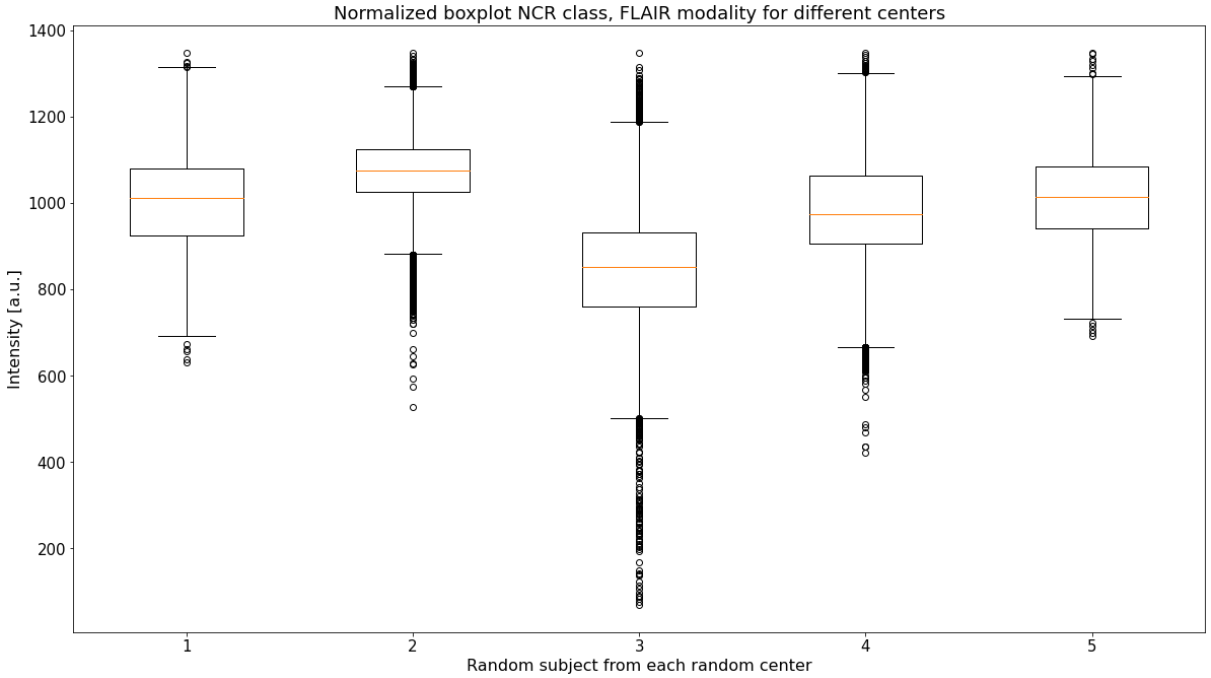


Figure 4.34: Box Plots of FLAIR images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

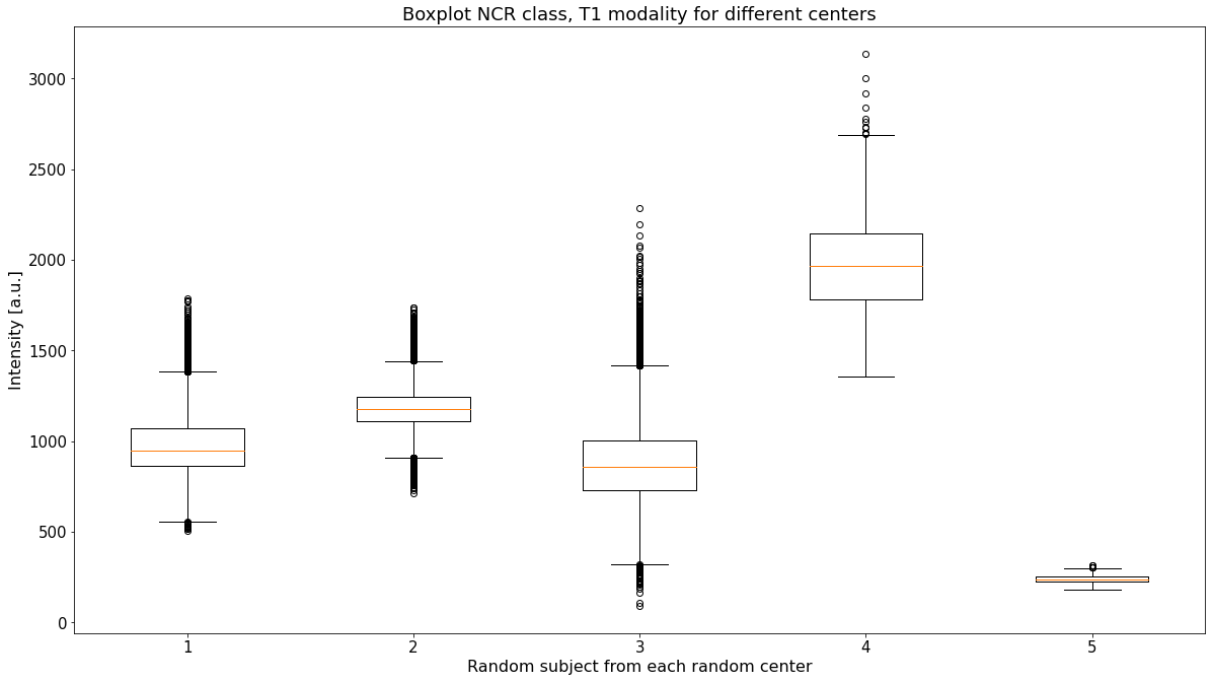


Figure 4.35: Box Plots of T1 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

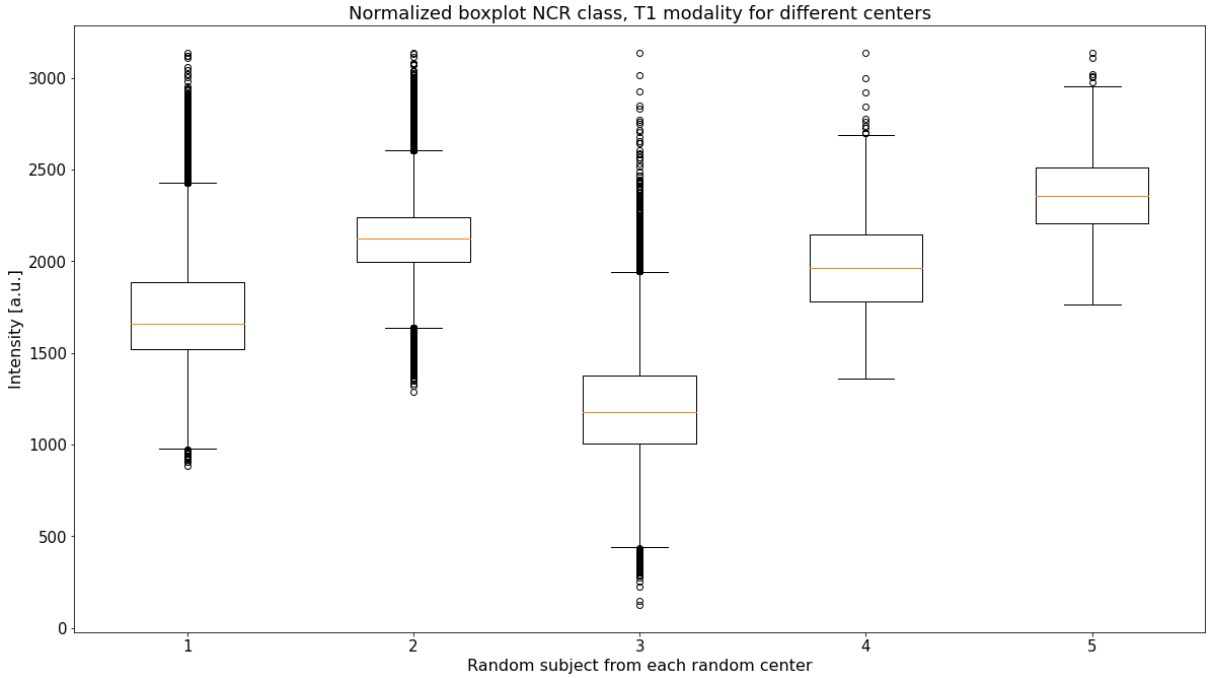


Figure 4.36: Box Plots of T1 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 4).

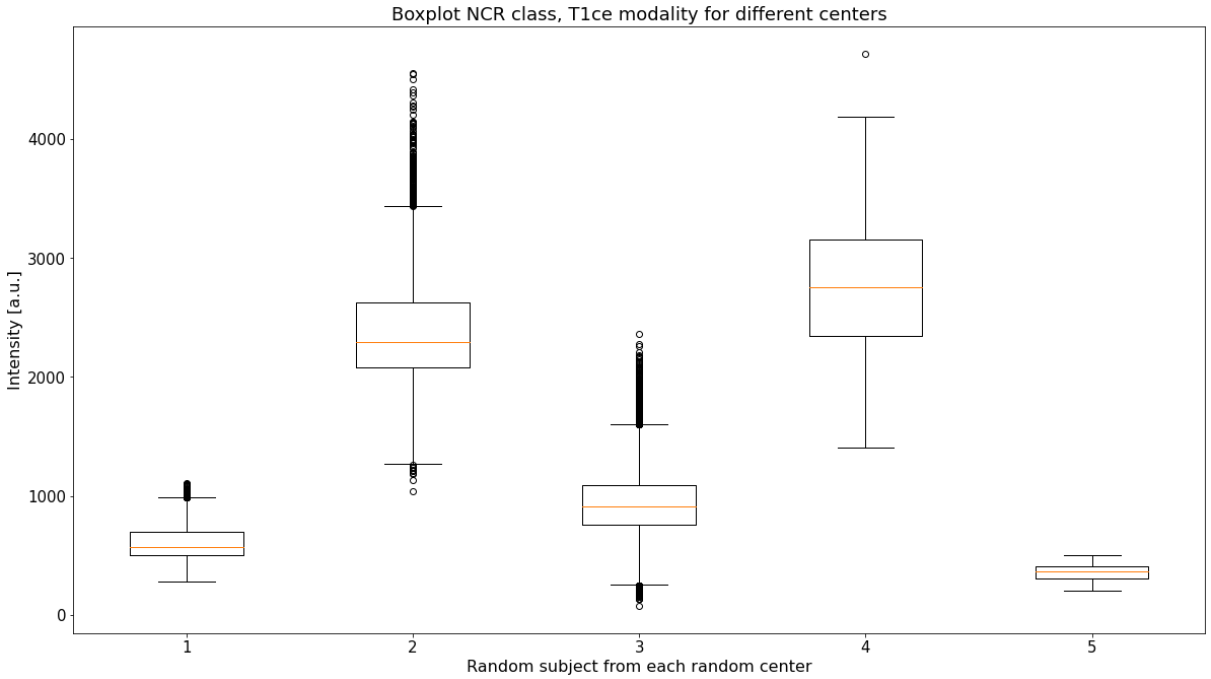


Figure 4.37: Box Plots of T1ce images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

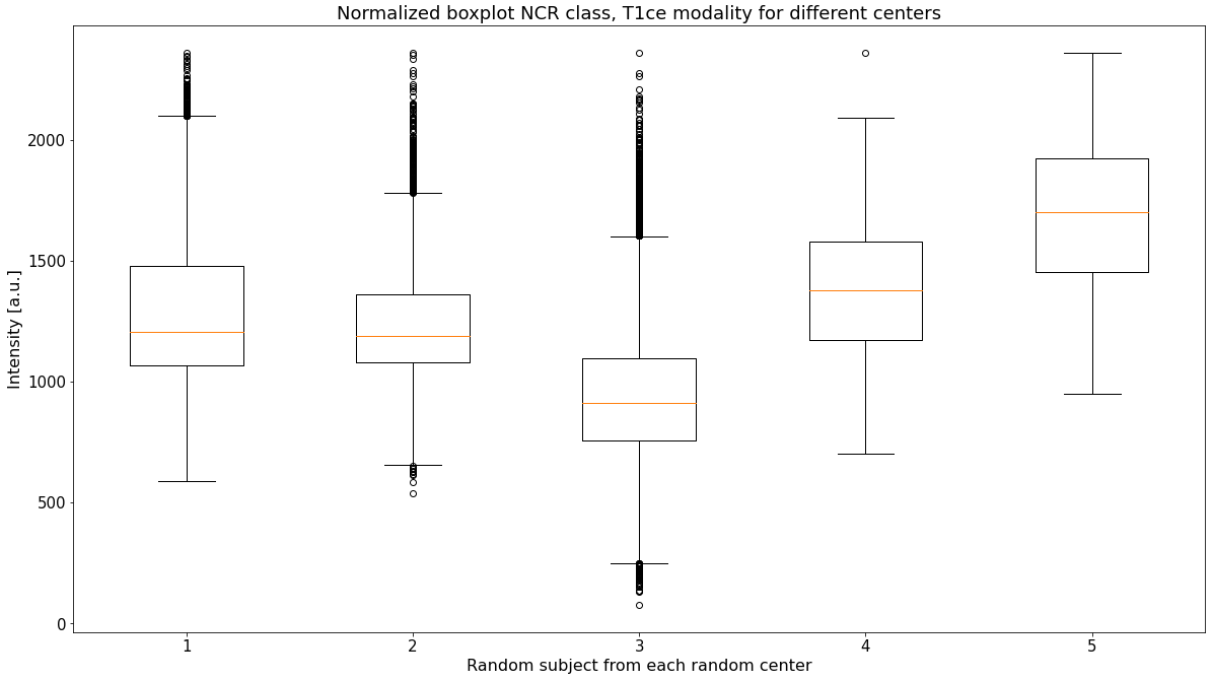


Figure 4.38: Box Plots of T1ce images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 2).

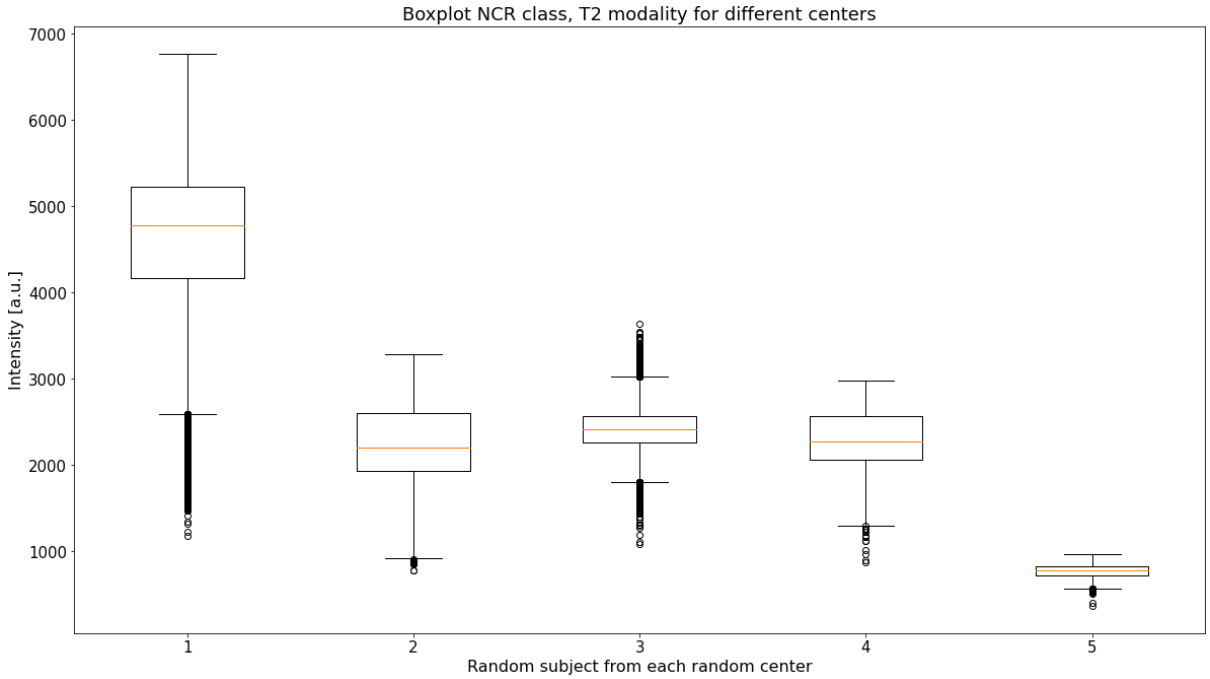


Figure 4.39: Box Plots of T2 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core) and having removed zero values.

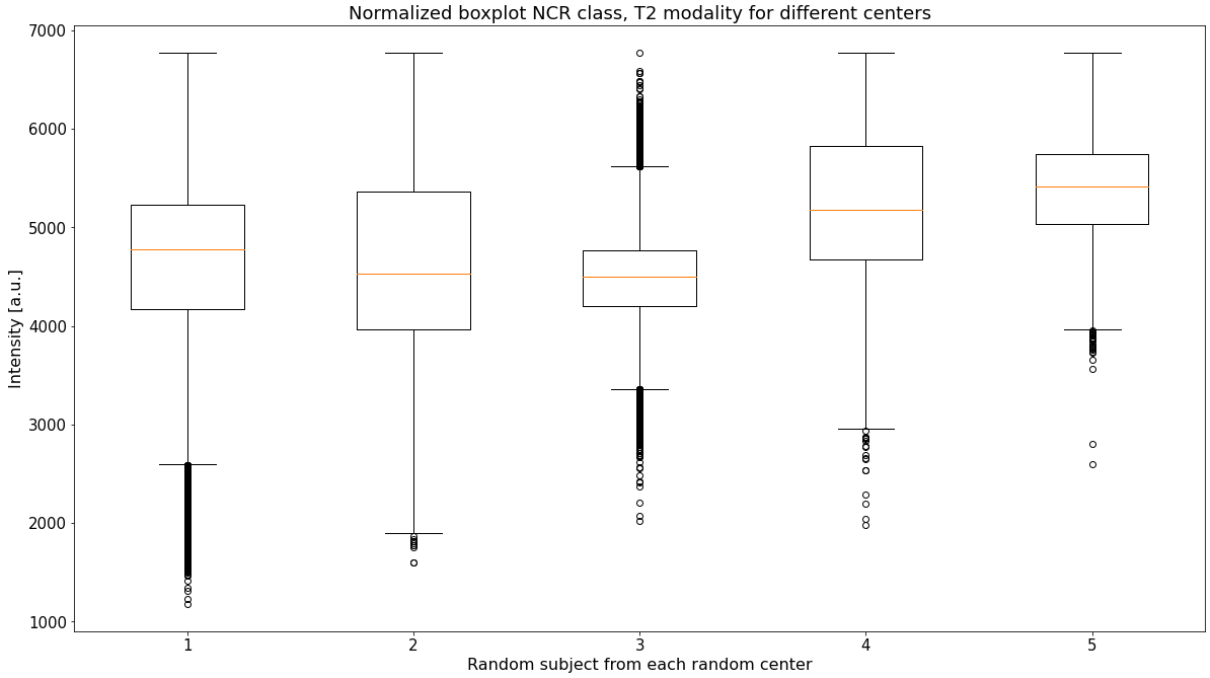


Figure 4.40: Box Plots of T2 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the second class (NCR: necrotic tumor core), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

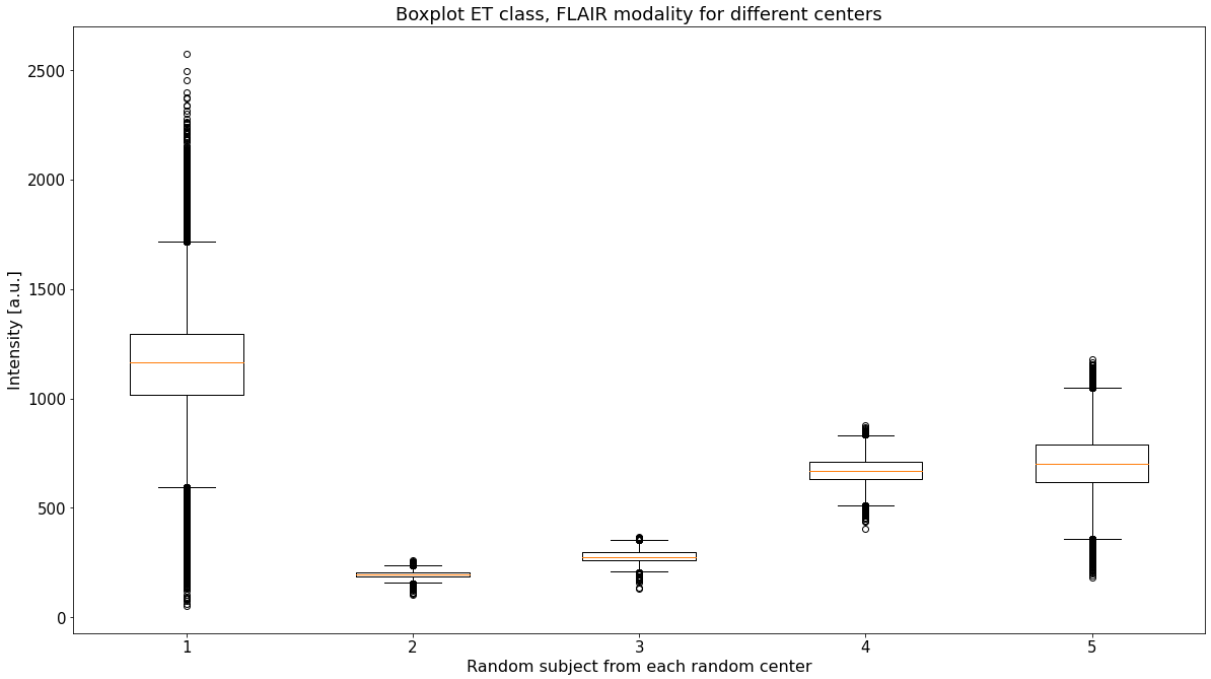


Figure 4.41: Box Plots of FLAIR images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor) and having removed zero values.

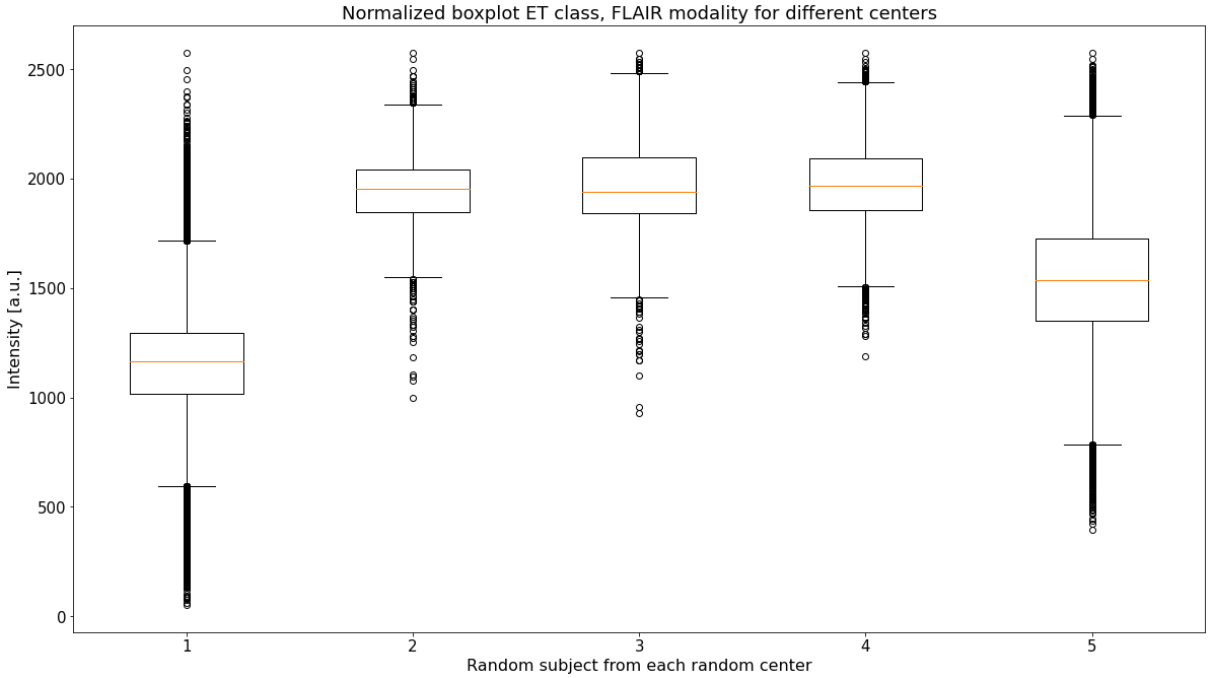


Figure 4.42: Box Plots of FLAIR images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

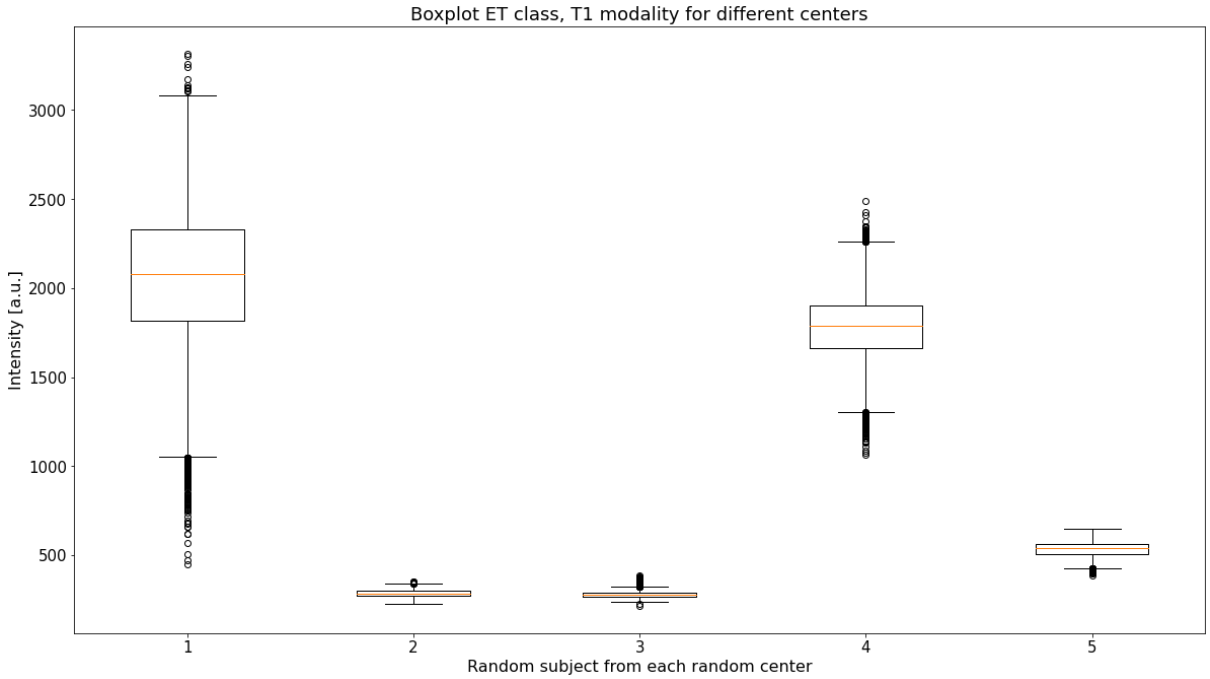


Figure 4.43: Box Plots of T1 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor) and having removed zero values.

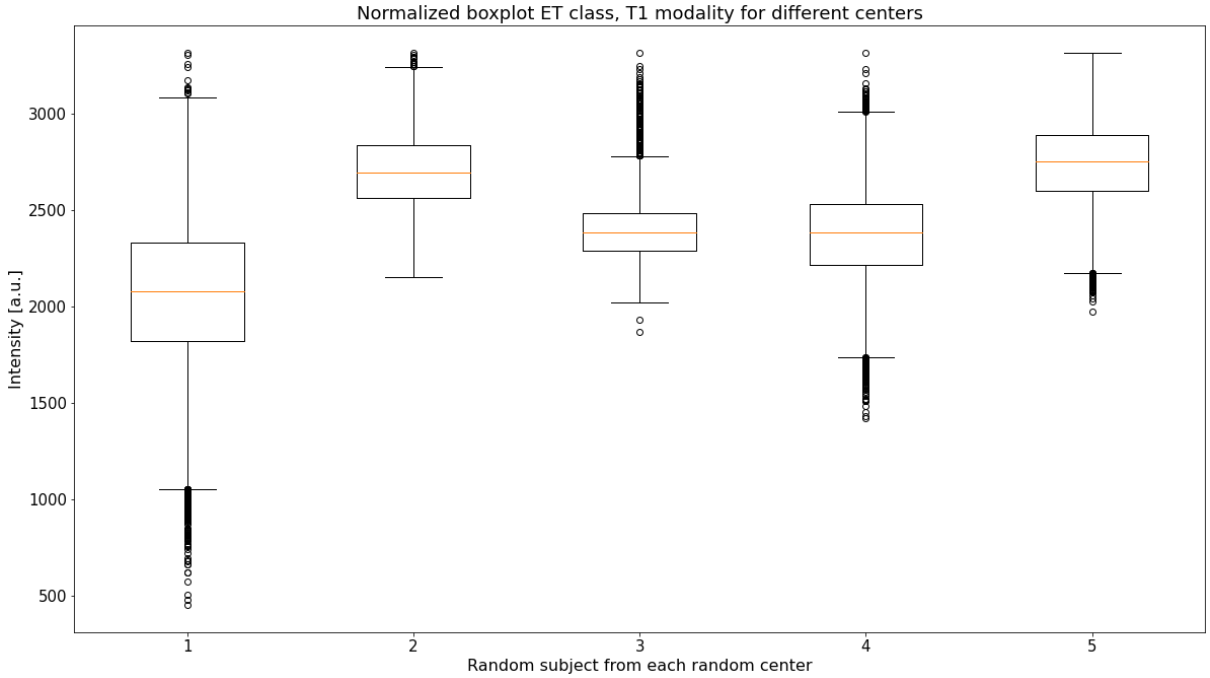


Figure 4.44: Box Plots of T1 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

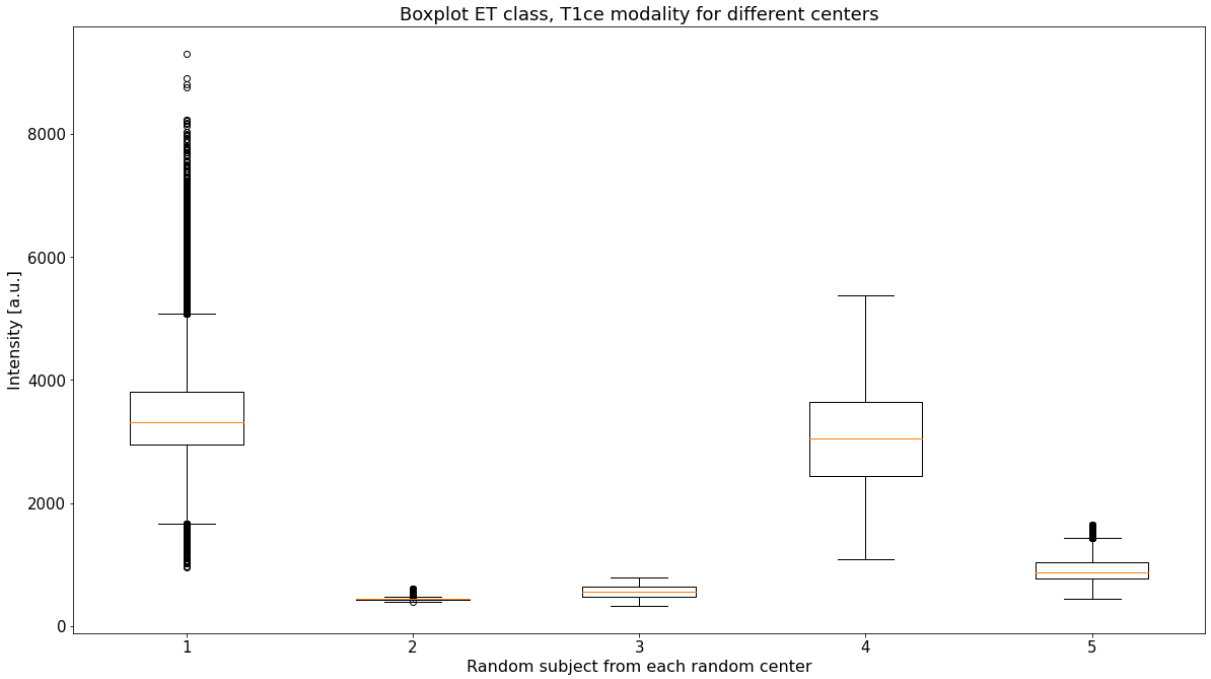


Figure 4.45: Box Plots of T1ce images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor) and having removed zero values.

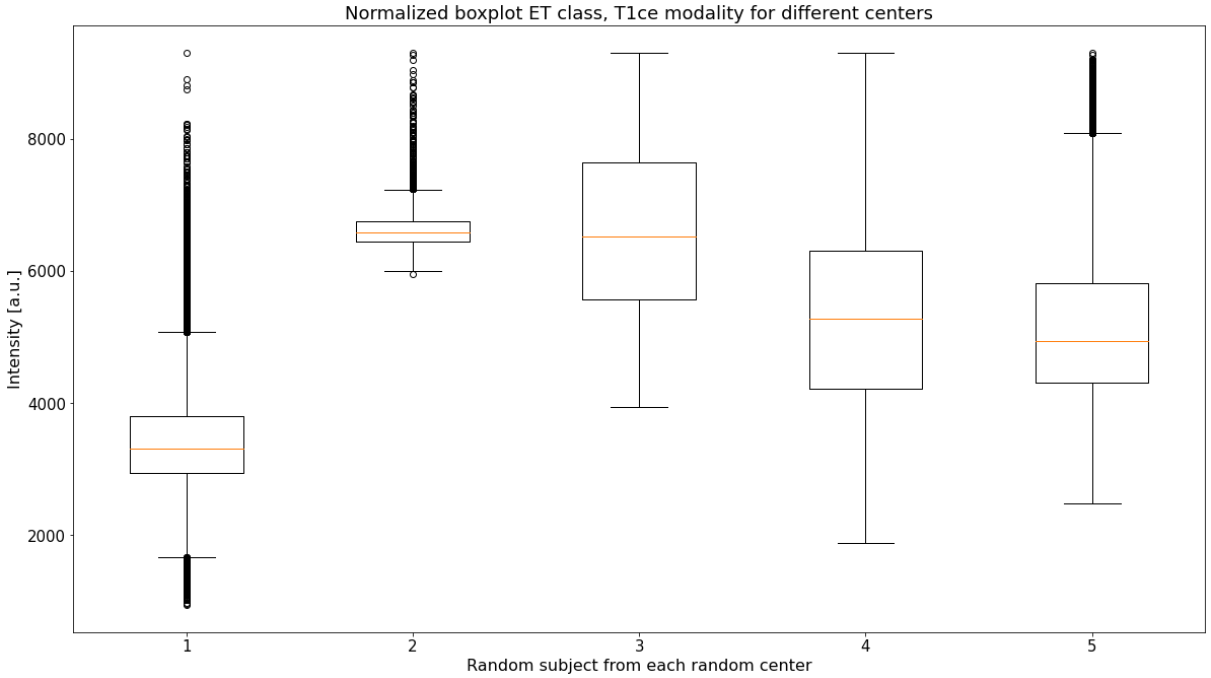


Figure 4.46: Box Plots of T1ce images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

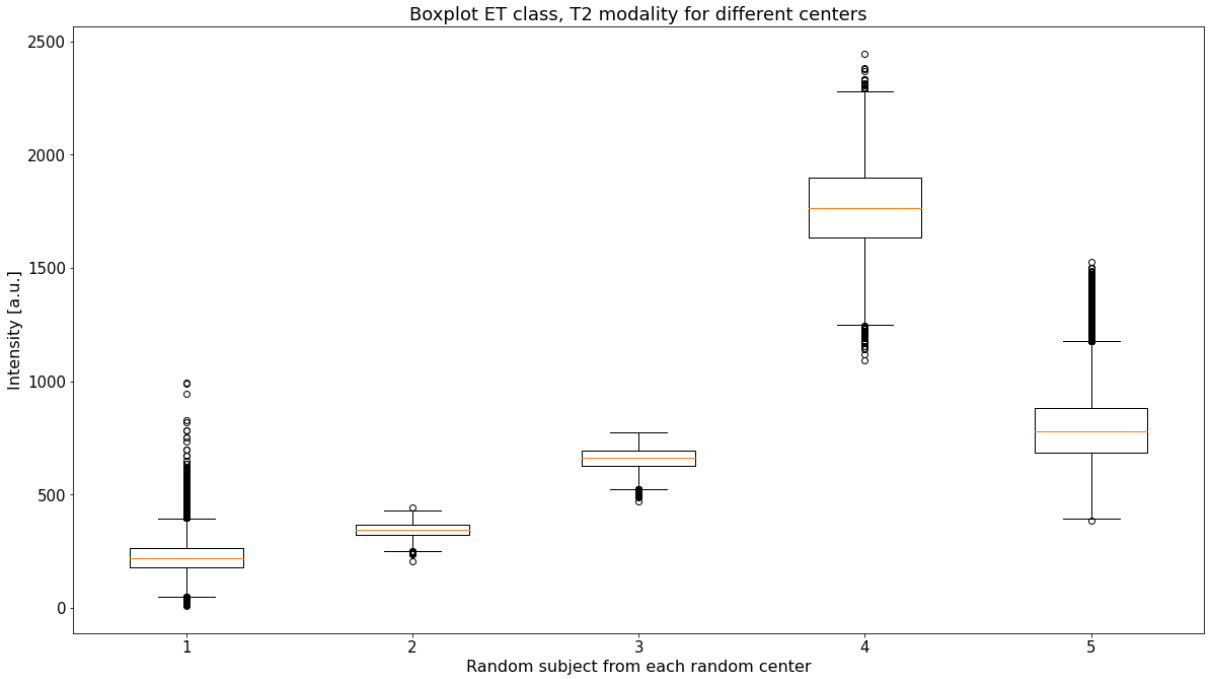


Figure 4.47: Box Plots of T2 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor) and having removed zero values.

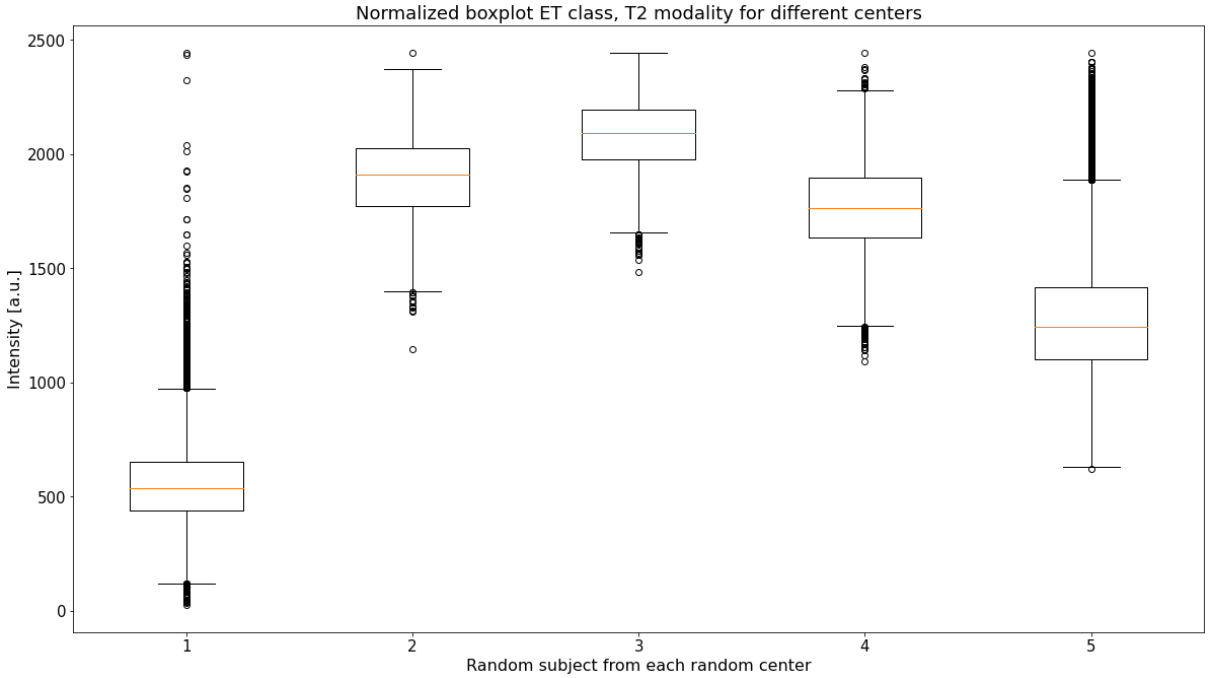


Figure 4.48: Box Plots of T2 images of 5 randomly extracted subjects from 5 randomly chosen center (one for each center), after having multiplied them for the corresponding segmentation masks for the last class (ET: enhancing tumor), having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 4).

4.1.3 ISLES 2022 dataset

For the exploratory analysis of the ISLES 2022 training set, 5 subjects were randomly drawn from the dataset, as for previous analyzed datasets. The following procedure was then repeated for the realization of representative box plots: for each modality, a new image was obtained by multiplying the segmentation mask of the ischemic stroke lesion with the corresponding scan; after, zero values were extruded because they were not informative, and finally box plots were derived from these concatenated images.

As before, for a better visualization and comparison between box plots, they were also normalized for the highest value reached by all analyzed images, for each modality.

The box plots of the 5 images, and their corresponding normalized box plots for the unique class of the stroke lesion, and for the different available modalities (ADC, DWI and FALIR), are showed respectively in *Figures 4.49, 4.50, 4.51, 4.52, 4.53* and *4.54.*

From the comparison between box plots of images of different modalities it's possible to notice that the appearance of the ischemic stroke lesion significantly changes from one modality to the other: the position and dimension of box and whiskers, the number of outliers and the value of the median are very different between subjects. Moreover, it can be noticed that some images are characterized by values which are so lower with respect to the other images acquired with the same acquisition setup, that their boxplots, compared to the others, are not even visible. Only after normalization it's possible to appreciate it. This peculiarity underlines the huge difference of values that it's possible to find inside images of the same datasets.

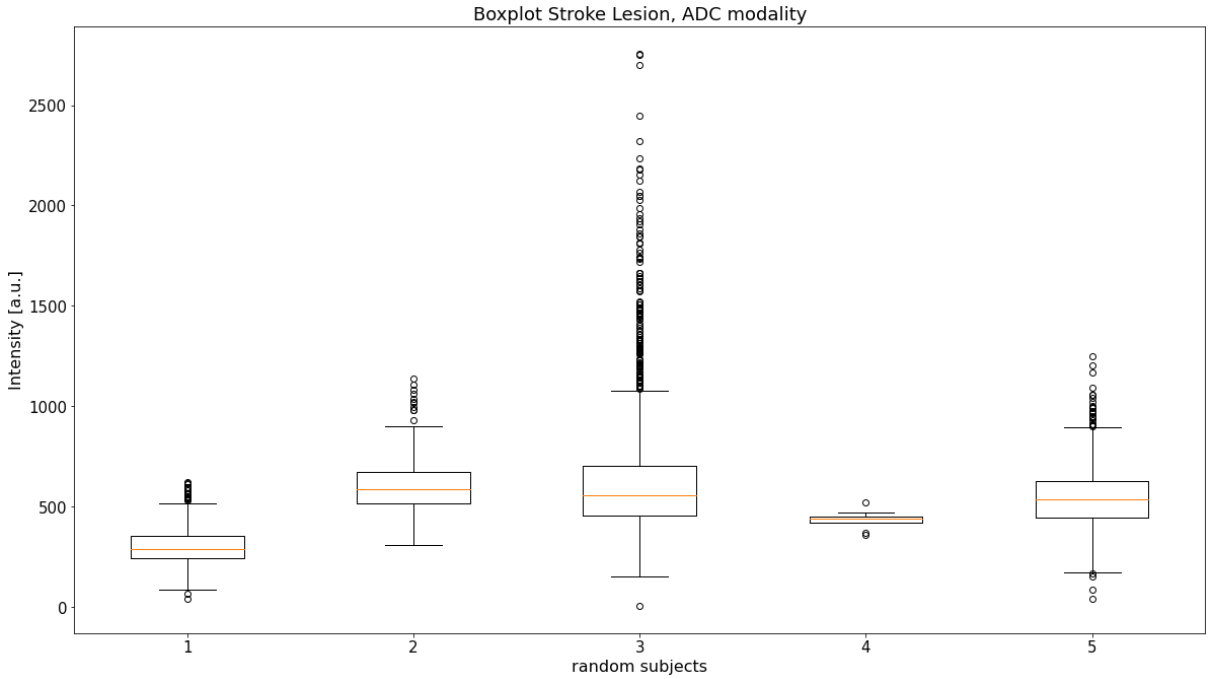


Figure 4.49: Box Plots of 5 randomly extracted ADC images, after having multiplied them for the corresponding segmentation masks for the stroke lesion and having removed zero values.

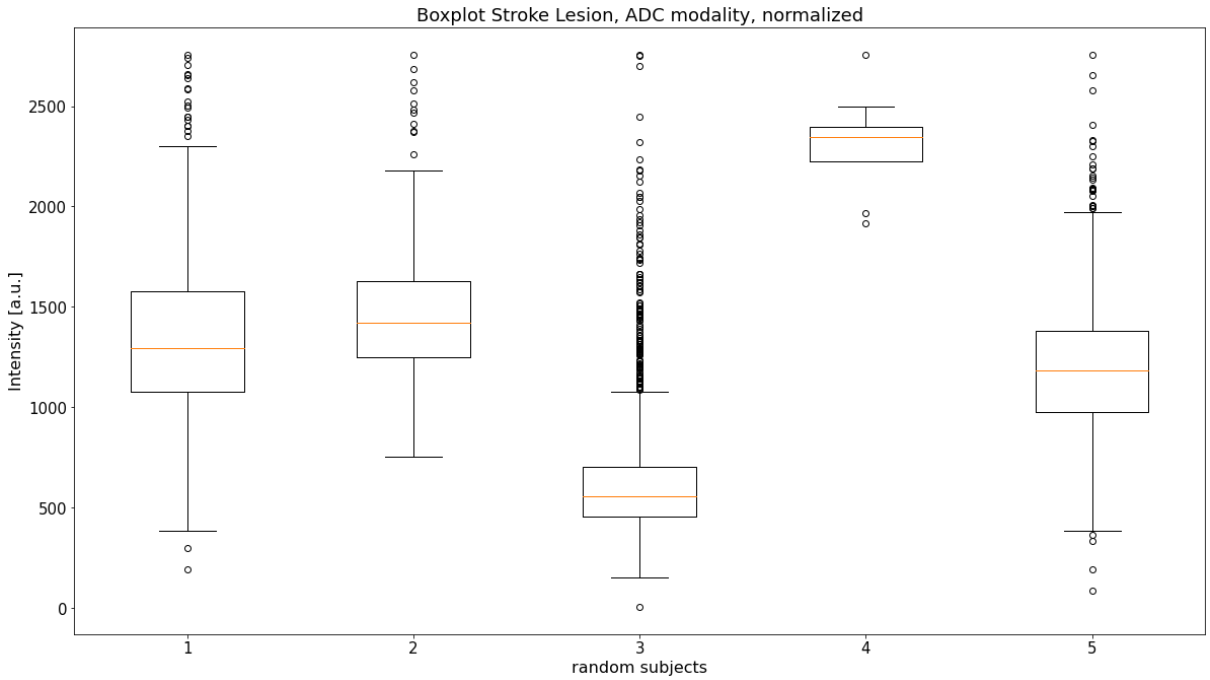


Figure 4.50: Box Plots of 5 randomly extracted ADC images, after having multiplied them for the corresponding segmentation masks for the stroke lesion, having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 3).

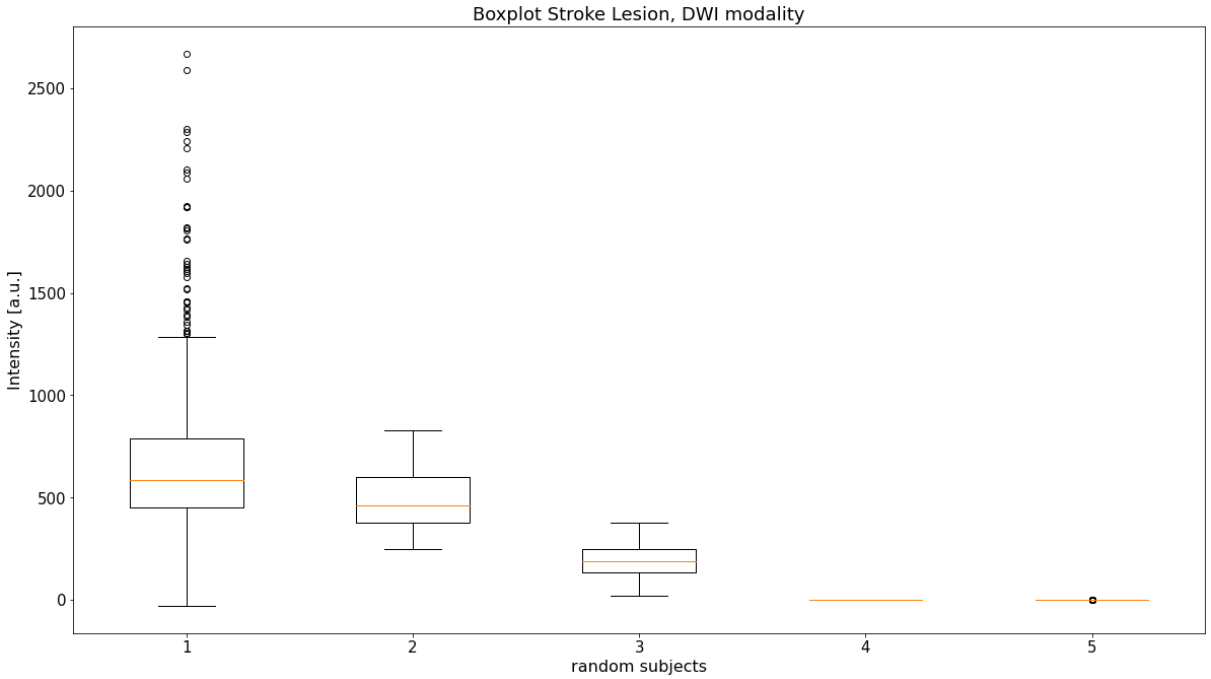


Figure 4.51: Box Plots of 5 randomly extracted DWI images, after having multiplied them for the corresponding segmentation masks for the stroke lesion and having removed zero values.

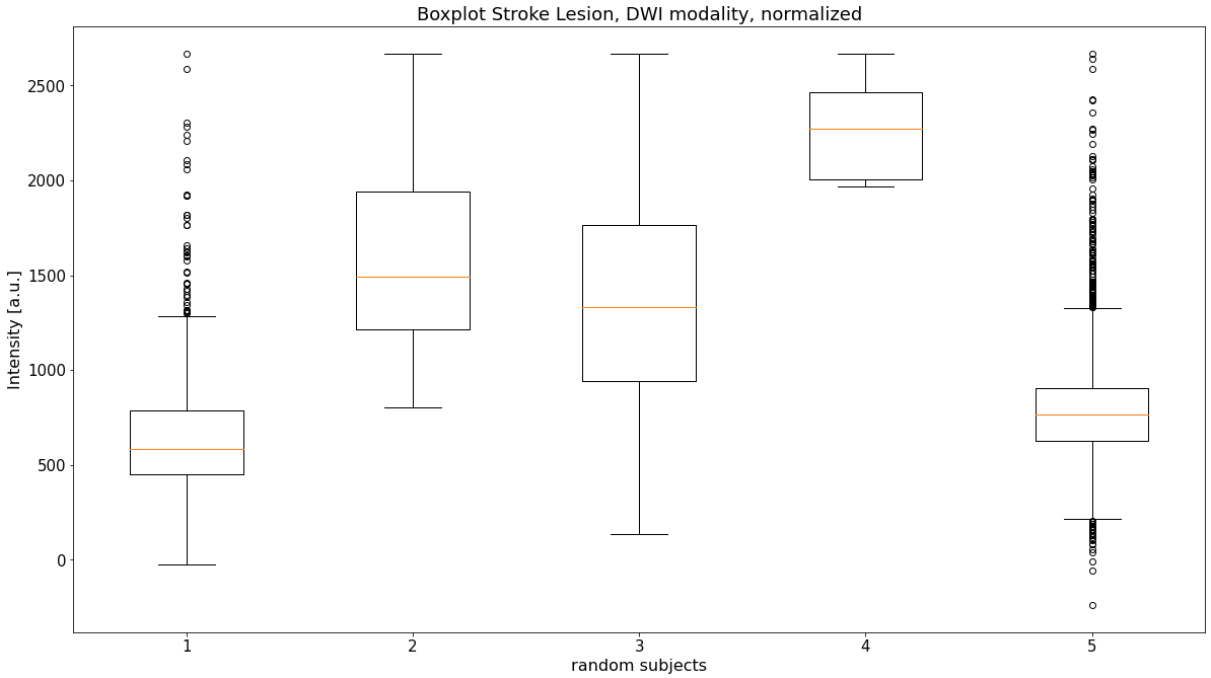


Figure 4.52: Box Plots of 5 randomly extracted DWI images, after having multiplied them for the corresponding segmentation masks for the stroke lesion, having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 1).

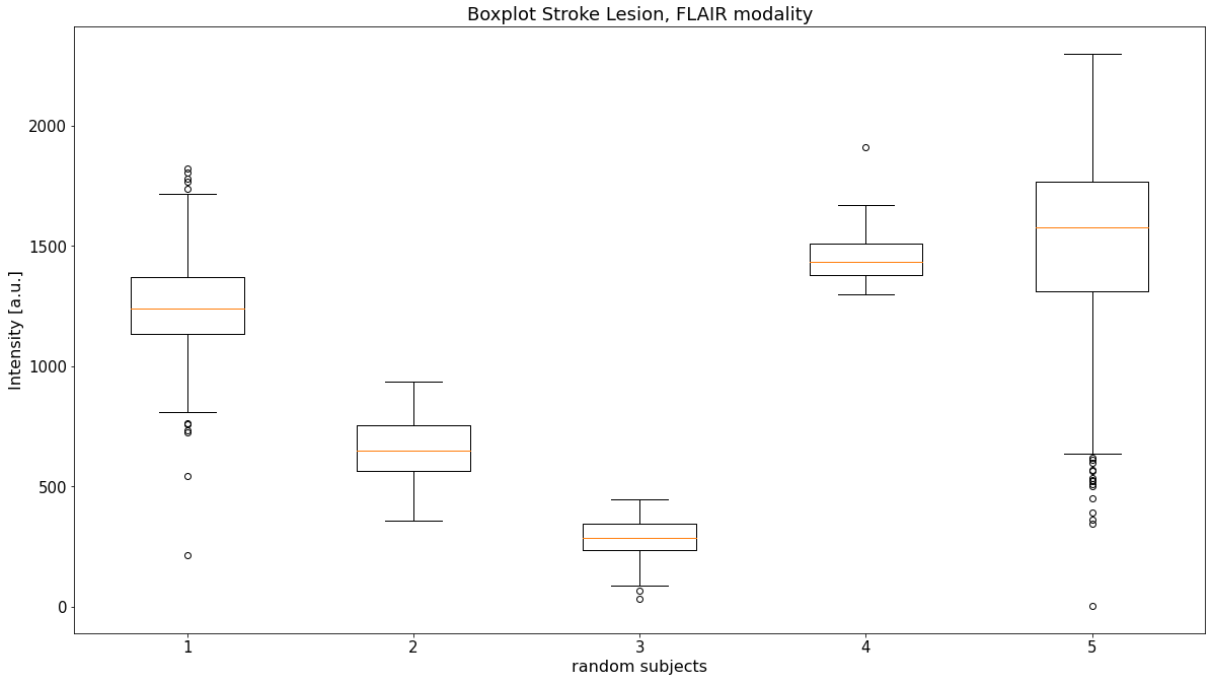


Figure 4.53: Box Plots of 5 randomly extracted FLAIR images, after having multiplied them for the corresponding segmentation masks for the stroke lesion and having removed zero values.

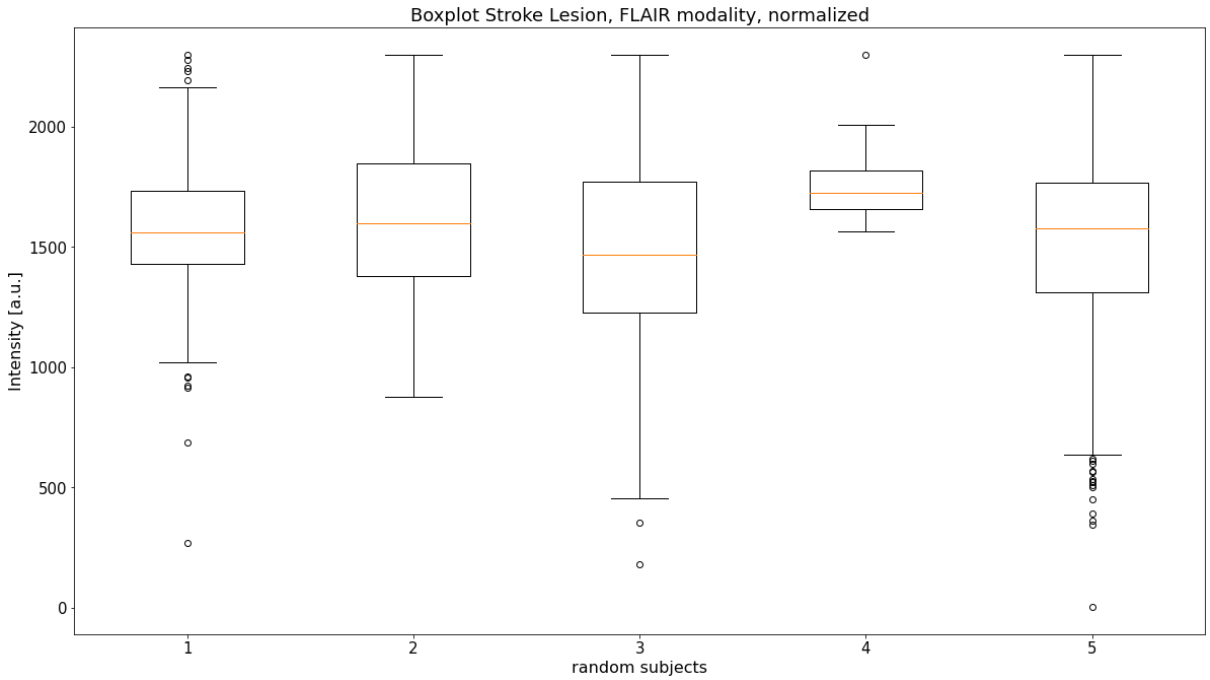


Figure 4.54: Box Plots of 5 randomly extracted FLAIR images, after having multiplied them for the corresponding segmentation masks for the stroke lesion, having removed zero values and having normalized them to the maximum value reached by all images (obtained by image 5).

4.2 Implementation of the best performing model

nnUNet was deeply analyzed and used as starting point for this thesis not only because it represents the state-of-the-art method for brain tumor segmentation, winning both BraTS 2020 and BraTS 2021 competitions, but also for its capability to adapt to every task and dataset automatically: nnUNet just needs to receive as input data organized in the correct manner and format, and it will be able to generate a dataset fingerprint independently. *Dataset.json* is a json file that must be generated once the data have been correctly organized, and it contains the number of training and test set, the modalities and labels; starting from it nnUNet extracts all the dataset characteristics during experiment planning and preprocessing, and automatically selects the best UNet configuration in an easy and straightforward way.

For these reasons nnUNet was chosen as baseline to perform further analysis. nnUNet architecture can obviously be modified, but it has been chosen to include only minor modifications to achieve my purposes, also to respect the ideal of the authors, which stated that small architectural modifications are not superior to a properly tuned model, and often a configuration carefully adapted to the given task performs better than a complex and computational heavy model.

Before proceeding with the analysis, a first, simple training was performed using BraTS 2020 data and the basic trainer (*nnUNetTrainerV2*) to understand computational time and costs. Training and validation performing cross-validation only on fold 0 (instead of all five folds) took about 24 hours using a small dataset (90 images for training and 30 for testing, where each image consists in four modalities). From the analysis of metrics and losses during training, it was possible to recognize that validation loss more or less stabilized after 100 epochs: from *Figure 4.56* it is possible to notice that even if the validation loss is still oscillating, the general trend is stable. To reduce training time and allow a comparison between well representative models, it was chosen to perform an initial study using this number of epochs (instead of 1000) and training only on fold 0. Furthermore, only the 3D full-resolution UNet was trained to additionally reduce training times, and because it represents the best configuration in the majority of cases. The network was trained from scratch, letting nnUNet learn directly from the dataset the optimal architecture.

As previously described in Chapter 2.3.2, nnUNet architecture is automatically adapted to the specific dataset; in *Figure 4.55* it's showed the structure of the best network generated for the BraTS 2020 challenge.

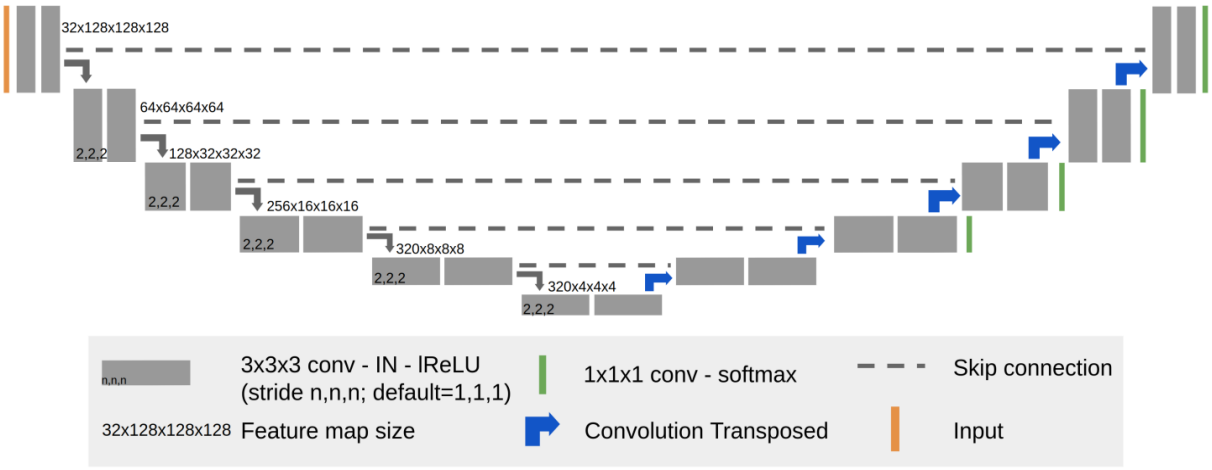


Figure 4.55: Network architecture generated from nnUNet from the BraTS 2020 dataset. Each gray block represents a 3x3x3 convolution, followed by instance normalization and Leaky ReLU; the input patch size is set to 128x128x128, and the initial number of feature maps is 32, doubled at each downsampling step. Feature maps sizes are showed only for the encoder, as they are symmetrically equal in the decoder. In this case the input is composed by 4 channels, one for each modality.

The performances using the basic trainer were also compared with a more BraTS-specific trainer, which was chosen to be *nnUNetTrainerV2_DA3_BN*. In this case a more aggressive data augmentation is performed, instance normalization is substituted with batch normalization, and batch dice loss is introduced. It was chosen not to apply region-based training and the postprocessing method from the modifications performed on nnUNet for the subscription to BraTS 2020 challenge, because they're specific changes which allow a better evaluation using BraTS metrics, and avoid obtaining small Dice scores, with the primary purpose of ranking in the best possible position in the challenge. The aim of this study is to develop a method directly applicable in clinics, able to exploit the dataset information to delineate lesions in the best possible way, with a minor computational cost and as fast as possible, so these modifications are not considered. The batch size was kept to two, because increasing to five would lead to small performances improvements, with large computational costs. The extensions introduced by H.M. Luu et al. aren't considered too, because they only slightly improved the performances, while widely increasing the computational time and complexity of the model.

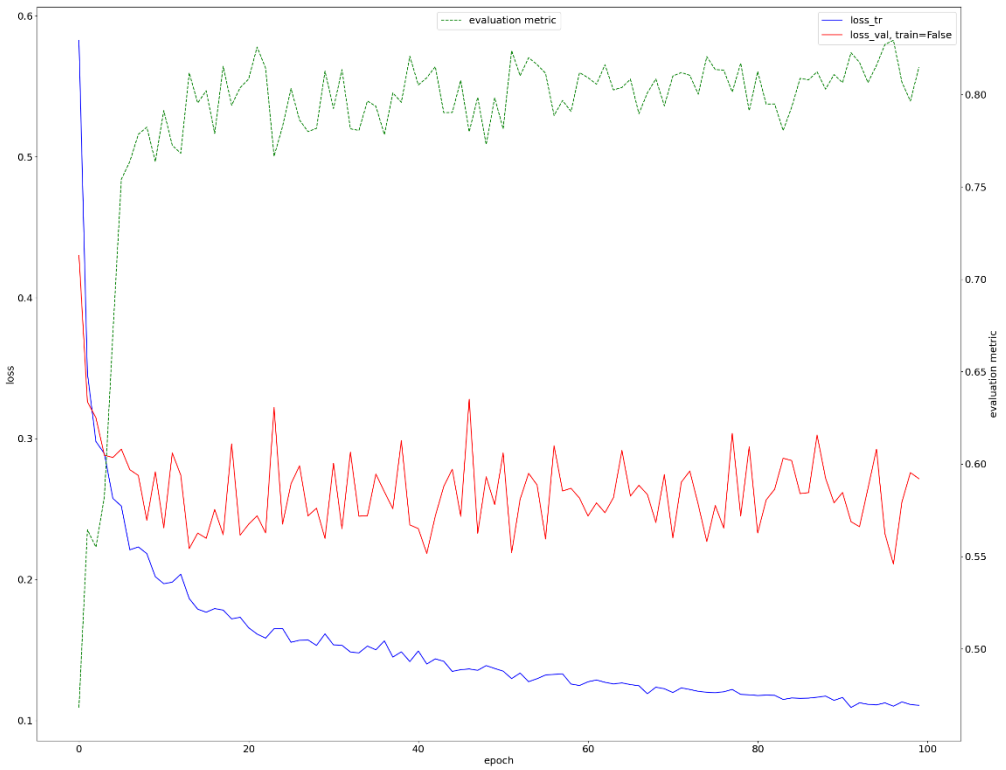


Figure 4.56: Trend of training (blue) and validation (red) losses for Task500, which uses the first 90 images of BraTS 2020 for training. It is evident that the model stabilizes after more or less 100 epochs, and can then be used for comparison with other networks. The green curve represents the average Dice score of the foreground classes; however, the same authors say that it isn't a reliable metric for model evaluation, because it is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.3 Introduction of Wassertian Dice Loss

nnUNet makes available different possible loss functions that could be used during training, like the General Dice Loss, Soft Dice Loss, Dice Loss based on Matthews Correlation Coefficient, Cross-Entropy Loss etc. But the best performing loss, and more suitable for brain tumors and stroke lesions segmentation task, is specified to be the combination between General Dice Loss (GDL) and Cross-Entropy (CE) loss.

A possible nnUNet modification which doesn't collide with the ideals of Isensee et al. and doesn't change significantly the computational cost is the exploitation of a different loss functions.

As previously stated, for the application of their method to BraTS 2021 challenge, L. Fidon et al. applied the Wassertian Dice Loss (Fidon et al., 2021), which they previously introduced in 2017 (Fidon et al., 2017) and exploited in the BraTS 2020 challenge and for other medical images segmentation tasks, showing superior segmentation performances when compared to the mean Dice Loss.

In particular, the generalized Wassertian Dice Loss, has been specifically designed for brain tumor segmentation task, because it takes advantage of the hierarchical structure of the set of classes in BraTS. The formula of the generalized Wassertian Dice Loss is:

$$\mathcal{L}_{GWDL}(\hat{p}, p) = 1 - \frac{2\sum_{l\neq b}\sum_i p_{i,l}(1 - W^M(\hat{p}_i, p_i))}{2\sum_{l\neq b}[\sum_i p_{i,l}(1 - W^M(\hat{p}_i, p_i))] + \sum_i W^M(\hat{p}_i, p_i)}$$

$$\forall i, \quad W^M(\hat{p}_i, p_i) = \sum_{l=1}^{L} p_{i,l}(\sum_{l'=1}^{L} M_{l,l'}\hat{p}_{i,l'})$$

Where L is the number of classes, i the index for voxels (N number of voxels), l the index for classes, $\hat{p}$ the predicted probability map, p the ground-truth probability map, $W^M(\hat{p}_i, p_i)$, is the distance between the predicted $\hat{p}_i$ and the ground truth p_i; b is the class corresponding to the background (0), while the peculiarity of the Wassertian Dice Loss is represented by the matrix M, which is a distance matrix between BraTS labels. Given the classes 0: background, 1: enhancing tumor, 2: edema, 3: non-enhancing tumor, M is set as:

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0.7 & 0.5 \\ 1 & 0.7 & 0 & 0.6 \\ 1 & 0.5 & 0.6 & 0 \end{pmatrix}$$

In this way, when the labeling of a voxel is too ambiguous, the matrix was designed to favor mistakes that remain consistent with the sub-regions used in the evaluation of BraTS (Fidon et al., 2021).

This represents just a particularization of the Wassertian Dice Loss for the application in the brain tumor segmentation field, but it was generally proposed for any multi-class segmentation problem, building the matrix M based on the Wassertian distance, which allows to represent the semantic relationship between classes as the comparison between label probability vectors (Frogner et al., n.d.), in the probabilistic label space (Fidon et al., 2017). Given these peculiarities of the generalized Wassertian Dice Loss, it doesn't make sense to apply it in the

segmentation of stroke lesions, because it's just a binary classification problem (0=background, 1=lesion), thus no matrix could be defined.

On the other hand, it could be easily applied as an alternative of the Dice Loss in the brain tumors segmentation problem, having a matrix already set for this aim.

nnUNet was then modified combining the Cross-Entropy Loss with the generalized Wassertian Dice Loss, using the experimental conditions previously mentioned. nnUNet was trained with a dataset of 90 images (each containing four modalities) extracted from the BraTS 2020 training set, and tested using 80 images randomly extracted from FeTS 2022 training set, so that training and test set used are completely independent, allowing a more robust and reliable evaluation.

As showed in *Table 4.1,* the introduction of Wassertian Dice Loss is not improving the performances when using the classic nnUNet training pipeline, because it achieves an average Dice Loss of 0.613, against the 0.658 of the classic model with general Dice Loss. Instead, when coupling the Wassertian Dice Loss with BraTS-specific trainer and modifications, the altered loss function is able to improve the results, obtaining a higher Dice Score for all three classes.

a)

	CE + Dice	CE + Wassertian Dice
ED	0.740	0.732
NCR	0.500	0.425
ET	0.733	0.683
Mean	0.658	0.613

b)

	CE + Dice	CE + Wassertian Dice
ED	0.732	**0.751**
NCR	0.426	**0.506**
ET	0.683	**0.720**
Mean	0.614	**0.659**

*Table 4.1: Comparison of the Dice scores obtained in the segmentation of the three classes (ED stands for Enhancing Tumor, NCR for Necrotic tumor Core, ET for peritumoral Edematous tissue), and their average, when using Cross-Entropy (CE) loss combined with Dice loss, or CE combined with Wassertian Dice loss. Using classic nnUNet trainer (nnUNetTrainerV2), the segmentation results don't improve **(a)**, while using BraTS specific trainer (nnUNetTrainerV2_DA3_BN) the performances increase on single classes and on average **(b)***

It follows that Wassertian Dice loss can be used as a valid alternative to classic Dice loss, when using nnUNet model in the brain tumor segmentation task, only when coupled with an appropriate training pipeline, which benefits of an increase data augmentation and the usage of batch normalization, that privileges and is more suitable with brain tumor segmentation task and data.

4.4 Analysis of the dependency of nnUNet training from different input modalities

nnUnet works by treating multi-modal images as color channels. In this way, different modalities of the same image are processed together by the network, which is fusing their information at early stages and extracting features from their conjunction. However, different modalities can contain really dissimilar, and in some cases totally opposite, information. In T1-weighted images, for example, white matter appears white, gray matter looks gray, while cerebrospinal fluid (CSF) appears dark; on the other side T2-weighted images cause the white matter to look more gray, gray matter to appear white and CSF to appear lighter (*Figure 4.57*), so they basically provide opposite information, highlighting regions in a completely different way. On one hand it makes sense feeding a neural network with multiple modalities, so that it can learn to identify the lesion (or tumor) from images in which it is emphasized in contrasting manners, but on the other hand using single modalities could avoid mixing contrasting information, and help extracting more intrinsic features of the image and especially of the lesion.

For these reasons, both for stroke lesions and brain tumors segmentation, it has been performed a single modality analysis, which consists in feeding the network (nnUNet) with images of a single modality at a time, and compare the segmentation results, to examine the dependency of the network training from the modalities, and identify possible modalities from which the network is better able to localize the lesion. For each analysis, it was necessary to create a new dataset (and so, a new task) with a single modality with ID 0000, specify in dataset.json the name of the modality, and nnUNet was able to extract all necessary information, adapting the framework and considering one single channel per time.

An ulterior analysis was also performed, training nnUNet by leaving out one modality at a time, allowing therefore the network to learn from all but one modality. This study was carried out to further optimize the identification of the dependency of the segmentation results produced by nnUNet by the different modalities, to point out possible modalities which are more remarkable, or other that are substantially useless, for the different tasks of brain tumors and stroke lesions segmentation.

4.4.1 Brain Tumor Segmentation data analysis

The investigation about the dependency of nnUNet from single modalities started from the brain tumor segmentation field. The representative dataset from which training images where extracted was BraTS 2020 dataset, from which the first 90 subjects (for a total of 360 images, considering the four modalities) were used for training these models. While, to encourage a representative and effective analysis, escaping overfitting due to the excessive similarity of training and testing images, 80 subjects were chosen from a completely different dataset, FeTS 2022, by extracting randomly images provided by different centers. In this way, by avoiding using images coming from the same dataset both for training and testing, it is possible to replicate a real-life application, in which these methods could be exploited to segment images generated from direct examinations, with nothing in common with the ones used for training. Moreover, it was chosen to randomly draw images from different providing centers to further increase their variability and the generalizability of the results.

As stated before, each input consists in four modalities: FLAIR, T1, T1ce, T2. In *Figure 4.57* it can be appreciated the difference between how each modality highlights the different brain regions and especially the tumoral portions, underlining the need to study separately the distinct modalities. In particular, from the visualization of the corresponding ground truth labels, it is possible to notice the intake of the information from the different modalities in the delineation of the distinct tumoral regions: the tumor core (necrotic tumor core and enhancing tumor) is more visible in T1ce scan, while the precise contours of the peritumoral edematous/invaded tissue are more visible in FLAIR scan.

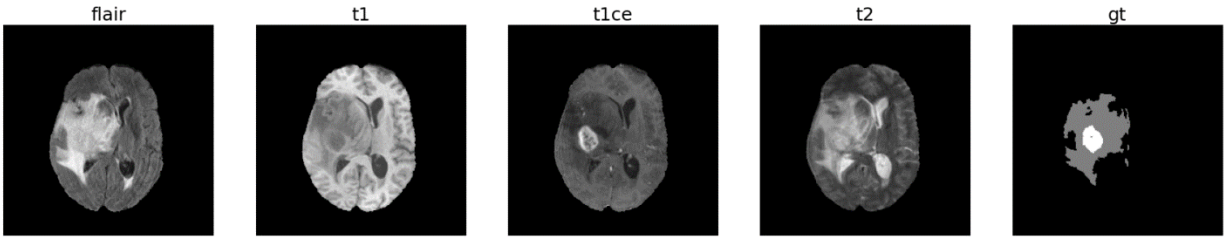


Figure 4.57: Available modalities (flair, t1, t1ce, t2) and corresponding ground truth segmentation for the first image (BRATS_001) in BraTS 2020 dataset. It's immediate to notice how the different brain regions, and especially tumoral regions, appear distinct, and in some cases opposed, in different modalities, providing complementary information.

First of all, a study removing one modality at a time was performed.

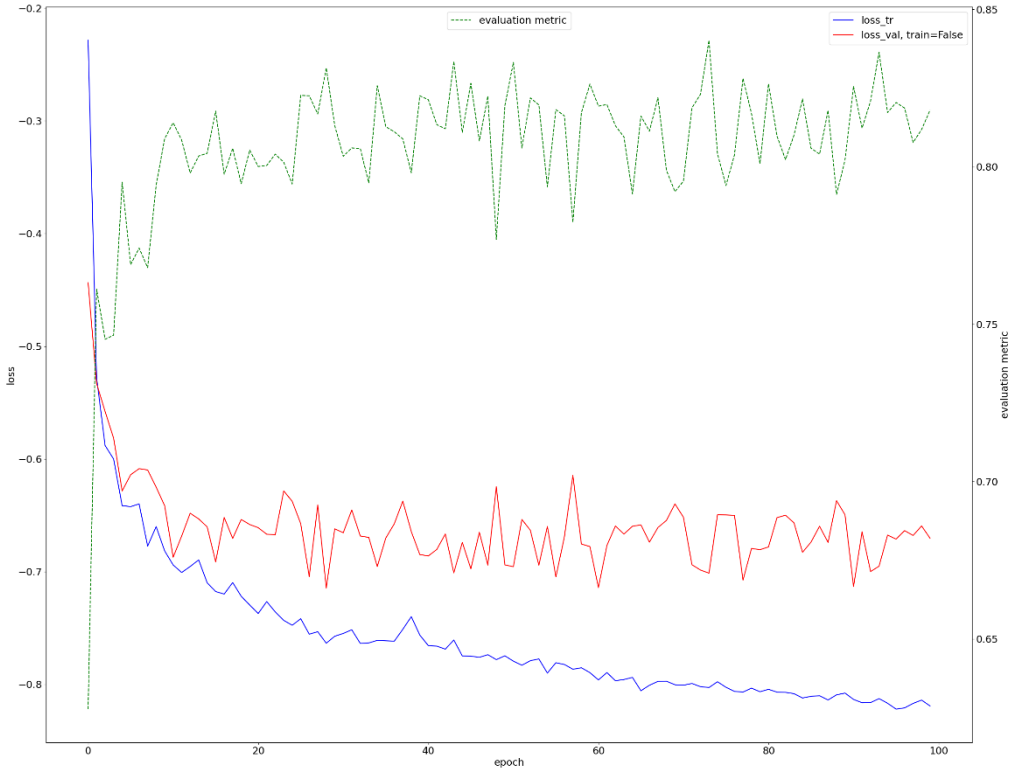
4.4.1.1 Training performed with all but one modality

The first research was performed by training models removing one modality at a time. Four new datasets were created, following the rules of the baseline model chosen, nnUNet, each containing all but one modality: the first model was trained removing FLAIR modality, the second was trained taking away T1 images, the third eliminating T1ce, and the fourth removing all T2 images. This study was performed to investigate how the segmentation results produced by nnUNet depend from the diverse modalities, with the aim of identifying possible modalities not useful for the task, and whose removal doesn't affect the performances of the model. To quantify the intake of each modality, all these networks were obviously compared with the complete model, which uses all modalities for training.

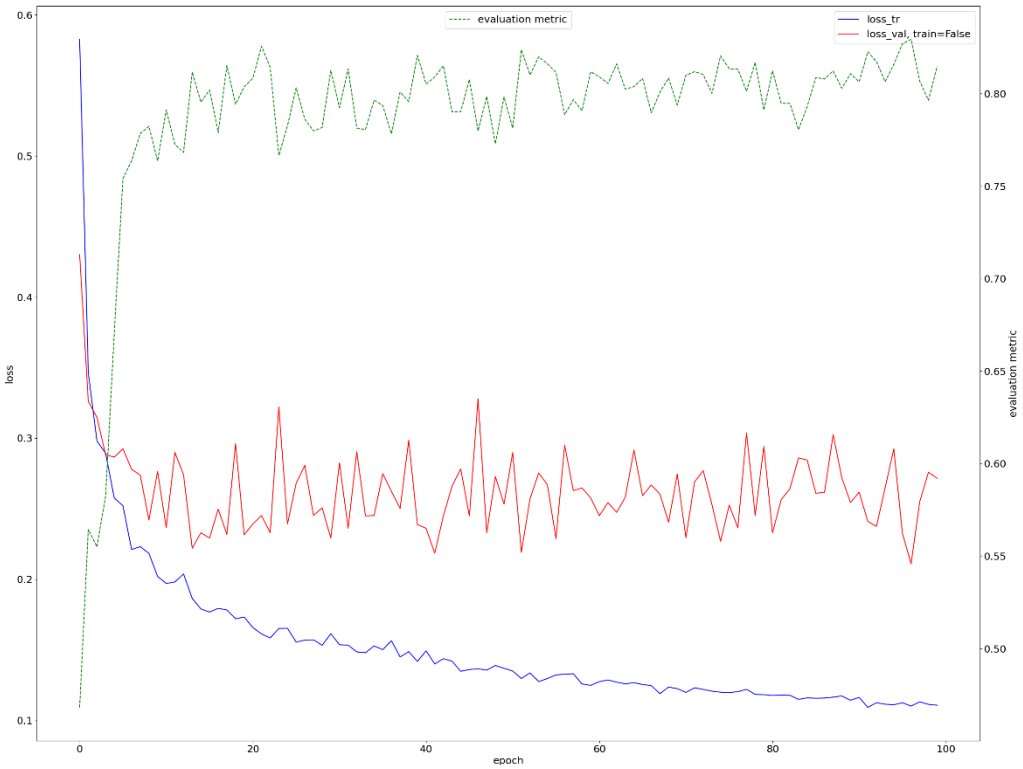
The progress of training and validation losses during training for all tasks can be visible in *Figures 4.58, 4.59* and *4.60*, from which it's possible to notice some peculiar characteristics: first of all, in all cases the validation loss is becoming more or less stable after 100 epochs; it's not possible to know if it will continue to decrease afterwards, but this demonstrates that 100 epochs can be used as a good comparison value between testing models. There aren't signs of overfitting or underfitting, but the validation loss shows noisy movements, oscillating a lot with respect to training loss: this can originate from the training-validation split, which for nnUNet is set to 80:20; considering that the training dataset is just composed by 90 images, only 18 are used for validation at each step. This small number of images can justify the oscillating trend of the validation loss, and also for this reason validation results weren't used for the comparison between models.

After the end of the training of all models, their performances were compared by testing on the same 80 images extracted from the FeTS 2022 dataset, and evaluating the results.

(a) Training and validation losses for the full model

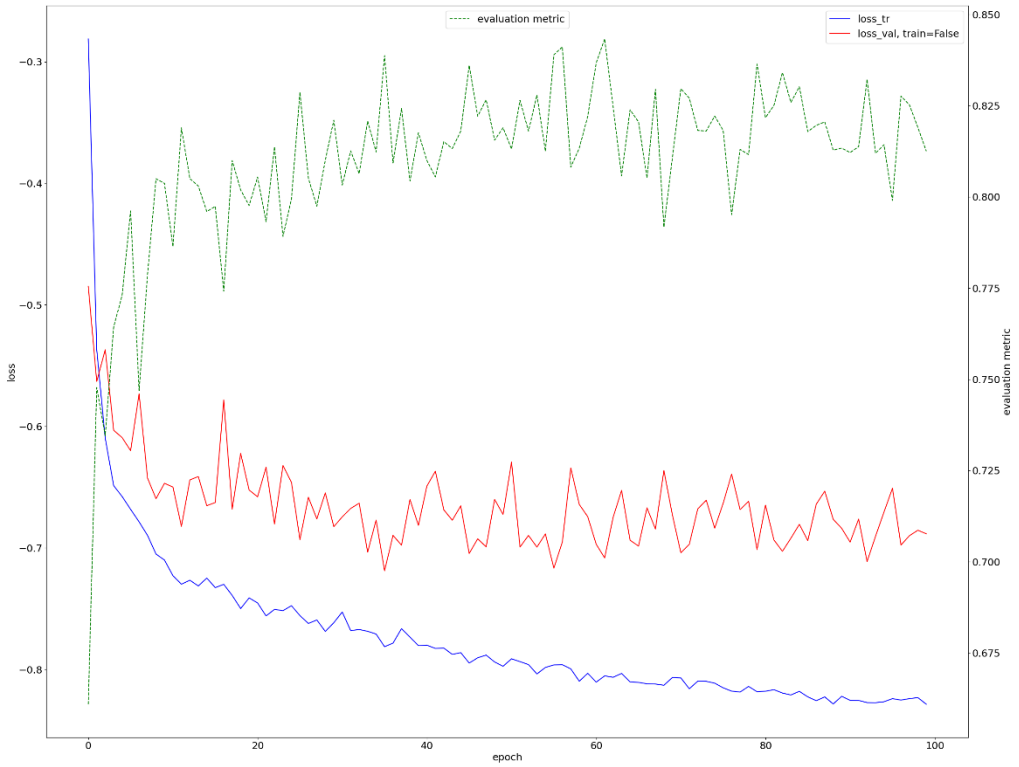


(b) Training and validation losses for the model trained with all but FLAIR images



Figures 4.58: ***(a)*** *Trend of training (blue) and validation (red) losses for the full model, trained with all available modalities, and for the network in which the training dataset is composed by all modalities but T2* ***(b)****. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.*

(a) Training and validation losses for the model trained with all but T1 images



(b) Training and validation losses for the model trained with all but T1ce images

*Figure 4.59: **(a)** Trend of training (blue) and validation (red) losses for the model trained with all but T1 modality, and for the network in which the training dataset is composed by all modalities but T1ce **(b)**. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.*

Training and validation losses for the model trained with all but T2 images

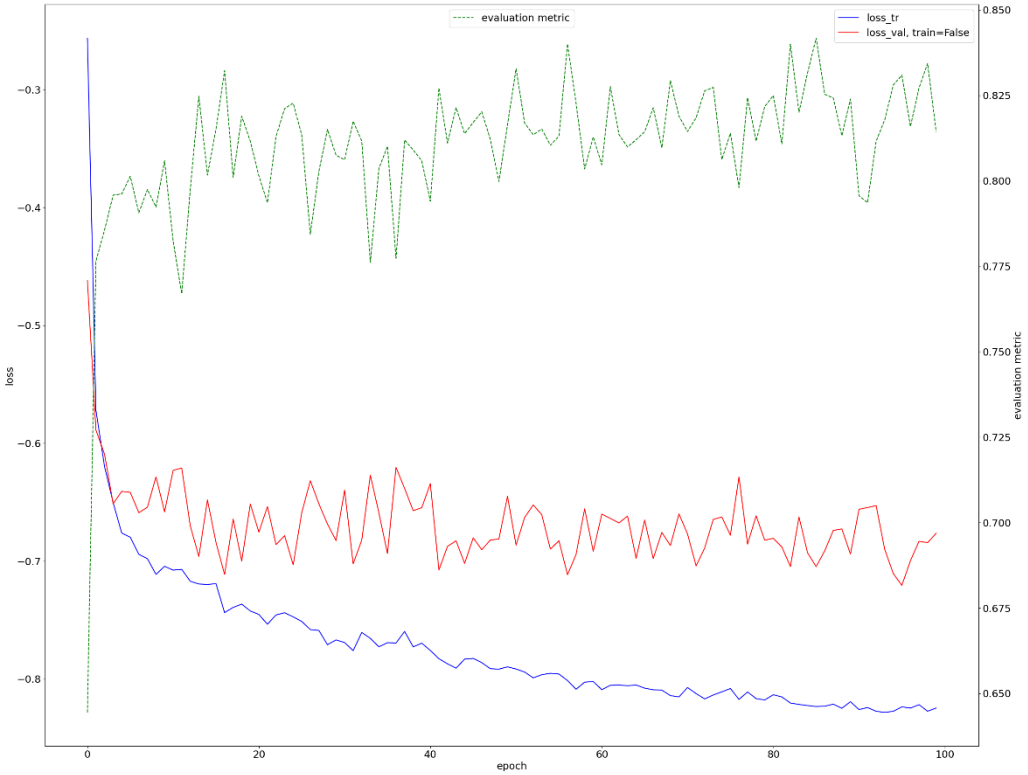


Figure 4.60: Trend of training (blue) and validation (red) losses for the model trained with all but T2 modality. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.4.1.2 Training performed with a single modality

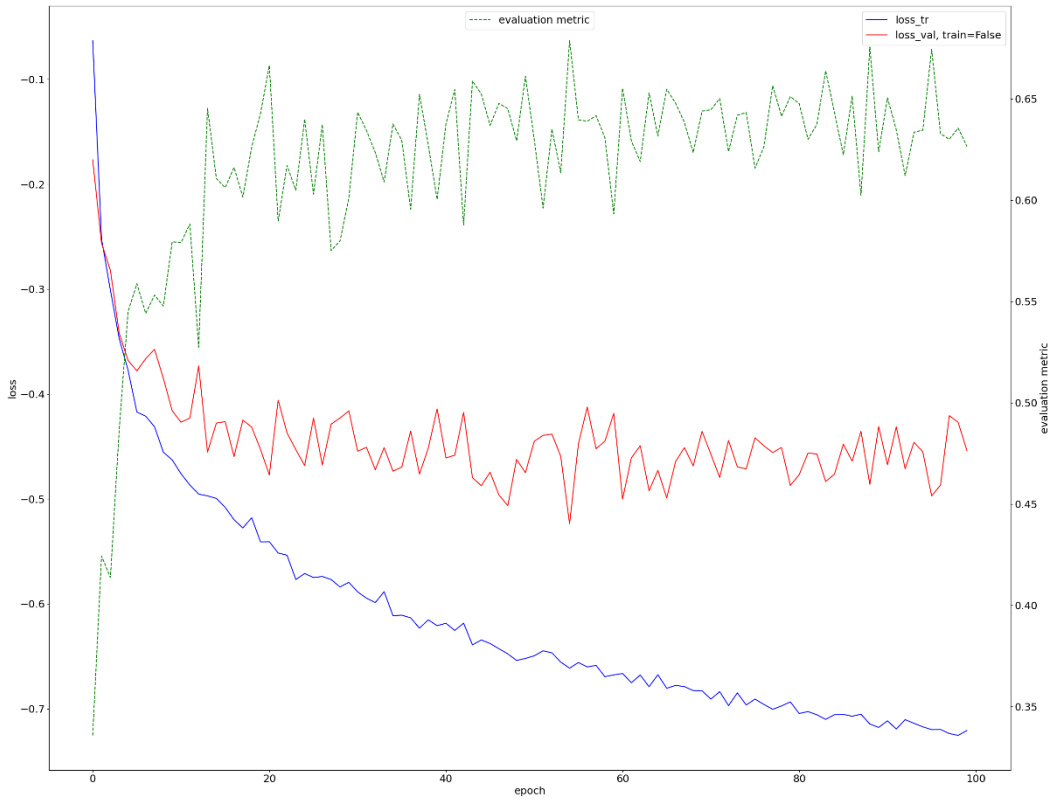
Following the pipeline of the previous analysis, to investigate the dependency of nnUNet from input modalities it was chosen to train different models, each with a single input modality. For this purpose, four new datasets were created, following the rules of nnUNet, and used to train four distinct models: the first one was trained using only FLAIR images; the second model using only T1 images; the third networ employing only T1ce modality; and the last model utilizing only T2 images.

As before, the performances of what were called 'single' models (trained with a single modality) were compared with the results of the 'full' model (trained with all modalities). This study was performed to further point out possible modalities which are more relevant than others for the segmentation of the brain tumor's regions, providing results closer to the full model, and thus capable to segment in an efficient way even if a single image modality is used.

On the other hand, all models trained with single modalities were expected to perform quite poorly with respect to the full model, because they were trained with a quarter of the images used for the complete model, and because they couldn't benefit from the fusion of features extracted from modalities providing different information.

The progress of training and validation losses during training can be appreciated in *Figures 4.61* and *4.62*; the progress of the full model is the same as exposed in section 4.4.1.1 so it's omitted. The same considerations as before can be made: the threshold of 100 epochs remains a good comparison value between models, while the trend of training and validation losses is really similar to the models trained with all but one modality, so can be considered acceptable.

(a) Training and validation losses for the model trained with only FLAIR images



(b) Training and validation losses for the model trained with only T1 images

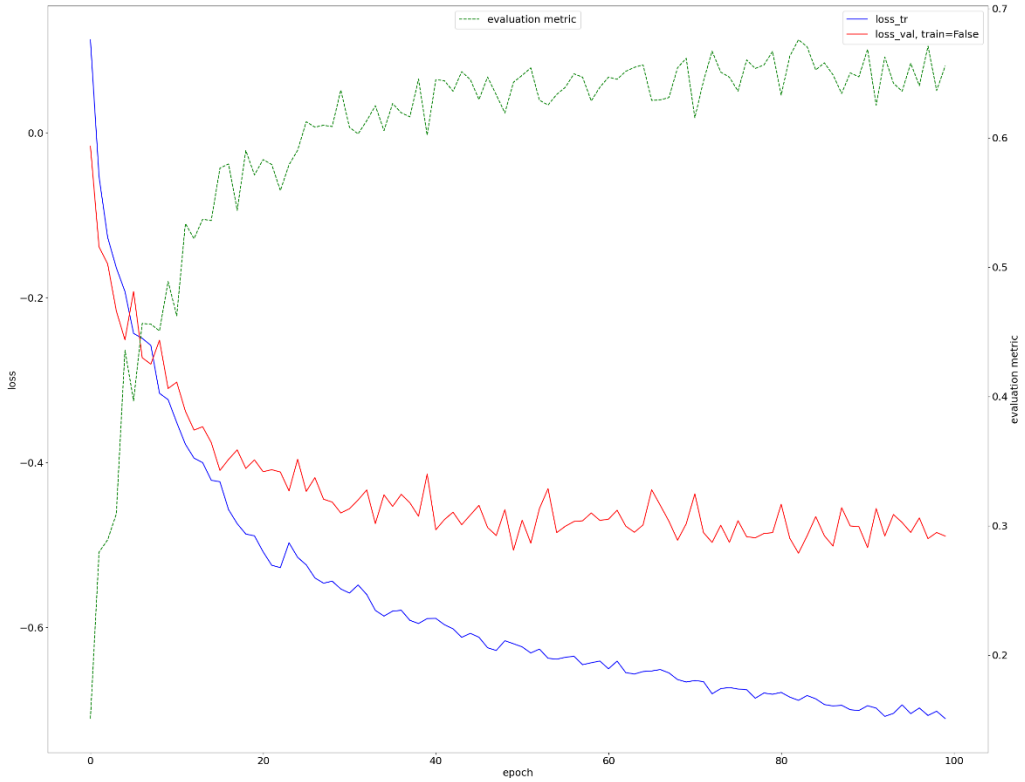
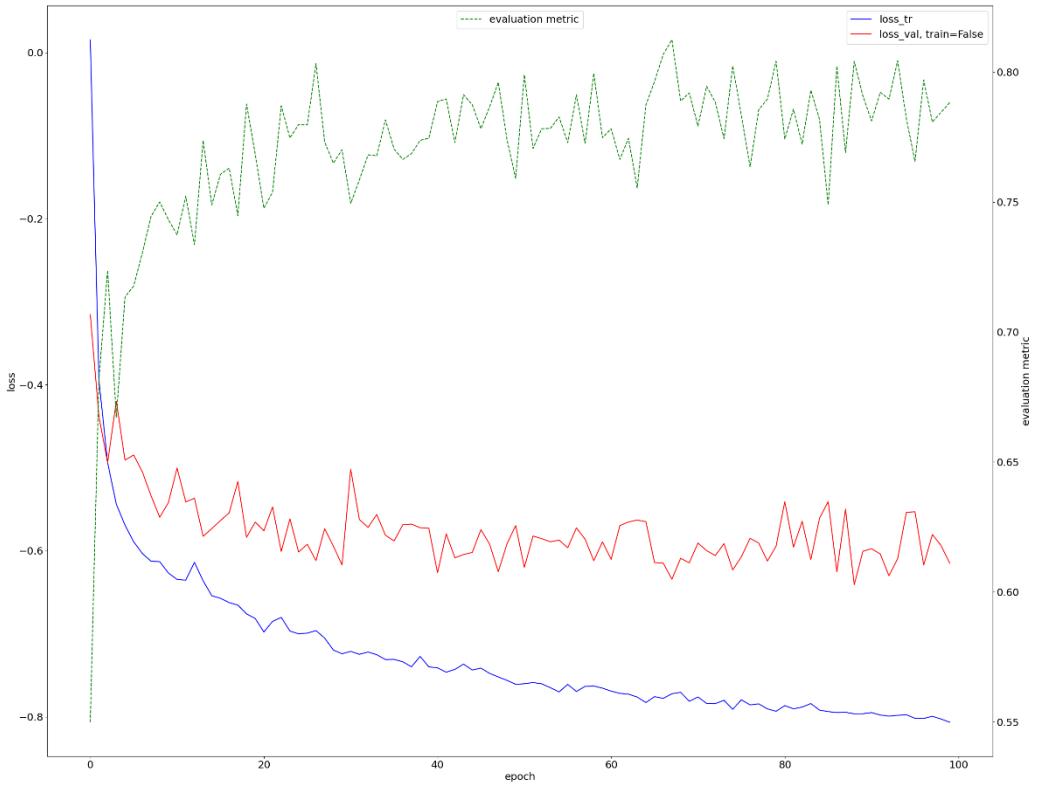


Figure 4.61: Trend of training (blue) and validation (red) losses for the network in which the training dataset is composed by only FLAIR modality ***(a)*** *and for the network in which the training dataset is composed by only T1 images* ***(b).*** *As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.*

(a) Training and validation losses for the model trained with only T1ce images



(b) Training and validation losses for the model trained with only T2 images

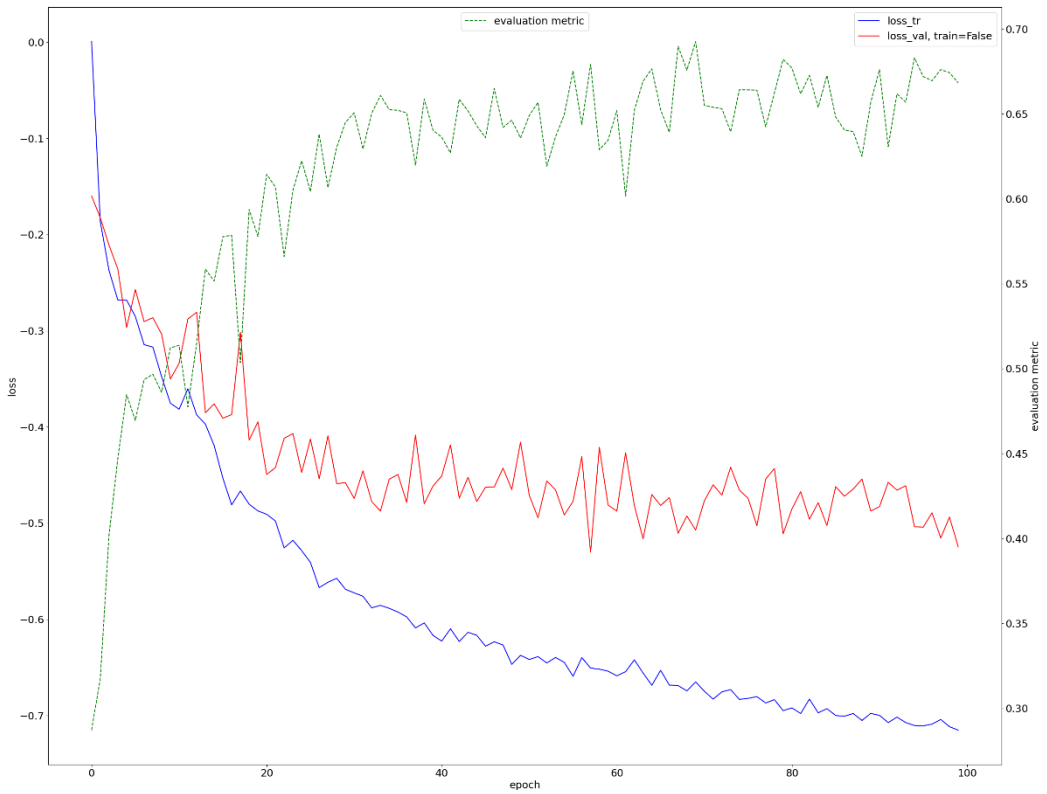


*Figure 4.62: Trend of training (blue) and validation (red) losses for the network in which the training dataset is composed by only T1ce modality **(a)** and for the network in which the training dataset is composed by only T2 images **(b)**. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.*

4.4.1.2.1 Equalization of computational times

As emerged from the previous investigations, models trained with single modalities are not able to reach the performances of the full model, trained with all available images. One of the motivations behind this is that the number of images used to train the full model is four times the number used for models trained with images of a single modality, and so it has a lot more inputs to learn from. A necessary step to correctly compare the results obtained by single models and by the complete model is to equalize their computations. There are tons of possible ways to make models comparable, a first method that was chosen to be exploited was to equalize the computational times between models.

Before doing that, training of networks was modified adding some lines of code to track the total time required. It was therefore observed that the time required for training models using images of a single modality was 2 hours and 19 minutes for the model trained with only FLAIR images, 2 hours and 26 minutes for the network trained using T1 scans, 2 hours and 35 minutes for the model using only T1ce modality and 2 hours and 24 minutes for the architecture learning only from T2 images. To complete the research, the time required by the full model, trained with all image modalities, was 6 hours and 47 minutes.

To make the segmentation results of the models comparable, it was chosen to compute the average training time of single models, which was estimated to be 2 hours and 26 minutes, and stop the training of the complete model, when that threshold training time was reached.

As can be visualized in *Figure 4.63*, the full model was thus stopped after 50 epochs. Moreover, it's possible to notice that even if the number of epochs is halved with respect to the standard value used for evaluation, the trend of validation loss can be considered analogous with the regular cases, making the results comparable.

At the end, an inference was performed on the common test set composed by 80 images extracted from FeTS 2022 dataset, and the Dice scores obtained by different models were compared.

Training and validation losses of the full model stopped at 50 epochs

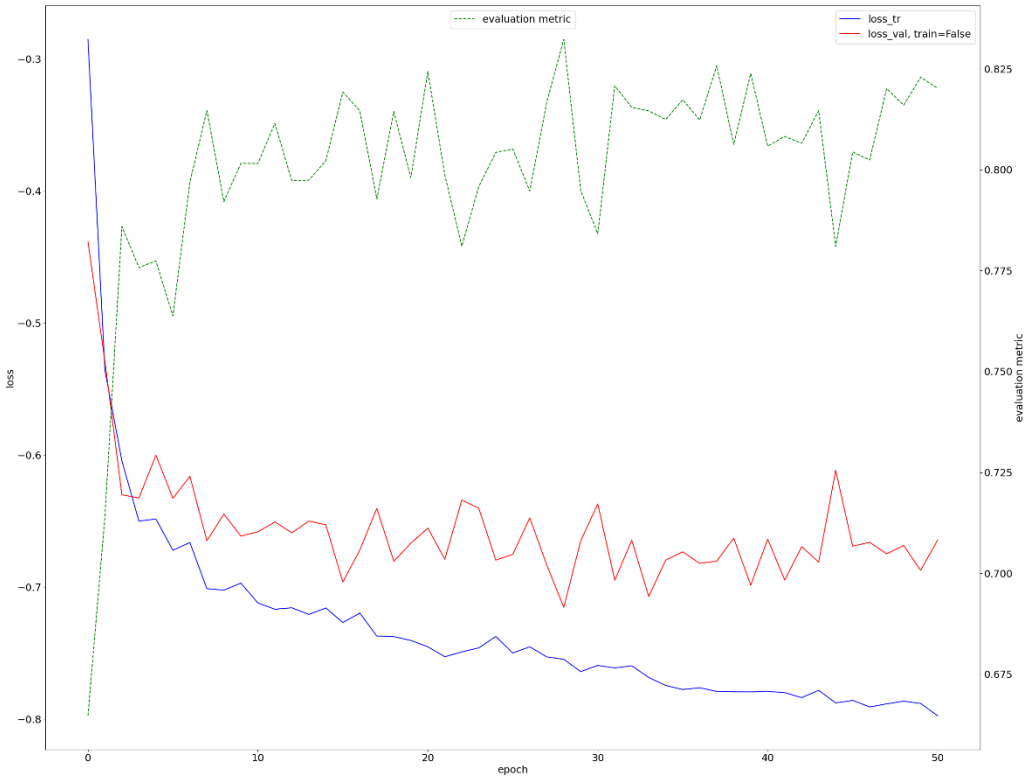


Figure 4.63: Trend of training (blue) and validation (red) losses for the network trained with all available images, but stopping its training at the time employed by models trained with a single modality. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.4.1.2.2 Equalization of number of training images

The alternative method that was chosen to be exploited to equalize the computations between models trained with a single modality and the full model, is the equalization of the number of training images. As previously stated, as long as the full model was trained with all available modalities (FLAIR, T1, T1ce, T2), the number of images used for its training is four times bigger than the quantity of images used to train models exploiting one single modality. For this reason, to equalize the computations between networks and allow single models to learn useful features from the same amount of images used by the full model, it was chosen to equalize the number of training images for all configurations. To be more specific, the number of subjects used to train the complete model was 90, each consisting of four modalities, for a total of 360

images contributing in the training of the networks. It was then chosen to create four new datasets, each with a corresponding TaskID, where each dataset was composed by 360 images of a single modality extracted from the training set of BraTS 2020 dataset.
In this way, it was given the possibility to each single model to learn from the same quantity of images with respect to the complete model, and segmentation results could be more comparable than previous cases.
As an example, in *Figure 4.64* it's reported the trend of training and validation losses for the model trained only with T1 modality, while the other single models have a really similar course. It can be seen that the model is pretty good, since the training and validation losses decrease for a larger amount of time, meaning that the model is learning for more time with respect to the previous cases, having more available images from which it can learn useful information. Also, the gap between training and validation losses is smaller, and the oscillations of the validation loss decrease, given that the number of validation images is higher, consisting of 72 images, and is therefore capable of providing informative and exploitable results.
After training models, inference was performed by testing their results on the same 80 images extracted from FeTS 2022 dataset and comparing the obtained average Dice scores.
To conclude this study and verify the reproducibility of the results, the same analysis was performed reducing in a proportional way the number of images used to train each model: starting from 360 images for single models and 90 subjects for the full one (examined in this case), passing to 176 images for single models, together with 44 subjects for the complete one, then 88 images for single models and 22 for the full one, after 40 and 10 and to conclude 20 and 5. This examination was done to check if segmentation results were consistent and to analyze how they changed with respect to the number of training images.

Training and validation losses for the model trained with 360 T1 images

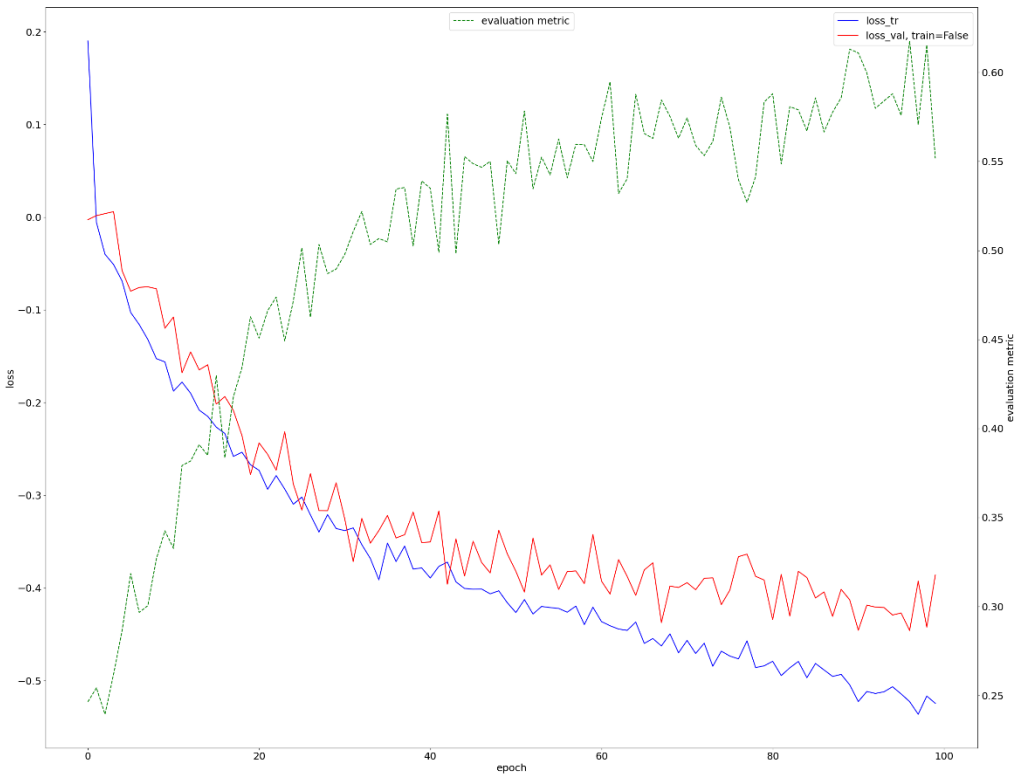


Figure 4.64: Trend of training (blue) and validation (red) losses for the network trained only with T1 images, but quadrupling their number, from 90 to 360. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.4.2 Stroke Lesion Segmentation data analysis

The same analysis with mutual purposes was realized on stroke lesion segmentation data. The dataset of ISLES 2022 challenge was used, because it entails the most advanced and recently released data in this field. ISLES 2022 challenge is the current year competition, so to access the data it was necessary to sign up to the challenge, but without submitting any method, with the only purpose of retrieving and study the data. The training data were released on 10th of May 2022, and they were requested and obtained on 15th July 2022.

For subsequent purposes, the number of training images chosen for this research was 70, while 40 images were chosen to compose the test set. The testing images were also extracted from the ISLES 2022 dataset, because the other more reliable ischemic stroke lesion segmentation

datasets are composed by different modalities, so they are not helpful for the current study. For this reason, with respect to the previous evaluation done on BraTS data, the testing results were expected to be closer to the validation results, because the images used to test the model are extracted from the same dataset of the training images (unlike for the brain tumor segmentation study), so the results will be less generalizable with respect to the previous case.

In this case, each image consists in three modalities: ADC, DWI and FLAIR, that can be visualized in *Figure 4.65*. Also in this case it is evident how the different modalities highlight cerebral and lesion regions in distinct ways: while flair images were already met in brain tumor segmentation task, DWI and ADC images are more informative for stroke lesions. DWI exploits the diffusion of water molecules, which can be quantitatively assessed using the apparent diffusion coefficient (ADC) value, calculated and displayed via parametric maps. By construction, the ADC value is opposite to DWI, because in ADC high values reflect high diffusivity of water and the opposite, while in DWI images the lower the degree of diffusion of a molecule, the higher (lighter) the signal. For these reasons, as can be seen in *Figure 4.64*, an ischemic stroke lesion hinders the diffusivity of water, reducing ADC, while appearing really bright in DWI. The lesion is then highlighted in different, and in some cases opposite, ways, according to the modality.

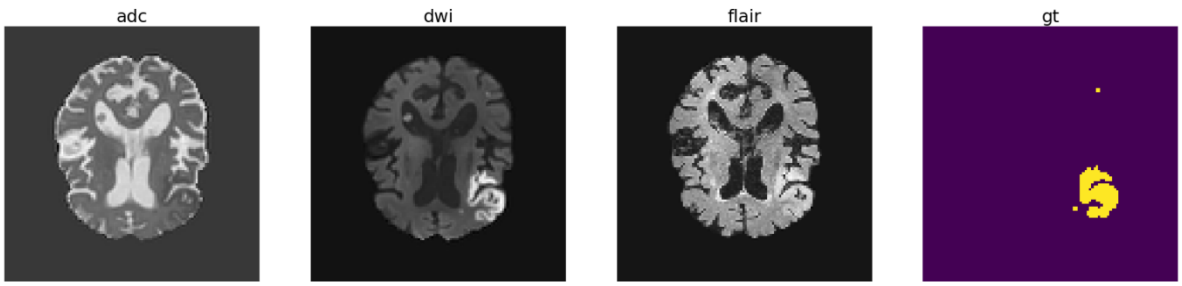


Figure 4.65: Visualization of a specific slice of ADC, DWI and FLAIR images of case 3 from ISLES 2022 dataset. The corresponding provided ground truth segmentation is also showed. From this case, which was chosen because of the magnitude of the stroke lesion, that can be easily seen, it's possible to appreciate how the lesion is differently highlighted in distinct modalities, being much more visible in DWI image.

Also in this study, a basic (full) model was created, associated to Task700 and containing all the modalities. A preliminary preprocessing of this dataset was sent to investigate its characteristics and evaluate its integrity. Unfortunately, unlike BraTS and FeTS datasets which are provided ready to be used, ISLES images are not coregistered. In particular, ADC and DWI images have the same dimensions as the ground truth segmentation, which is variable between images but usually has size 73x112x112; while flair images come in a totally different size,

which is really incompatible between provided images, for example in the first patient, its size is 352x352x281, while in the seventh patient it's 230x352x352. It can be also observed that in some cases the number of slices is the first dimension, while in others it's the last dimension. Moreover, because of the strength of nnUNet technique, it was also possible to detect and quantify a mismatch between origins, spacings and directions of all images (not only FLAIR2) with respect to the masked segmentations.

As a first step, flair images were transposed so that they could match segmentations, having the number of slices as first index in all cases. After that, resizing images would not be sufficient because the mismatch with respect to the segmentations would remain. For this reason, it was chosen to coregister the images of the three modalities with respect to the ground truths, which haven't been changed at all; not only FLAIR images, but also ADC and DWI were fixed, to correct their geometry. For the coregistration, it was used SPM12 software in MATLAB, arranging one patient at a time to the corresponding segmentation. It was automatically performed by finding parameters that either maximize or minimize an objective function; because it consisted in a multi-modal registration, the preferred and default function was Normalized Mutual Information. The images were then interpolated using 4^{th} Degree B-Spline, which is slow but more accurate because it considers more neighbors.

At the end of this process, corrected images were saved according to nnUNet's canons and they were ready to be used.

To start, a study removing all modalities but one was performed.

4.4.2.1 Training performed with all but one modality

The analysis performed for Brain Tumor Segmentation was replicated on Stroke Lesion Segmentation: three new datasets were created, each consisting of images of all but one modality, and three new models were trained from them. Each dataset was built and organized coherently with nnUNet rules, so that it could be able to automatically extract datasets information and adapt its architecture: the first model was trained with all but ADC modality; the second network, was trained without considering DWI images; while the last was trained removing FLAIR images. The segmentation results where then compared between each other and with the full model, trained with all modalities.

This study was performed to identify possible modalities unnecessary or limitedly useful for the segmentation of stroke lesions, analyzing how the segmentation results are dependent from the specific modality, and how the generated maps are affected by their selective remotion.

Before proceeding with the study, a previous investigation was performed on the full model to identify possible modifications of the training procedure: as it can be seen in *Figure 4.66,* training and validation losses have a different trend with respect to brain tumor segmentation models. It must be specified that, unlike brain tumor segmentation problems, stroke lesions tasks are characterized only by two classes (0: non-lesion, 1: lesion) and the number of modalities is lower; for these reasons, training took significantly less time with respect to the previous analysis. It was then attempted to increase the number of epochs to 200, because training times continued to remain low, and could thus allow a more precise comparison between testing models. In reality, doubling the number of epochs increased the performances on the test set of barely 0.03 (average Dice score) while, obviously, doubling the training time; it was then decided that it wasn't worth it. Moreover, it was also tried to implement the BraTS specific settings (increased data augmentation, batch normalization instead of instance normalization), to identify if they were able to improve the performances also for the segmentation of stroke lesions. In this case, the average Dice score on the test images increased of 0.02, while almost doubling computational times; it was thus chosen not to use this configuration and to remain to the basilar architecture.

Training and validation losses for the model trained with all modalities

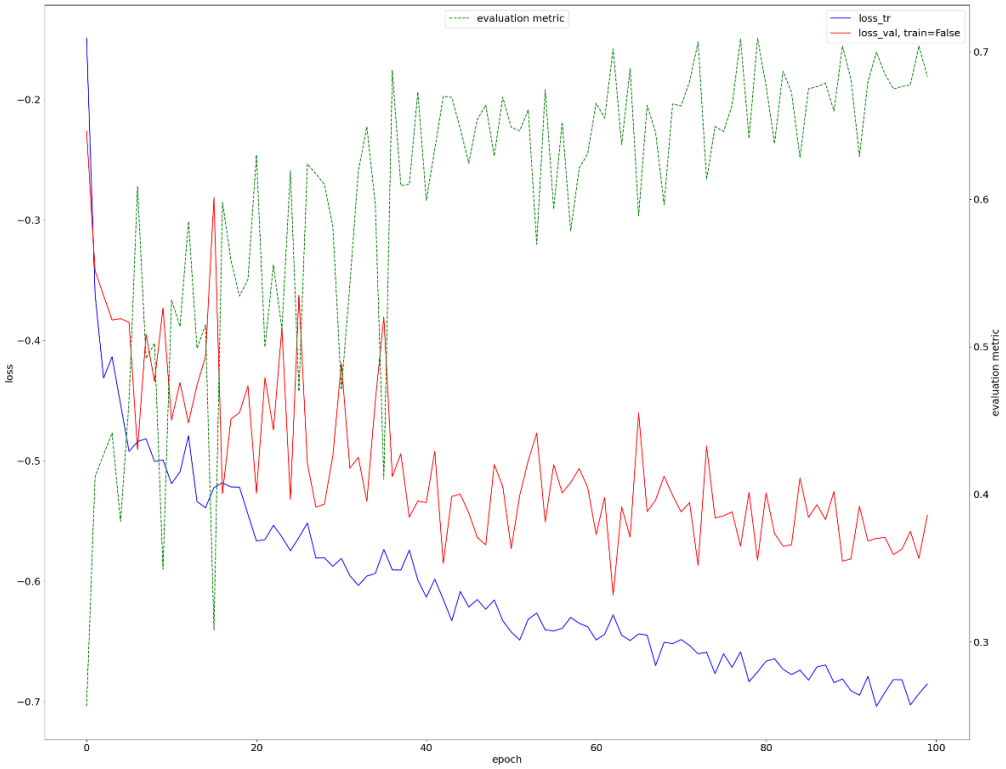
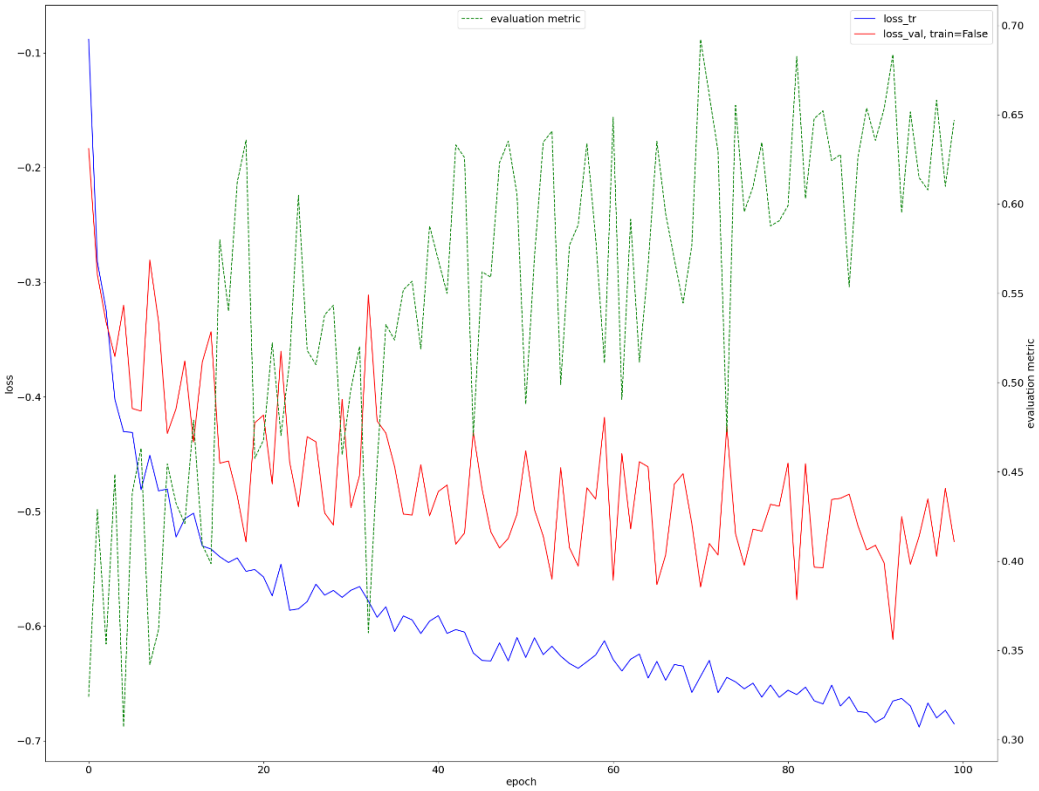


Figure 4.66: Trend of training (blue) and validation (red) losses for the full model, in which the training dataset is composed by all modalities (ADC, DWI, flair). As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

In *Figures 4.67* and *4.68* it's possible to visualize the trend of training and validation losses for models trained with all but one modality. A required observation is that the validation loss oscillates much more than the brain tumor segmentation cases, and the reason can be found in the number of training cases: as pointed out before, the number of chosen training images was 70, and 20% of them were used for validation. It means that only 14 images were used to validate the model, and this really small number explains why validation results fluctuate so much and aren't highly reproducible.

(a) Training and validation losses for the model trained with all but ADC images



(b) Training and validation losses for the model trained with all but DWI images

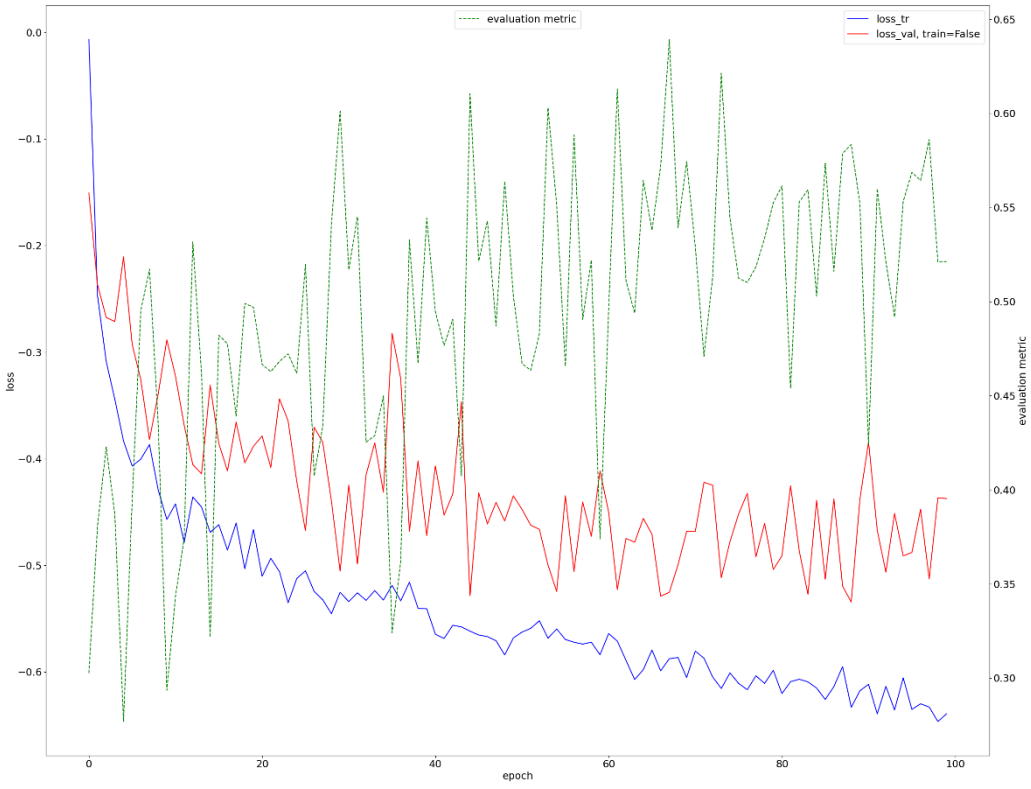


Figure 4.67: Trend of training (blue) and validation (red) losses for the model trained with all but ADC images ***(a)****, and for the network trained removing DWI modality* ***(b)****. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.*

Training and validation losses for the model trained with all but FLAIR images

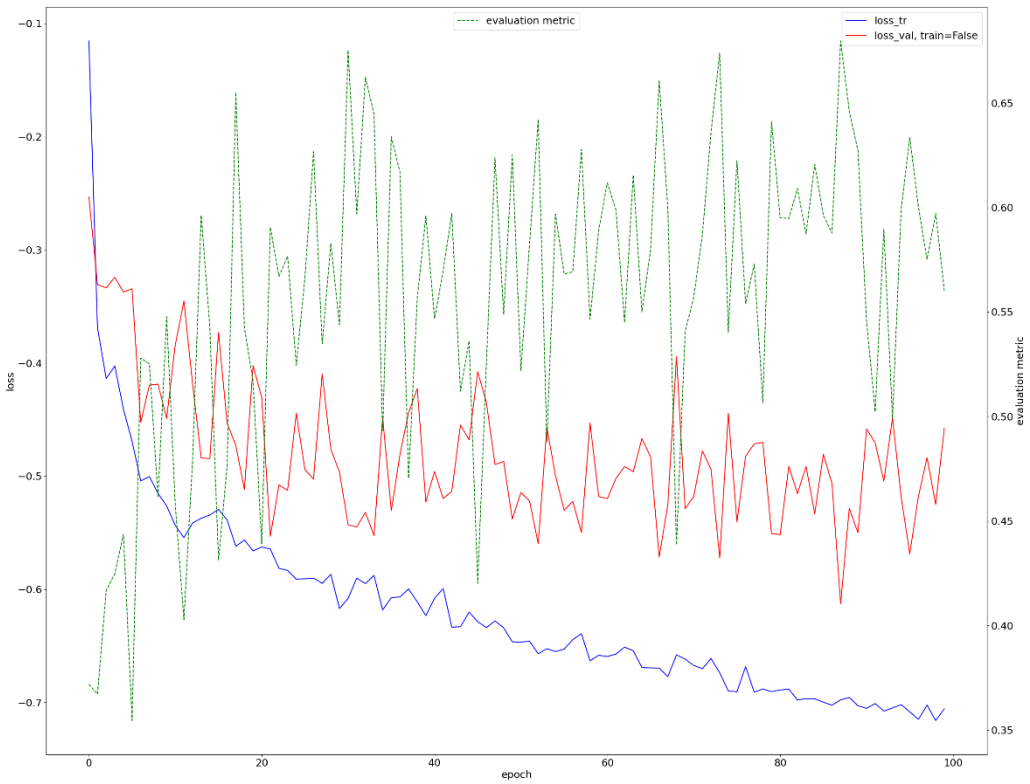


Figure 4.68: Trend of training (blue) and validation (red) losses for the model trained with all but FLAIR images. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.4.2.2 Training performed with a single modality

Following the same procedure than brain tumor segmentation analysis, a research about the dependency of nnUNet from single modalities was performed also in the stroke lesion segmentation field. In particular, three new models were designed, each trained with images of a single modality, and their performances were compared between each other and with the full model, trained with all provided modalities. For the training of each single model, three new datasets, associated to a corresponding task ID, were generated, following as before the rules set up by nnUNet: the first dataset was composed by only ADC images, the second one constituted by only DWI scans, and the last one was formed by only FLAIR images.

This study was performed to complete the previous analysis for the identification of the most relevant modalities for nnUNet when employed for ischemic stroke lesions segmentation. The

expectations were that models trained with images of a single modality would perform poorly with respect to the full model, on one side because they can learn useful information from one third of the images used by the model trained with all modalities, so the size of the training set is considerably lower, on the other because they couldn't exploit the fusion of deep features extracted by different modalities, which could underline different lesion characteristics and provide complementary information.

In *Figures 4.69* and *4.70* the trends of training and validation losses for the different tasks are depicted; the same considerations of previous cases can be carried out, giving special attention on underlining that the excessive oscillations of the validation losses are probably due to the very limited number of images used for validation. Validation results were then chosen not to be showed, focusing more on test set results, which are more reliable and more generalizable.

Training and validation losses for the model trained with only ADC images

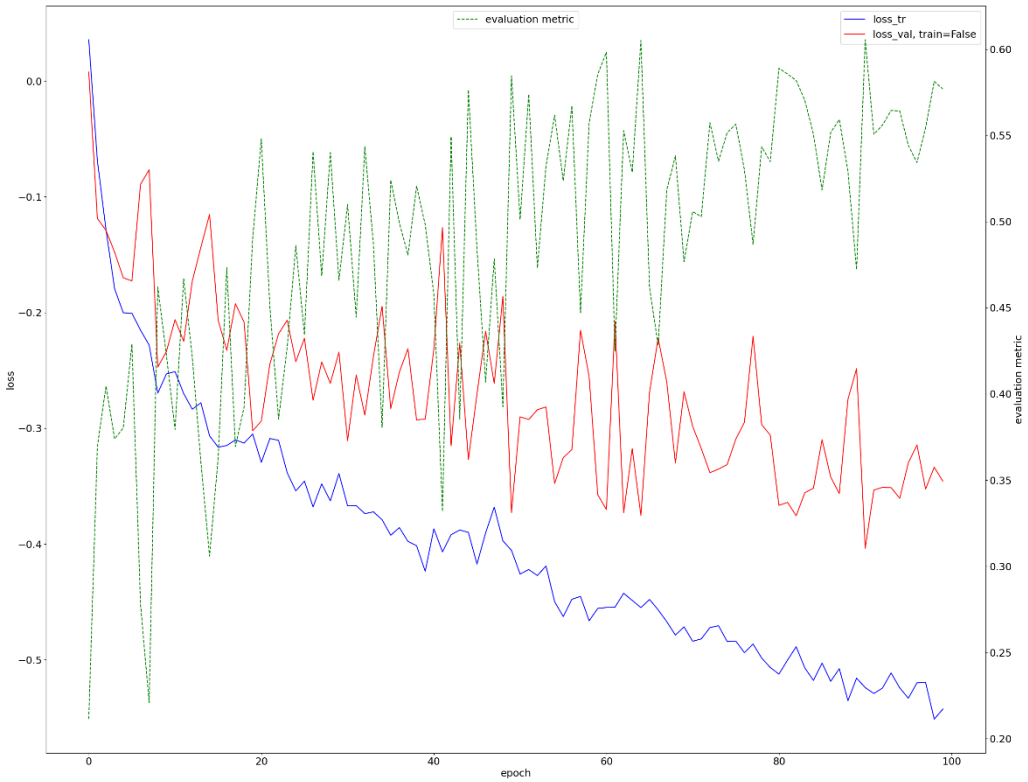
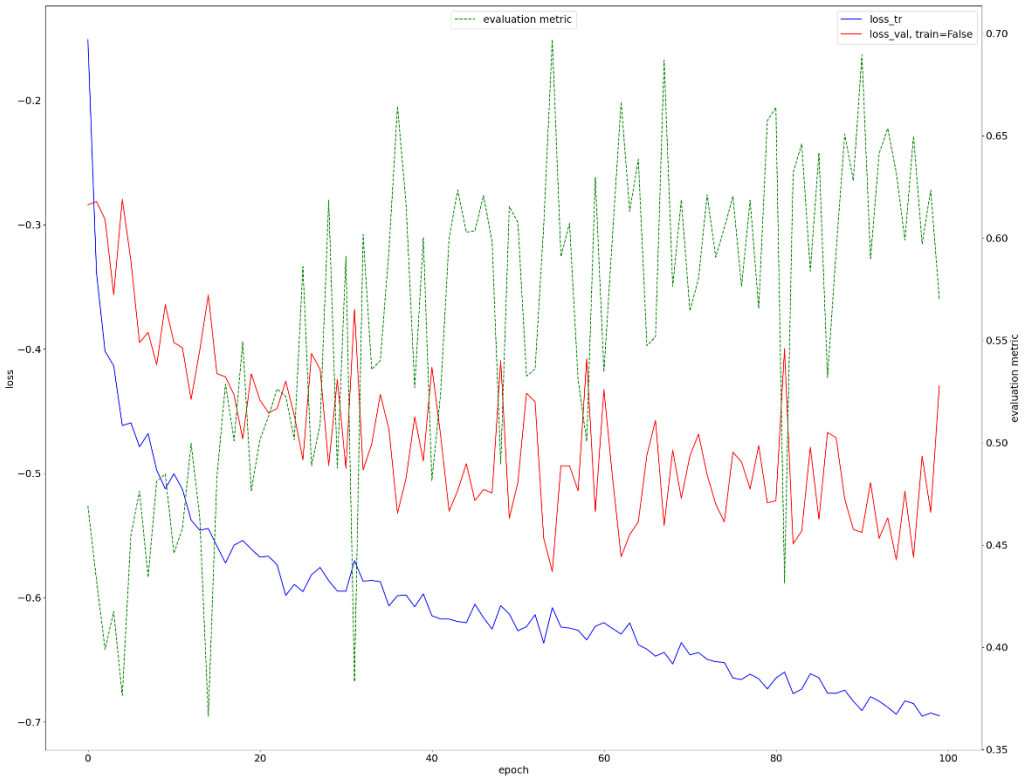


Figure 4.69: Trend of training (blue) and validation (red) losses for the model trained only with ADC images. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

(a) Training and validation losses for the model trained with only DWI images



(b) Training and validation losses for the model trained with only FLAIR images

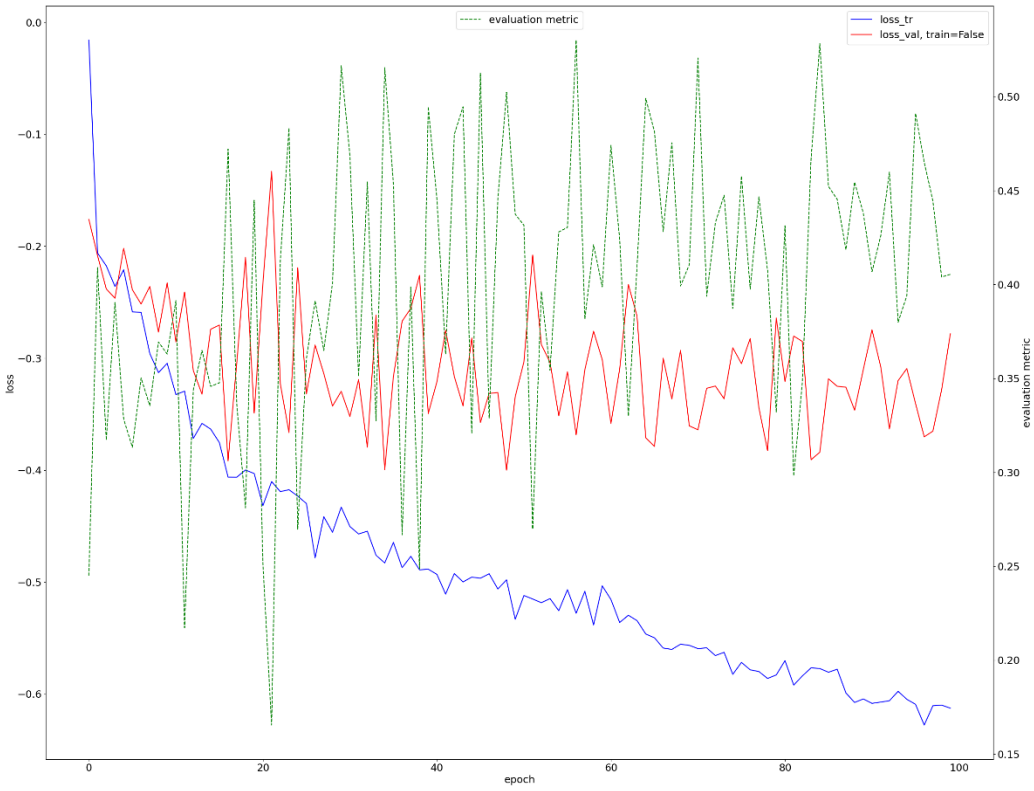


Figure 4.70: Trend of training (blue) and validation (red) losses for the model trained only with DWI images ***(a)*** *and for the network trained with only FLAIR modality* ***(b)****. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.*

4.4.2.2.1 Equalization of computational times

To keep up with the analysis performed for Brain Tumor Segmentation and allow a better comparison between the segmentation results obtained from the examination of nnUNet models trained with images of a single modality and the full model, trained with all available images, an equalization of computations is needed. As a matter of fact, likewise it happened for brain tumor segmentation, the complete model was trained with a number of images which is, in this peculiar case, three times the amount of images fed into single models, because there are three available modalities: ADC, DWI and FLAIR.

This reduces the fairness of comparison of the performances of these models, because the full model has a lot more images in input from which it can learn and extract useful features for the accurate identification of the stroke lesion. There are many possible techniques that can be used to equalize the computations between models and allow a correct comparison between them.

As a first test, it was chosen to equalize the training times required by the models, so that, even if the complete model has many more images from which it can learn, it is given the same amount of time to all networks to learn useful information from their provided inputs.

To get more in details, it was tracked the time required by training all models, in particular the network fed with only ADC images took 1 hour and 26 minutes to be trained, the model which used only DWI images had a training time of 1 hour and 31 minutes, while the configuration with only FLAIR modality took 1 hour and 32 minutes. On the other side the full model, fed with all these three modalities, employed 2 hours and 38 minutes to train. It must be specified that the training time required by all models is significantly lower than the brain tumor segmentation task, because for the stroke lesion segmentation problem, it can be appreciated a reduction of both the number of modalities (three instead of four) and the number of segmentation classes (one instead of three). This is not directly correlated on a simplification of the task, because stroke lesions have a really variable shape and location.

To equalize the computation between single models and the full model, the training of the complete model was stopped when the average time required by networks trained with a single modality was reached, and this threshold was estimated to be 1 hour and 30 minutes.

In *Figure 4.71* it's shown the trend of validation and training losses for the full model, when the training was cut off at the specified threshold, corresponding to a total number of epochs of 66. It is visible that, even if the number of epochs is almost halved, the validation loss can be considered already stable, and segmentation results are therefore expected not to vary so much from the complete model.

As before, after training all models, inference was performed using a common test set composed by the last 40 images of ISLES 2022 dataset, and their performances were compared.

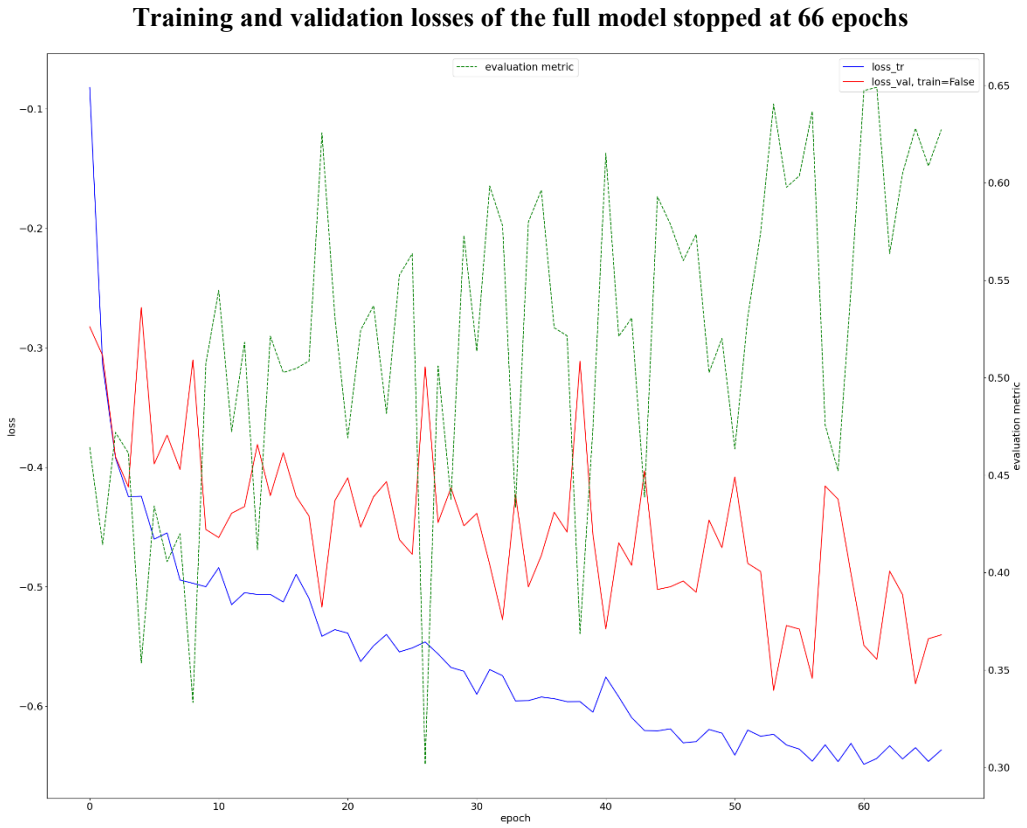


Figure 4.71: Trend of training (blue) and validation (red) losses for the network trained with all available images, but stopping its training at the time employed by models trained with a single modality. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.4.2.2.2 Equalization of number of training images

The second method chosen to equalize the computations between models was the equalization of the number of training images. As previously disclosed, the full model, fed with all available images, has a lot more content from which it can learn useful features for the identification of stroke lesions, with respect to networks trained with images of a single modality. In details, the complete model is fed with 70 training cases, each consisting in three images, one for each modality (ADC, DWI, FLAIR), for a total of 210 images used as input. On the other hand, for

single models the number of training images is just 70, because only scans of a single modality are passed inside the network. To allow a fair comparison between segmentation results obtained by all models, and force all networks to learn from an equivalent input, it was chosen to triple the number of images used to train single models. Three new datasets were then created, each consisting of 210 images of a single modality extracted from the ISLES 2022 training dataset. This was the main reason for which, at the beginning of the study, the number of training subjects was set to 70: knowing that the total amount of images available for the study was 250, which is the length of ISLES 2022 dataset, the size of the test set was set to 40 images, so that fixing the training set at 70 images could allow to perform this specific study in which the number of training images was tripled, using all available images, 210 for training and 40 for testing. In this way, even if all images come from the same dataset, training and test set remain independent and no images used to train are also used to test the models. One additional consideration that come from this analysis was that four cases from the ISLES 2022 dataset, subjects 98, 150, 151 and 170 respectively, don't have any type of stroke lesions, so they represent just healthy patients, not useful for this task.

In *Figure 4.72* it's showed the trend of training and validation losses of the model trained with 210 ADC images, which is similar to the progress of other single models. The same considerations made on the brain tumor segmentation field are here valid: the training and validation losses decrease is longer than before, meaning that the model is learning for more time, and the oscillations of the validation loss are reduced due to the increased number of images used to validate, yielding to more reproducible results. The full model was untouched for this study.

At the end of training, inference was performed using the last 40 images of ISLES 2022 dataset as test set, and performances of models were compared.

To conclude this study and verify the reproducibility of the results, the same analysis was performed reducing in a proportional way the number of images used to train each model: starting from 210 images for single models and 70 subjects for the full one (examined in this paragraph), passing to 108 images for single models together with 36 subjects for the complete one, then 54 images for single models and 18 for the full one and to conclude 27 and 9. This examination was done to check if segmentation results were consistent and how they changed with respect to the number of training images.

Training and validation losses for the model trained with 210 ADC images

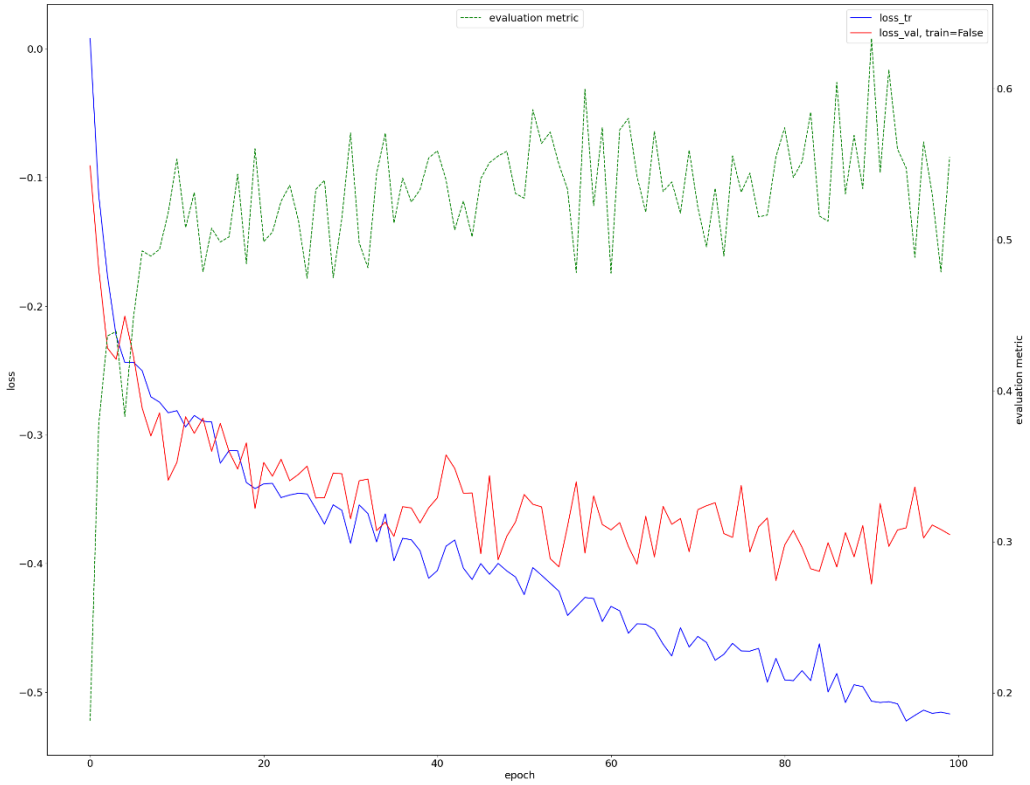


Figure 4.72: Trend of training (blue) and validation (red) losses for the network trained only with ADC images, but tripling their number, from 70 to 210. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

4.5 Ensemble of models: a new perspective

The results from the study of nnUNet models trained with images of a single modality show that, even if the full model kept having the best overall performances, the analysis of single configurations had promising results from some points of view. First of all, the training procedure of those models took considerably less time than the full model, reducing their computational cost. Moreover, focusing on some specific modalities: the network trained with T1ce images, for the brain tumor segmentation task, was able to achieve results more or less comparable, on average, with the full model, being able to overcome the latter in the segmentation of two classes out of three. While, for the ischemic stroke lesion segmentation task, the model trained with DWI modality was able to obtain comparable performances, and in some cases to overcome the complete model.

As a last consideration, the analysis of models trained with images of single modalities was performed to avoid to mix up information and features extracted from images in which cerebral and especially tumoral regions are highlighted in a different and in some cases completely opposite way, risking the model to elaborate opposite features in a joint manner.

For these reasons, to complete the research about images of single modalities, it was then chosen to carry out an ensemble learning strategy.

Ensemble learning is a branch of deep learning which involves combining a finite number of algorithms, with the conception that they could obtain a better performance than any of the single networks used on the ensemble. Literally, it consists on building models on many simpler ones, with the aim of exploiting the different learning capabilities of dissimilar networks, to improve the overall accuracy and efficiency, even if increasing the computational cost, due to the need of training more than one model.

There are two main ensemble learning strategies: bagging (bootstrap aggregation), in which constituent models are independent and trained on bootstrap samples of training data, allowing their errors to be independent as well, hence reducing variance but without affecting bias; and boosting, in which models are sequential and thus dependent, trained in sequence by reducing the errors of examples that were misclassified from previous models. The most used ensemble learning method with deep learning algorithms is bagging, because bagging technique is able to parallelize training of single models, reducing their training time (which is really high with neural networks). Moreover, neural networks often have high variance and low bias, and their differences seem to be mostly due to variance: by reducing variance without affecting bias, bagging is the ideal method.

Once the constituent models have been trained independently, there are several different ways in which they can be combined, and many of these techniques have been tried during this study, therefore they will be analyzed subsequently.

In this thesis, ensemble of models has been used to combine models trained with images of a single modality both for brain tumor and stroke lesion segmentation, to keep modalities separated and allow single models to learn specific features, combining them at inference time. The aim of this research was to identify if the road of the ensemble was passable, opening a new perspective and offering a new possible baseline in these segmentation tasks, allowing to treat images of different modalities independently but still combining their capabilities.

4.5.1 Implementation of ensemble learning

As introduced before, it was chosen to implement the ensemble of models at inference time; for this reason, training of models wasn't modified: the used networks were the ones already trained with images of a single modality.

Going into more detail, a new Python script was created for the prediction using the ensemble learning technique, and obtained by modifying the standard prediction method used by nnUNet. This baseline method requires to specify the folder containing the testing images, the output folder (automatically created it if it doesn't exist yet) where predictions will be saved, and the folder automatically generated by nnUNet containing the trained model.

The new prediction technique developed, on the other hand, requires the following inputs, which must be specified when calling the script:

- The input folder, which is the *nnUNet_raw_data* folder containing images used to test the model, without requiring to separate the different modalities as if they were to be fed into single models, but organizing images as if they were to be used by the full model, always following the structure of nnUNet;
- The output folder, which will be automatically created if it doesn't exist yet, where predictions will be saved;
- The *3d_fullres* folder, automatically created by nnUNet and containing all the 3D full-resolution UNets trained for the different tasks. For this study it was assumed that only 3D UNets configurations could be used for prediction, given that they were the only ones trained;
- *-z,* that, if mentioned, allows to save the softmax probabilities as numpy arrays in the output folder;
- *--only_ensemble*, which can be specify to avoid making predictions from single models, if they were already done or available, and simply perform the ensemble of models.

A new function was then created, which, once this method is called, automatically identifies the existing modalities inside the dataset by exploring the *dataset.json* file, which was previously necessarily created to run nnUNet.

After that, another function was designed to use the previously extracted information to select only images of a specified modality, from a dataset composed with all available images and following the nomenclature of nnUNet (e.g., BraTS_001_0000 for subject 1, modality 0000=FLAIR). This function was specifically defined to deal with images of BraTS challenge

(with modalities FLAIR=0000, T1=0001, T1ce=0002, T2=0003) or ISLES challenge (with modalities ADC=0001, DWI=0002, FLAIR=0003), because it can retrieve the specific modality from its associated index. This function is focalized on the specific purpose of this thesis, but it could be expanded in future to be able to deal with every dataset.

Once the prediction starts, if it wasn't previously chosen the *–only_ensemble* option, first of all predictions of images of single modalities using the corresponding single models are performed: images of single modalities are extracted from the whole dataset using the previously defined function, while single models are extracted from all trained nnUNet models, identifying networks which contain in their specific TaskID, isolated, the name of the modality they were trained with (e.g., Task600_FLAIR). Another necessary requirement to run this ensemble method was therefore to associate to models trained with single modalities, a task name containing only, or in any case isolated, that specific modality.

The prediction pipeline follows then the one defined by Isensee et al.

Once prediction of single models is completed, or if the *–only_ensemble* option is inserted, the function for ensemble of models is called. This function iteratively selects one subject at a time, and considers only the softmax files generated by the prediction of single models for all the present modalities. These softmax files are necessary to apply the ensemble strategy, and thus the corresponding option (*-z*) must be specified during prediction of single models. Analyzing also the properties files of each segmented image generated during inference of single models, it was possible to notice a detail: the softmax images of different modalities but corresponding to the same subject, often have slightly different shapes between them, due to the cropping of non-zero voxels applied during preprocessing. To ensemble single models softmax images are required to have the same shape, so they were chosen to be reshaped: it was tried to reshape them considering the information about the size of the cropped region inserted inside the properties of each image, and resize to the original shape by zero padding. Anyway, this technique didn't perform so well, and it was then chosen to use a different method depending on the type of dataset, applying the zero padding and switching back to the original shape only after the ensemble is completed.

After that, the proper ensemble is performed and the final segmentation map is produced and saved.

4.5.2 Output of the ensemble model

Over the years, several different ensemble techniques have been developed and tried in many biomedical domains and not only. Two main ensemble methods have been explored in this thesis: prediction averaging, which is rather intuitive and involves on simply averaging the predictions of the considered models, and majority voting, that consists on assigning to each voxel the value which has been voted by most networks.

To improve the overall performances, several different alternatives have been tried. In some cases, the simplest methods are also the best performing ones: first of all, it was in fact chosen simply to try averaging softmax probabilities produced by model trained with images of single modalities, and then to assign each pixel to the class with highest probability, as implemented in many segmentation methods developed over the years (Marmanis et al., 2016; Kumar et al., 2017; Gu et al., 2022; Moon et al., 2020).

Instead of simply averaging the predictions generated by single models, another more accurate possibility could be to perform a weighted average, assigning a different weight to each model. These weights could be trained during the training phase, to optimize a given loss function, but given that it was chosen to implement the ensemble method during inference, once models have been already trained, this technique wasn't chosen. Different values could be used as weights, and the complicated part was to select the ones leading to the optimal performances: Devan et al. suggested to perform grid search between 0 and 1 during training, to identify weights leading to better performances of any contributing model and of the ensemble model trained with equal weights (Shaga Devan et al., 2022). M. Sewell chose to train all single models, obtain the corresponding predictions and weight each voxel prediction by the model's posterior probability, and then perform the average of these weighted results (Sewell & Tat, 2011). Wan et al., like many other authors, identified a specific weight function to be trained, to find the optimal weights (Wan & Yang, 2013).

Based on literature, on intuition and on the most used metrics for semantic segmentation, it was chosen to implement a weighted average between model ensembled, where the softmax probabilities produced by each model for each class (1=ED, 2=NCR, 3=ET for brain tumor, 1=Lesion for stroke lesion), excluding the background, for all voxels, were multiplied for the average Dice Scores obtained during the validation phase for that specific class. Meaning that, after the validation phase, the average Dice scores obtained in the prediction of the three classes

were computed and saved as weights; after that, during inference, the probability of a specific pixel to belong to a given class was multiplied for the corresponding weight.

The only inconvenient was that the Dice scores, as previously cited, were available only for the foreground classes. Therefore, in the case in which the posterior probabilities of a specific foreground class and of the background were really close for a given pixel, which was wrongly classified as background, multiplying the probability of the foreground class for a weight between 0 and 1 can't be able to improve the performances, because it would reduce its probability. For this reason, it was also chosen to try to double the Dice scores obtained in the validation phase and use them as weights.

As suggested by Zhou et al. and by Quek et al., in some cases it is better to ensemble only some instead of all learners, considering only the most important ones for the segmentation purposes (Z.-H. Zhou & Tang, n.d.), (Quek et al., 2003). For this reason, it was tried to include in the prediction averaging only the base models able to achieve an average Dice score across all classes in the validation phase above a certain threshold, which was chosen to be varied between 0.5 and 0.7.

Another tried technique was to perform majority voting: each single model votes for a specific class for each pixel, and the given pixel is then assigned to the most voted class. It was also tried to modify this method to improve the performances: first of all, before performing majority voting, the posterior probabilities of each class, for each model, were multiplied for specific weights (average Dice scores obtained in the validation phase for each class, these are the weights used also in the following methods); after that the class with the highest probability was identified for each model, and majority voting was performed for each pixel.

It was also attempted to perform majority voting by doubling the vote of models with an average Dice score across all classes in the validation set higher than 0.7, or also 0.5 was tried. To be more specific, it was then tried to double the vote of models voting for a specific class and achieving an average Dice score above the previously specified threshold for that specific class, in the validation set.

The same procedure was repeated, but following the ideas of Zhou et al. and Quek et al., and performing majority voting considering only the votes of models able to achieve a Dice score on the validation set (on average or considering separately each class) over 0.7.

Finally, it was tried the following method: for each class, the model achieving the best performances on the validation set on that specific class was identified. If, for a specific pixel, the model with highest Dice score for a given class predicts that specific class, then this value is considered also for the ensemble; on the other hand, if this happens for more than one class

at the same time, it is followed the vote of the model which has a higher average Dice score on the validation set for the voted class (e.g., the model with highest performances on label ED votes for ED and has an average Dice score of 0.8 on the validation set for ED class, and the model with highest performances on ET label votes for ET and has an average Dice score of 0.7 on the validation set for ET class, then the pixel is assigned to class ED). If none of the previous cases happens, majority voting is applied.
The same procedure was then repeated but, if at the end majority voting is applied, and a specific class is chosen, but the model with the highest weight for that label doesn't vote for it, then the pixel is simply assigned to the background.
All these alternatives were tried for the identification of the optimal ensemble technique, both for brain tumor and stroke lesion segmentation tasks.

4.5.3 Ensemble of Brain Tumor Segmentation models trained with single modalities

As it was anticipated in previous chapters, before proceeding with the ensemble technique, the shapes of the softmax images of different modalities of the same subjects generated during training were compared between each other, leading to the identification of a lack of overlap between them, due to cropping to a different size during preprocessing. Analyzing more in detail the difference between the shapes of the softmax files of the available modalities (FLAIR, T1, T1ce and T2), it was observed that it was really small and not always present, removing the necessity of identifying the optimal shape: it was then chosen simply to resize all images to the first softmax file shape (it could be chosen randomly given the small differences between them), corresponding to the FLAIR image. It was also tried to resize each probability file to its original shape, considering that all images had the same shape before cropping, but this method lead to worse performances and it was therefore discarded.
Ones the sizes of all softmax images of the same subject were modified to be the same, the different ensemble techniques previously exposed were tried, considering both prediction averaging and majority voting techniques.
Only at the end of the prediction the segmentation map was brought back to its original shape, using the information included in the properties file of the first modality image (FLAIR).

4.5.4 Ensemble of Stroke Lesion Segmentation models trained with single modalities

The same analysis was repeated for the stroke lesion segmentation task, considering, however, some adjustments, due to the difference with the brain tumor segmentation task: in this case the number of modalities is three instead of four (ADC, DWI and FLAIR); but especially, the number of classes is reduced, having a single label related to the stroke lesion instead of three of the previous study. For this last reason, the ensemble techniques that required to identify the best performing models for the different classes, based on the average Dice scores obtained in the validation phase, in this case are simplified by obtaining a unique model, associated to a specific modality, which is the one obtaining the best results in the validation set with respect to the single label of the stroke lesion.

After that, in this case it was discovered that the difference between softmax files shapes of images of different modalities wasn't negligible. Resizing all images to their original shape before cropping didn't lead to optimal results, and reshaping all files to the size of the first image was considered too simplistic and carried to too many errors. It was thus chosen to resize all softmax files of each subject, to the shape of the modality image whose corresponding model obtained the highest average Dice score on the validation set.

Ones the sizes of all softmax images of the same subject were unified, the different ensemble techniques previously exposed were tried, considering both prediction averaging and majority voting techniques.

Only at the end of the prediction the segmentation map was brought back to its original shape, using the information included in the properties file of the most important modality previously identified.

4.6 Adaptation of nnUNet to IVD-Net structure: Dense Multi-path nnUNet

As introduced in Chapter 2.4, a relevant and interesting architecture for the purposes of brain tumors and stroke lesions segmentation, is the one introduced by Dolz et al. for the segmentation of ischemic stroke lesions (Dolz, Ayed, et al., 2018), whose baseline architecture was equivalent to IVD-Net, implemented by the same authors for the purpose of localization and segmentation of intervertebral disc (IVD) (Dolz, Desrosiers, et al., 2018), but whose code was publicly released.

It basically consists of a UNet network with a multi-path architecture, where the encoding path is split into a number of streams equal to the amount of acquisition modalities used for the study, allowing to feed each stream with images of a specific modality. To allow a better flow of information through the network, hyper-dense connections were also developed within and between multiple paths, and the different streams were then concatenated to form a bridge that allows to convey all the extracted information into a single decoder for the segmentation of the image.

The importance of this architecture is straightforward for this thesis: disentangling the input data based on their different modalities has the same purpose of the ensemble learning methods previously implemented, that is to separate the different modalities allowing not to fuse their information at early stages inside the network and to separately extract features from images of the same subject but with intensities that would be, in some cases, even opposite.

On the other hand, this method has the drawback of not being able to capture complex relationships between modalities, which could supply a relevant addition to better segment medical images. To make up for this lack of the model, hyper-dense connections between multiple paths and within the same ones were introduced inside the original architecture, allowing to better model the relationships between modalities. In the original paper, hyper-dense connections were implemented by feeding each level of the UNet encoders with the concatenation of the outputs of the different encoders from the previous level. To increase the performances and the regularization effect, they also chose to concatenate feature maps in a different order for each branch:

$$x_l^a = H_l([x_{l-1}^a, x_{l-1}^b, x_{l-1}^c, x_{l-1}^d])$$
$$x_l^b = H_l([x_{l-1}^b, x_{l-1}^c, x_{l-1}^d, x_{l-1}^a])$$

Where x_l^a is the output of the layer *l* from stream *a,* H_l defines the specific dense block, which consists on a convolutional layer, followed by batch or instance normalization and by a non-linear activation, while $x_{l-1}^a, x_{l-1}^b, x_{l-1}^c, x_{l-1}^d$ represent the outputs of the previous layer from the different streams (*a, b, c, d*). For each path, the feature maps are concatenated sorting them starting from the output of the corresponding stream indeed. This was performed starting from the second level, so that the images of different input modalities are still fed into separated streams; knowing that the original UNet was developed by Dolz et al. with a total of four levels. Moreover, to increase again connectivity and the flow of information inside the network, starting from the third level they also decided to concatenate to the previously analyzed tensor, for each encoder, the input of the precedent layer of the same network, cropped to the matching size.

This network seemed to match the investigations performed in this thesis, and the will of this research, but it was developed for 2D images, needing thus to be fed with 2D scans. Instead of modifying the input data to directly apply the IVD-Net architecture on brain tumor and stroke lesions data, it was chosen to follow a different road: given the countless advantages of building a 3D network, and knowing the great capabilities of the nnUNet model, it was decided to adapt the nnUNet architecture to the IVD-Net one, trying to merge the power of both models.

To do that, it was necessary to manually override the automatic setting of nnUNet and fix the number of channels to just one, knowing that each encoder must deal with images of a single modality. After that, before proceeding with the implementation of the IVD-Net architecture, the first network implemented was developed by just disentangling the input data and create N different encoders, one for each input modality, inside the nnUNet architecture. The outputs of the N streams were then concatenated and fed into a bridge, whose architecture was taken directly from IVD-Net released code, and which allows the connection between the N encoders and the decoder. After this first adaptation, different variants were tried, especially starting from the brain tumor segmentation aim and adapting to the stroke lesion segmentation one.

The produced architecture was then called *Dense Multi-path nnUNet.*

4.6.1 Brain Tumor Segmentation Task

The number of streams, which corresponds to the amount of input modalities, was obviously set to 4 (FLAIR, T1, T1ce and T2). The number of levels, each of which consists in two stacked

convolutional blocks (convolution followed by instance normalization and leaky ReLU activation function), was automatically set by nnUNet to be equal to 5.

After creating 4 different encoders, the input data were separated in the different available modalities and fed into the corresponding streams.

Following this first, simple architecture, it was chosen to properly adapt the nnUNet model to the IVD-Net one, implementing the hyper-dense connections, while skip connections between encoders and the decoder were already generated by nnUNet, with the only difference that, at each level, the proper skip connection was obtained by averaging the skip connections coming from the four paths.

To accurately implement the IVD-Net architecture, it was necessary to modify the number of input and output channels, which was again automatically set by nnUNet, to follow the dimension of the different concatenations performed, and it was also used a function developed by Dolz et al. to crop input feature maps, so that they could be concatenated with the input of the following level.

However, the development of this architecture to deal with 2D scans, has an important throwback in its 3D adaptation: the number of parameters, and thus the operations and the computational cost, increase consistently, leading to occupy the whole available GPU, and going out of memory when trying to train this adapted architecture. By removing a level, passing from 5 to 4, the computational cost remains still excessive.

To reduce the computations, it was therefore tried to decrease the number of channels at different levels, in two distinct ways: by manually reducing the number of channels in some levels of the encoders, and by introducing some convolutions with kernel size of 1x1x1 with the aim of reducing the dimensionality of feature maps by compressing their information into a lower feature size. Implementing these modifications, the problem was solved and it was also possible to add back the last 5^{th} level, going back to the original nnUNet configuration.

Other adjustments were proposed to increase the efficiency and decrease the computations of the network: first of all, following the structure of IVD-Net, the input at each level was initially obtained by concatenating all the outputs of the previous level from the different encoders, leading to obtain four (one for each modality) pretty big tensors, which contain the same information in different order. It was thus decided to neglect the idea that the ordering of the concatenated tensors is relevant, and create one single array for each level, independent from the order, being able to reduce the computational cost. After that, the produced tensor was fed into a 1x1x1 convolution, with the same aim as before: reduce the number of channels and decrease the computations. The compression factor was set at 1/4 and, after that, the input of

the following level of each encoder is obtained by concatenating that tensor with the specific output of that path, instead of the input: in this way, the input tensor for each level of each encoder contains a 50% of path-specific information, and a 50% of scrambled and compressed cross-path information, in order to maintain the dense connections between different streams but also to keep streams separated and to allow them to extract different features. A schematic representation of the stream of information from consequent levels in the encoder can be appreciated in *Figure 4.73* (courtesy of L. F. Tshimanga).

The last proposed improvements were related to the skip connections: by integrating the structure of IVD-Net and nnUNet, encoders' outputs were saved inside specific containers and, at the corresponding level in the decoder, the skip connections of different paths were averaged; the averaged skip connection was then concatenated with the corresponding tensor fed into the localization pathway. A schematic representation of those considered skip connections can be visualized in *Figure 4.73*.

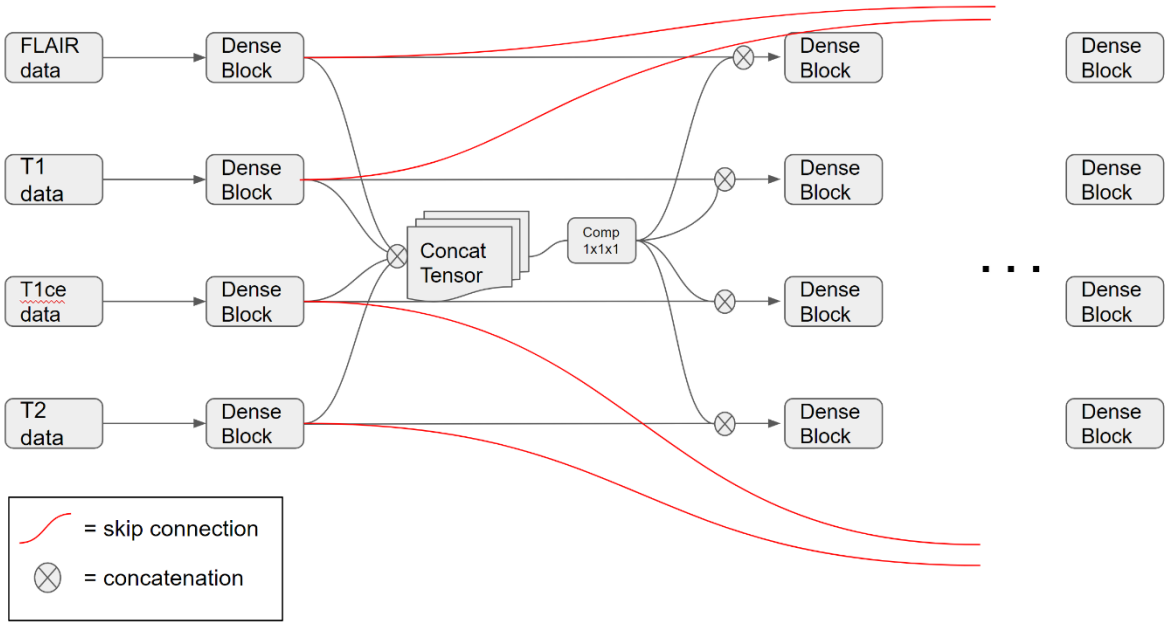


Figure 4.73: Schematic representation on how information is passed through different levels of the encoders, maintaining the skip connections (red lines, as can be seen in the legend) how they were conceived in the original IVD-Net. Outputs of the same levels of different streams are concatenated, and the produced tensor, in this first trial, was chosen to be compressed with a unique 1x1x1 convolution; for each path, the result is then concatenated with the output of that stream, obtaining the input of the following level. The outputs of each level (Dense Blocks) are used as skip connections. L. F. Tshimanga is warmly thanked for the realization of these graphs.

Giving that the information contained inside the skip connections correspond to the one inside the concatenated tensors created at each level, to save memory and computations it was tried to use the concatenated tensors as skip connections as well, after their compression with 1x1x1

convolutions (*Figure 4.74*). In this way, at each level of the decoder, no average between skip connections coming from the different streams must be performed, because the compressed tensor is ready to be used as skip connection and concatenated with the corresponding tensor passed inside the decoder. Another method that we tested consisted in using four different 1x1x1 convolutions of the concatenated tensor at each level, one for each stream, so that, during training, the learnt weights for each stream could be specific for that path. This compressed tensors, could then be used as specific skip connections for each path, and averaged at each level of the decoder, as suggested in the original architecture (*Figure 4.75*). This method will be the one resulting in the higher performances, and so chosen as final architecture.

A network for each technique was trained, repeating the analysis if the model was able to obtain good performances, to verify the reproducibility of the results.

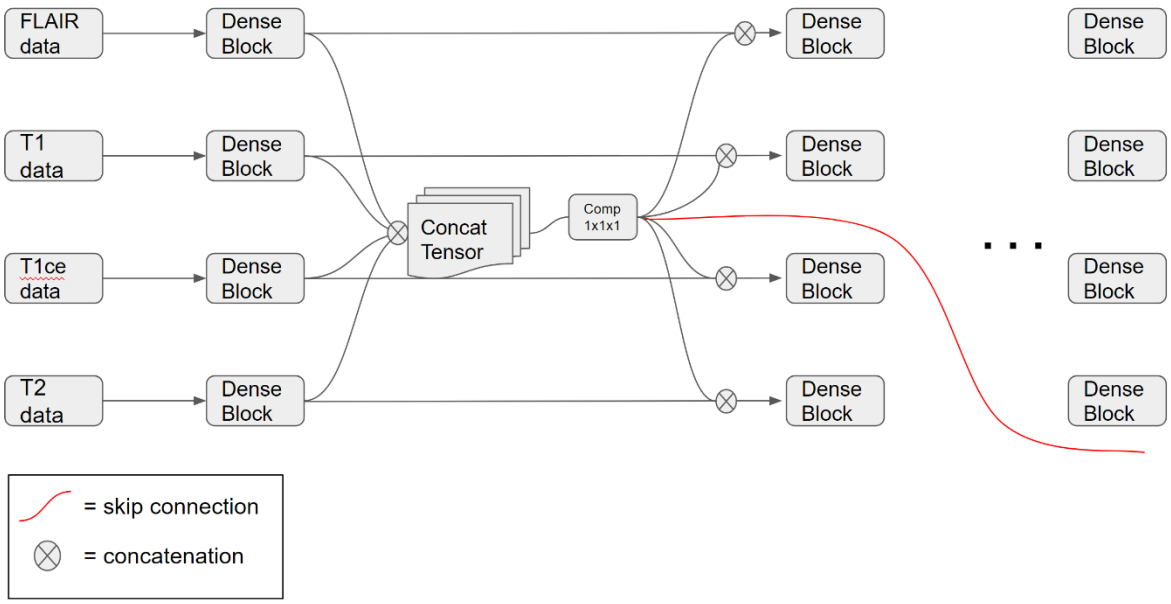


Figure 4.74: Schematic representation on how information is passed through different levels of the encoders, modifying the skip connections with respect to the original architecture. The structure and flow of information inside the encoders is unchanged, while, for each level, the tensor obtained by concatenating the outputs of the previous level, is compressed with a factor of ¼, and this compressed tensor is used as skip connection, and concatenated with the corresponding tensor at the decoder level. L. F. Tshimanga is warmly thanked for the realization of these graphs.

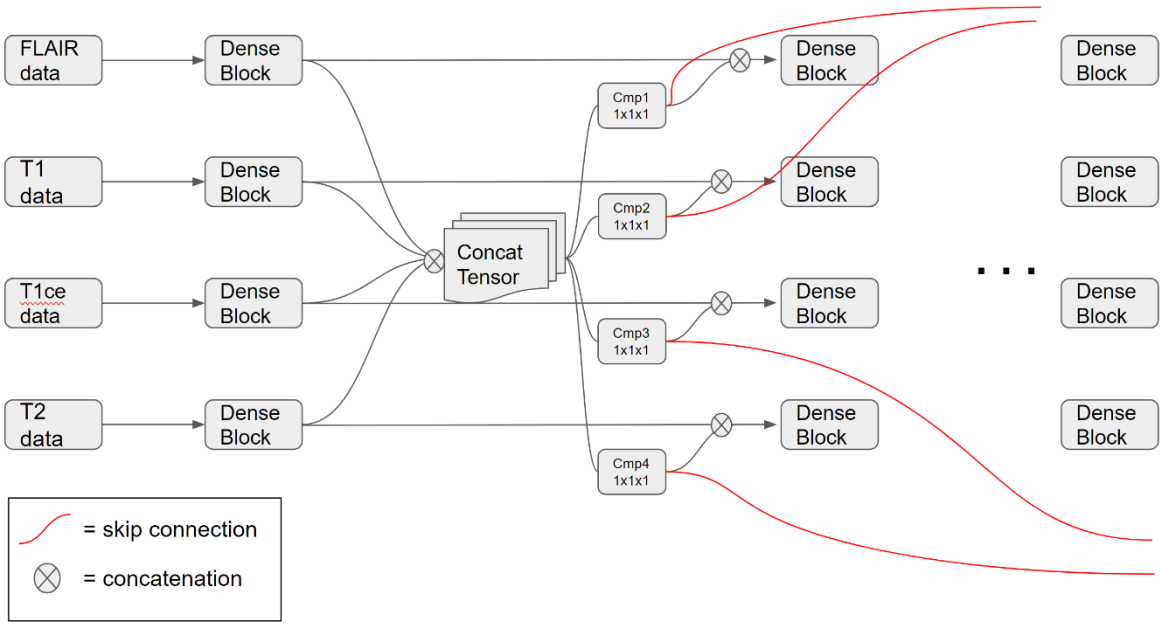


Figure 4.75: Schematic representation on how information is passed through different levels of the encoders, modifying the skip connections. The flow of information inside the encoders is slightly modified because, for each level, the tensor obtained by concatenating the outputs of the previous level, is compressed with a 1x1x1 convolution specific for each path. The resulting compressed tensors for the different paths are thus used as skip connections and averaged at the corresponding level of the decoder. L. F. Tshimanga is warmly thanked for the realization of these graphs.

4.6.2 Stroke Lesion Segmentation Task

The number of encoders created for the Dense Multi-path nnUNet for stroke lesion segmentation was three, having three available modalities (ADC, DWI and FLAIR). In this case, the number of levels, each consisting as before of two stacked convolutional layers, automatically set by nnUNet was four, one less than brain tumor segmentation. As before, an initial architecture was created by modifying nnUNet to have three encoders, one for each input modality, whose outputs are connected through a bridge to a single decoder.

After this, it was chosen to implement the IVD-Net structure into nnUNet, developing the dense connections between and within paths, always remembering that the skip connections were already implemented by nnUNet; they were only modified so that skip connections coming from the three input encoders are averaged at the corresponding level of the decoder. Moreover, the difference in the initial architecture with respect to brain tumor segmentation allowed to avoid to encounter the same problems met before, so that the network obtained by inserting the

IVD-Net structure into nnUNet could be run without meeting memory or computations problems.
The number of channels was then left unchanged. On the other hand, the best performing techniques for brain tumor segmentation, obtained by increasing the efficiency and decreasing the computations of the architecture with changes in the concatenations and skip connections, and corresponding to the structures visualized in *Figures 4.74* and *4.75* for brain tumor segmentation, were also adapted and tried for stroke lesion segmentation, to identify the network able to obtain the best results in this field too.

4.7 Inter-pathology Learning

The main purpose of this thesis is to divide the input data based on the different acquisition modalities, and keep them separated while they are fed inside the network, so that the information extracted from them is not fused at early stages. This therefore allows not to merge features extracted from images which contain the same information which is, in some cases, highlighted in opposite ways. This procedure has been applied both for brain tumor and stroke lesion segmentation, but without finding a way to combine these two tasks.
Analyzing the provided and studied datasets for the two tasks, ISLES 2022 is the main dataset used for stroke lesion segmentation, as well as the most recently released, and is composed by images of ADC, DWI and FLAIR modalities, while FeTS 2022 is the most used dataset for brain tumor segmentation, as well as the most recently released in this case too, and comprehends FLAIR, T1, T1ce and T2 images.
Given that images of FLAIR modality are useful and used for both tasks, it was chosen to perform a technique called "*Inter-pathology Learning*" between the model trained with only FLAIR images for brain tumor segmentation, and the corresponding network trained as well with only FLAIR images for stroke lesion segmentation. "*Inter-pathology Learning*" is a term that refers to the application of the Transfer Learning methods between tasks involved in different pathologies.
Transfer Learning can be defined as a machine learning technique based on exploiting knowledge gained on solving one problem, to solve a new one. When dealing with deep learning models, its definition can be adapted as tuning a network pre-trained and pre-designed for a given task, to perform on a similar one. The advantages of transfer learning are that, if the

two tasks are similar between each other, it allows to take advantage of what the model has already learnt from the first task, reducing the time of building and validating a second network, which is usually really time consuming. On the other hand, if the two tasks are not identical, there might be information that the original model has learnt which is not useful for the second one, and new information that the new model needs to learn from the data. There are two ways in which transfer learning can be applied: if the two tasks are really similar between each other and/or the data available are limited, after the training of the first model, the training of the second network is performed by freezing the weights of all but the last layers (or only last layer, depending on the task), so that only the last layers, which features are linked to task specific information, are updated. While if the tasks are dissimilar and/or the amount of available data is huge, all weights can be updated, but starting from the final weights of the previous task, instead of a random initialization. Transfer Learning works well only if the low and intermediate level features learnt from the first task are general, and can be meaningful for the second one, and so if the first task is linked and can represent a generalization of the second one.

In this thesis the two analyzed tasks were brain tumor and stroke lesion segmentation. No assumptions about the similarity between the tasks could be made, given that brain tumors and ischemic stroke lesions are two different pathologies, with disparate causes, manifestations and characteristics. Anyway, given that images of the same acquisition modality (FLAIR) are used for both brain tumor and stroke lesion segmentation, it was tried to apply Transfer Learning between these tasks, assuming that brain tumor segmentation can be considered a more general task, given the greater number of classes and the general better performances in this field.

It was therefore performed the Transfer Learning technique between two different pathologies, from brain tumor to stroke lesion segmentation, to identify if the architecture and the weights learnt for the former task could be useful also for the latter, and thus provide some additional information which could in some way link these pathologies; for these reasons this method was called ‘Inter-pathology learning’.

Given that nnUNet automatically sets the preprocessing methods for each task and during this procedure derives the dataset fingerprint used to create the model architecture, a preliminary step was to preprocess the stroke lesions FLAIR data with the same plans used to preprocess the brain tumors FLAIR data, so that nnUNet could automatically generate the same architecture as the first task for the second one. After the complete training of the first model, trained for brain tumor segmentation using only FLAIR data, the correspondent weights were saved and used for the stroke lesion segmentation model, in turn trained with only FLAIR

images. The first analysis was performed by using the pretrained weights coming from the brain tumor segmentation task, only as an initialization of the weights for stroke lesion segmentation, given that the two tasks are not so similar and the amount of data is quite high. After that, it didn't make sense to freeze all layers with the exception of the last one, because dealing with UNets, this means that all the features learnt in the encoder and decoder pathways are related to brain tumors, and using them to segment stroke lesions can't lead to good performances, given that also the relationship between image intensities is different. A sensible analysis, which was chosen to be performed, was to freeze only the weights of the encoder from the brain tumor segmentation model, and learn during training the weights of the decoder, initializing them to the corresponding weights learnt for brain tumor segmentation: in this way, it is possible to identify if the features learnt from brain tumors can be relevant also for the localization of stroke lesions.
The brain tumor segmentation model was trained using 90 FLAIR images extracted from BraTS 2020, while the stroke lesion segmentation model with 70 FLAIR images from ISLES 2022.
This study allows to understand if a model already pretrained for the brain tumor segmentation task with FLAIR images, could be useful also for the identification of the optimal model for stroke lesion segmentation using images acquired with the same acquisition modality.

4.8 Final models

The final training of the best performing models was then carried out, based on some considerations: other than the original nnUNet models, for both stroke lesion and brain tumor segmentation, the chosen architectures were the previously tested ones able to reach or even overcome the performances of the correspondent nnUNet model in the previous analyzes, showing promising results. Therefore, the chosen final networks were:

- Original nnUNet for Brain Tumor Segmentation;
- Original nnUNet for Stroke Lesion Segmentation;
- Dense Multi-path nnUNet for Brain Tumor Segmentation;
- Dense Multi-path nnUNet for Stroke Lesion Segmentation;
- Original nnUNet trained with FLAIR images for Brain Tumor Segmentation (baseline model to be used for Inter-pathology Learning);

- nnUnet trained using Inter-pathology Learning technique with FLAIR images for Stroke Lesion Segmentation.

With respect to the previous tests realized, these models were trained and evaluated using larger datasets, and increasing also the number of epochs, leading to a net increase of the training time but also of the overall performances.

For the stroke lesion segmentation task, it was chosen to use the most recently released dataset, which was also used for the previous analyzes: the ISLES 2022 dataset. Having only 250 available subjects, it was decided to randomly extract 210 subjects for the training set, and 40 for the test set. Remembering that images of three modalities (ADC, DWI and FLAIR) were acquired from each subject, the training set is composed by a total of 630 images.

The FeTS 2022 dataset was used for brain tumor segmentation, which is the most recently released in this aim and in which the amount of subjects is much bigger than the ISLES 2022 dataset: 1254 subjects, randomly divided between training and test set with an 80/20% split (1003 subjects in the training set, 251 in the test set). Knowing that each subject consists in images of four modalities (FLAIR, T1, T1ce and T2), the total dimension of the training set is 4012 images.

It can be therefore derived that the Inter-pathology Learning technique is applied between a model trained for brain tumor segmentation using 1003 FLAIR images, and a model trained for stroke lesion segmentation with 210 FLAIR images.

The number of training epochs was initially set to 1000, which was the default value chosen by Isensee et al. for nnUNet. After the completion of the training of the first model, the original nnUNet for stroke lesion segmentation, it was noticed that the validation loss was pretty much static, without improving after more or less 600 epochs, as can be seen in *Figure 4.76*. Moreover, the Dice score obtained didn't change so much with respect to the one obtained using 100 epochs; it was therefore chosen to use 600 epochs to train the other final models.

The average Dice scores obtained in the test set by models trained for the same tasks were then compared between each other.

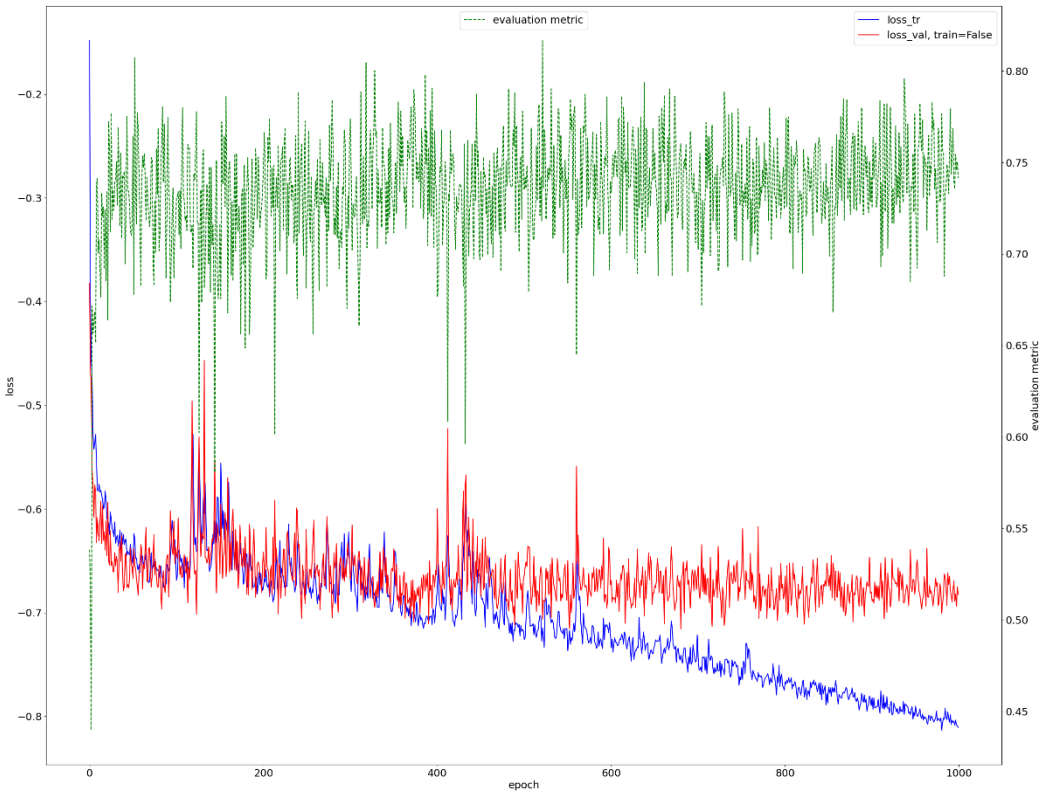


Figure 4.76: Trend of training (blue) and validation (red) losses for the original nnUNet model trained for stroke lesion segmentation, using 1000 epochs. As previously stated, the green curve (global Dice score) is displayed by way of example, and is computed on patches randomly drawn from the validation set at the end of each epoch, treating the patches as if they all originate from the same image. It is computed because it's easy to calculate during training and is still able to identify if the model is training or not.

5. Results

All the results of the previously exposed methods are presented in this section, comparing between each other the different studied methodologies on common test sets uniquely determined at the beginning of the analyses both for brain tumor and stroke lesion segmentation. To be more specific, the test set used during the investigations for brain tumor segmentation was composed by 80 images extracted from the FeTS 2022 dataset, while the one used for stroke lesion segmentation was composed by 40 images sampled from the ISLES 2022 dataset. The evaluation metric used to evaluate the results obtained and compare the different segmentation methods was the Dice score, or Dice similarity coefficient (DSC), which measures the similarity between two segmentation maps based on their overlap (Li et al., 2019). It ranges between 0 and 1 and a higher value means a better match between the segmented image and the ground truth mask, underlining a better segmentation performance. For each class, being $P_i \in (0,1)$ the binary image produced by segmentation models for class i, while $T_i \in (0,1)$ the corresponding ground truth map, the Dice score is defined as:

$$Dice(P,T) = \frac{2 \times |P_i \cap T_i|}{(|P_i| + |T_i|)}$$

Where $|P_i \cap T_i|$ calculates the amount of elements which are common in both sets, meaning the pixels which are equal to 1 in both P_i and T_i. $|P_i|$ and $|T_i|$ are instead the cardinalities of the two sets, defined as the number of pixels where P_i=1 and T_i=1.
The results are then presented in the order exposed in Chapter 4.

5.1 Brain Tumor Segmentation analysis

5.1.1 Training of nnUNet with all but one modality

The Dice Scores for the three classes, and their average, obtained by all the evaluated models on the test set are visible in *Table 5.1*.

It is possible to notice that the removal of T1 and T2 images doesn't affect the segmentation results, instead the Dice scores for classes 2 and 3, and the average between classes, increase with respect to the full model. On the other hand, the segmentation outcome drastically decreases when removing T1ce modality, especially for classes 2 and 3, suggesting the large importance of T1ce images on the delineation of the tumor core, which comprehends the enhancing tumor (label 3) and the necrotic tumor core (which is reminded that is associated to label 2 and not 1, because of the switch performed by nnUNet for labels adjustment).
The removal of FLAIR modality decreases slightly the performances exclusively for the first class, while only the deletion of FLAIR or T1ce images causes a strong reduction of the Dice score for the first label, which defines the general contour of the tumor edematous region.
This study suggests that the most contributing and effective modalities in the segmentation of the different brain tumor subregions are T1ce and FLAIR, while the intake provided by T1 and T2 modalities seems to be significantly lower compared to the first two, reducing their impact for the delineation of the brain tumor.

	Model trained without FLAIR images	Model trained without T1 images	Model trained without T1ce images	Model trained without T2 images		Full model
ED	0.669	0.733	0.667	0.733		0.735
NCR	0.547	0.549	0.285	0.545		0.523
ET	0.725	0.779	0.360	0.758		0.716
Mean	0.647	0.687	0.437	0.679		0.658

Table 5.1: Dice scores obtained for the three classes (ED stands for peritumoral Edematous tissue, NCR for Necrotic tumor Core, ET for Enhancing Tumor), and their average, by the four models, each trained with all modalities except one, compared with the full model (last column) trained using all modalities. The greatest Dice scores with respect to the full model, underlining a small need of the corresponding modality, are highlighted in green, while the smallest Dice scores, underlining a great need of the corresponding modality, are highlighted in red.

5.1.2 Training of nnUNet with a single modality

Table 5.2 shows the Dice scores of the three classes, and their average, obtained by the single models, compared with the full model. In most cases, the behavior of single models reflects the

expectations: networks fed with images of a single modality are not able to accomplish comparable results with respect to the full model, on one side because the number of images used to train single networks is much lower than the ones used for the full model, on the other because the single networks can't take advantage of the fusion of features from different modalities, that can bring complementary information.

But, on the other hand, it has been previously demonstrated that not all modalities are really useful for obtaining good segmentation results, and, furthermore, some modalities have a huge impact on the results with respect to the others. This assumption was strengthened with this subsequent analysis: it is evident that, even if all modalities perform quite poorly with the segmentation of the first class, which represents the general burden of the tumor, the model fed with T1ce modality obtains excellent results on the segmentation of second and third classes, overcoming the full model. This means that T1ce modality is more useful than all the others in the identification of the tumor core, and its performances even get worse due to the influence of other modalities: it seems that, in this specific case, it's better to use the single T1ce modality for the identification of the tumor core, because it performs better than fusing all modalities.

	Model trained only with FLAIR images	Model trained only with T1 images	Model trained only with T1ce images	Model trained only with T2 images		Full model
ED	0.664	0.515	0.577	0.613		**0.735**
NCR	0.283	0.300	**0.560**	0.374		0.523
ET	0.352	0.342	**0.762**	0.387		0.716
Mean	0.433	0.386	0.633	0.458		**0.658**

Table 5.2: Comparison of the Dice scores of the single (trained using a single modality) models with respect to the full one (trained with all modalities) in the segmentation of the three tumoral regions (ED stands for peritumoral Edematous tissue, NCR for Necrotic tumor Core, ET for Enhancing Tumor). Segmentation results obtained by single models are much worse than the full model, except for the model fed with T1ce modality, which is able to overcome the performances of the full model for classes 2 and 3; the best values obtained for each class are highlighted in bold

5.1.2.1 Equalization of computational times

The Dice scores obtained by models trained with images of a single modality, and the full model, equalizing its training time with respect to the single models, are showed in *Table 5.3*. It's evident that this result is not so much informative, because it doesn't provide additional information compared to the previous analyzes. As reported inside the table, indeed, the segmentation results obtained by the full model by equalizing the computational time, and so cutting off its training at 50 epochs, are even better than the results obtained by the full model, trained completing the 100 epochs.

These results can be explained with the oscillatory trend of training and validation losses, and it also represents an additive proof that the model stabilizes before the chosen threshold of 100 epochs. The remarks that can be inferred from these results are the same carried out for the analysis of nnUNet models trained with images of a single modality for brain tumor segmentation task (previous paragraph). Anyway, it must be highlighted that the comparison between results obtained by models trained with images of a single modality with the ones obtained by the full model is more meaningful and more correct in this last case, because the computations have been equalized and all the models had the same amount of time available to learn from the given input images.

	Model trained with only FLAIR images	Model trained with only T1 images	Model trained with only T1ce images	Model trained with only T2 images		Full model equalizing training time with single models	Full Model trained with 100 epochs
ED	0.664	0.515	0.577	0.613		**0.747**	0.735
NCR	0.283	0.300	**0.560**	0.374		0.555	0.523
ET	0.352	0.342	**0.762**	0.387		0.775	0.716
Mean	0.433	0.386	0.633	0.458		**0.692**	0.658

Table 5.3: Comparison of the Dice scores of models trained with images of a single modality with respect to the full one (trained with all modalities), and with the full model but equalizing the training times, in the segmentation of the three tumoral regions (ED stands for peritumoral Edematous tissue, NCR for Necrotic tumor Core, ET for Enhancing Tumor). Segmentation results obtained by single models are worse than the full model, except for the model fed with T1ce modality, which is able to overcome the performances of the full model for classes 2 and 3; the best values obtained for each class are highlighted in bold

5.1.2.2 Equalization of number of training images

The other chosen method to equalize the computations between full and single models was by equalizing the number of images used to train each architecture, and therefore quadrupling the amount of training images for single models. The Dice scores obtained by models trained following this pipeline, and by the full model, are indicated in *Table 5.4*. By increasing the number of images from which single models can learn, the performances improve a lot with respect to the equalization of the computational time. These results can be explained on one hand because, as previously stated, the full model stabilizes before the threshold of 100 epochs, so limiting its training time didn't affect the results (which indeed increased); on the other hand, even limiting the training time, the full model has the possibility to learn useful features from four times the number of images available from single models, increasing a lot the variability that the full model can learn, and thus the possibility to generate more detailed segmentations.
In consequence, the results obtained using this equalization of computations method are more informative, because it was provided to each model the possibility to learn from an input of the same size. Observing *Table 5.4*, the difference between Dice scores obtained by single models and by the full model are much less marked than before. The gap between performances of models trained with only FLAIR, T1 or T2 images still exists, but it's less evident.
On the other side, while the network fed with T1ce scans was able to obtain comparable results with the full model, even when the number of training images was four times lower, in this case its segmentation results largely exceed the outcome of the full model, having a comparable Dice score for the segmentation of the peritumoral edematous/invaded tissue (ED), but overcoming the performances of the full model for the segmentation of the tumor core (NCR+ET) and on average.
It represents a strong result, suggesting that, if it was available a large amount of T1ce images, an nnUNet network trained with them could perform better than using images of four modalities (FLAIR, T1, T1ce and T2) in the segmentation of brain tumor subregions.

	Model trained quadrupling FLAIR images	Model trained quadrupling T1 images	Model trained quadrupling T1ce images	Model trained quadrupling T2 images		Full Model
ED	0.787	0.680	0.726	**0.749**		0.735
NCR	0.541	0.573	**0.791**	0.611		0.523
ET	0.561	0.586	**0.820**	0.601		0.716
Mean	0.630	0.613	**0.779**	0.654		0.658

Table 5.4: Comparison of the Dice scores of models trained with 360 images of a single modality with respect to the full one, in the segmentation of the three tumoral regions (ED stands for peritumoral Edematous tissue, NCR for Necrotic tumor Core, ET for Enhancing Tumor). Segmentation results obtained by single models are more or less comparable with the full model, except for the model fed with T1ce modality, which is able to largely overcome the performances of the full model for classes 2 and 3 and on average; the best values obtained for each class are highlighted in bold

5.1.2.3 Analysis of the consistency of results

The analysis of models trained with images of single modalities was repeated with a proportionally decreasing trend to analyze the consistency of the results and identify if the relationship between single models was stable when varying the number of training images. For each class (ED, NCR and ET) and on average, on *Figures 5.1* and *5.2* are showed the Dice scores of the single models trained respectively with 20, 40, 88, 176 and 360 images of a unique modality, compared with the full model trained with 5, 10, 22, 44 and 90 subjects, each consisting of four images of the different available modalities, and thus with the same amount of images of single models.

From *Figure 5.1* it's possible to notice that the performances of single models for the ED class have a continuous and almost linear increase, proportional with the number of training images, while the full model tends to be more stable, and is indeed overcome by the FLAIR model already with 176 training images, but also by the T2 model with 360. For the second class (NCR), it's visible the superiority of the T1ce model with respect to all the others, including the full one, even for a small number of training images, but the peculiar aspect is that, when using 360 images for training, all single models are able to beat the performances of the full one. On the other hand, the T1ce model is even more separated from the other networks for the last class (ET), always overcoming the full model. But, unlike the previous case, the full model

is here superior to all the other single networks for the whole range of number of training images used.

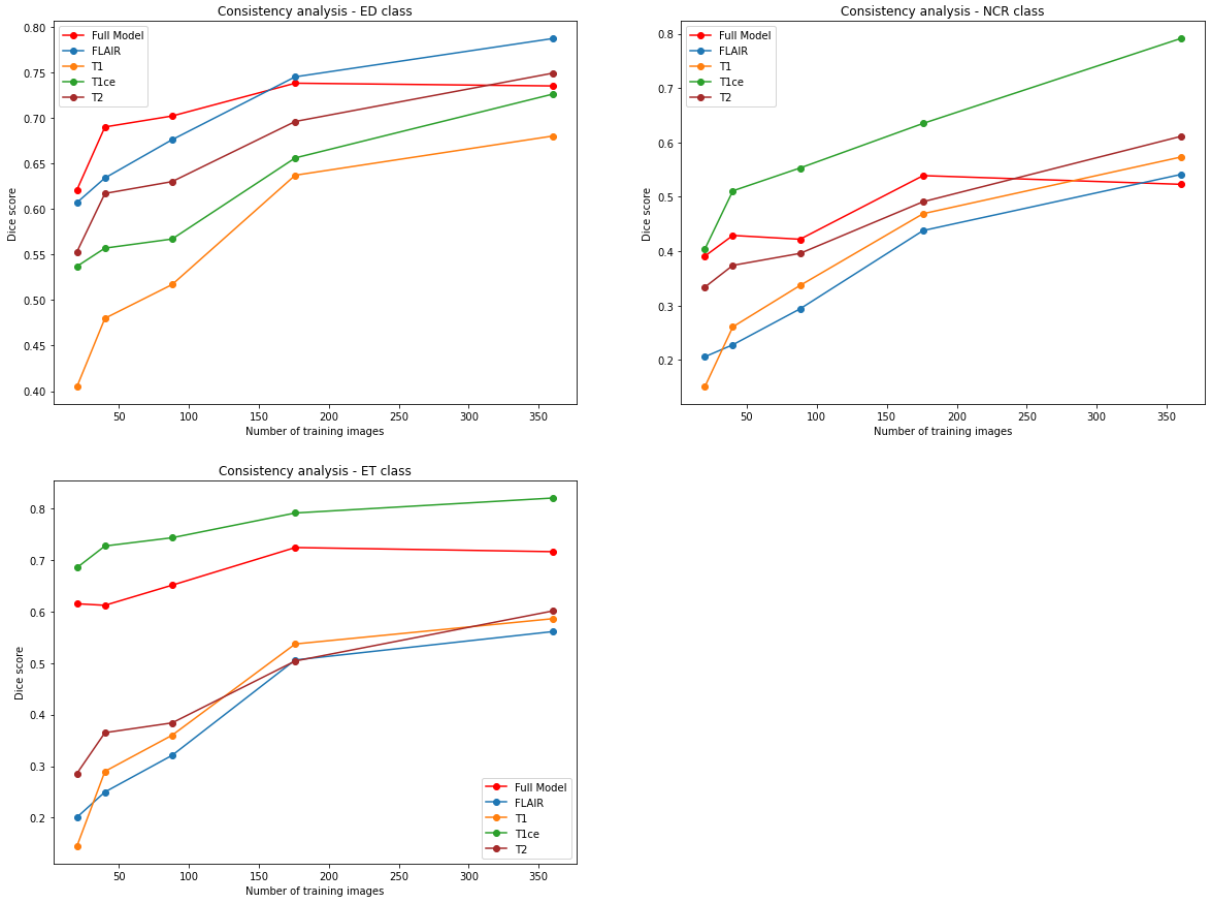


Figure 5.1: Consistency analysis for the different brain tumor classes. Five models were trained (four with images of a single modality, which can be FLAIR, T1, T1ce or T2, and a full model using all those) with an increasing number of training images, specifically: 20, 40, 88, 176 and 360. After that, their performances were compared and the trend of the Dice scores for the different number of training images is showed for the three tumoral subregions: peritumoral Edematous/invaded (ED), Necrotic tumor core (NCR), Enhancing tumor (ET)

Figure 5.2 shows the average Dice scores over the three classes. The T1ce model keeps being better than the full one for all the range of training images, being exactly equal to it for the lower number of images tried (20). Moreover, if the performances of the single models keep increasing with the number of training images used, it seems like the full model tends to stabilize: the difference between the full model and FLAIR, T1 and T2 models is huge with a small number of training images, but keeps decreasing until it can be considered more or less at the same level with 360 images.

These results can suggest that models trained with images of a single modality have higher capability to continue to increase while increasing the number of training images, with respect to the model trained with images of all modalities.

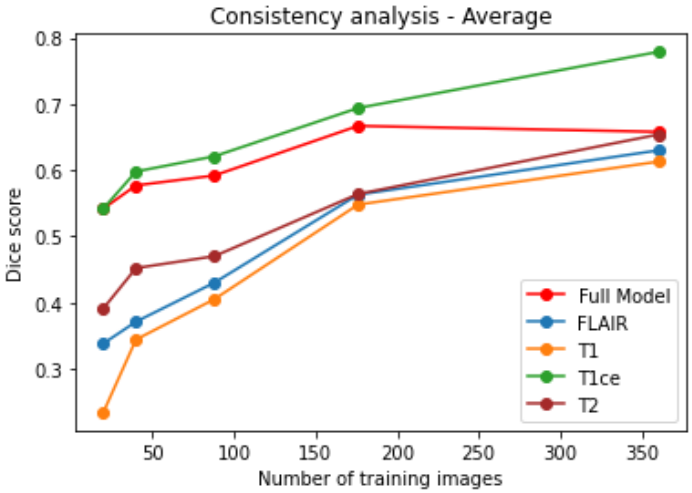


Figure 5.2: Consistency analysis on average. Five different models were trained (four with images of a single modality, which can be FLAIR, T1, T1ce or T2, and a full model using all those) with an increasing number of training images, specifically: 20, 40, 88, 176 and 360. After that, their performances were compared and the trend of the average Dice scores for the different number of training images is showed.

5.1.3 Ensemble of nnUNet models trained with a single modality

As analyzed in Chapter 4, many different ensemble techniques were tried for the brain tumor segmentation task. The results of the most representative ones, together with the Dice scores obtained by single models and by the full model are visible in *Table 5.5*. The first ensemble methods tried weren't able to obtain optimal results: simply averaging, for each class, the softmax probabilities produced by models trained with images of a single modality, and then assigning to each pixel the class with the highest probability, lead to an average Dice score of merely 0.453, against the 0.658 obtained by the full model.

On the other hand, by continuing to use the averaging techniques, but employing as weights the average Dice scores obtained for each class during the validation phase, multiplied by two; that is, by multiplying the softmax probability produced by a single model for each class, for the average Dice score obtained for that specific class in the validation set, the Dice score increased to 0.500, which was still really low.

It was discovered that, in this field, the majority voting techniques were able to produce much better results. All the previously specified techniques were therefore tried, leading to increasing results, until the final ensemble model, the one obtaining the best results among all the tried

methods, was chosen: it consists in the procedure based on saving the best performing model on the validation set for each class; then, during prediction, if the model obtaining the best performances for a given class, predicts that class for a certain pixel, then that pixel is assigned to the voted class; while if this happens for more than one class (meaning that if more than one network obtaining the best Dice scores on the validation phase for a given class predicts the corresponding class), it is followed the prediction of the model having the highest weight. If none of this happens, majority voting is performed; but if the voted class for a given pixel isn't voted by the best performing model for that class, then the pixel is assigned to the background. With this complex method, the model was able to reach an average Dice score on the test set of 0.647, which was really close to the full model but wasn't able to reach it.

Positively, the ensemble model was able to beat, on average, the single model trained with only T1ce images (and therefore all the other single models), but on the other hand the T1ce network was still able to obtain better Dice scores for NCR and ET classes: these results suggest that, for those two classes, in the ensemble strategy, the T1ce model was corrupted by the presence of the other models, without allowing the ensemble model to reach optimal results. On the other hand, this is still a good result, meaning that there is still room for improvements of the ensemble technique.

Moreover, as it can be seen in *Table 5.5*, the final ensemble model is able to beat the original architecture for the segmentation of the second class, corresponding to the Necrotic Tumor Core (NCR).

	Model trained with only FLAIR images	Model trained with only T1 images	Model trained with only T1ce images	Model trained with only T2 images	Ensemble (predictions averaging)	Ensemble (predictions averaging with 2*DS weighting)	Final Ensemble		Full model
ED	0.664	0.515	0.577	0.613	0.510	0.596	<u>0.711</u>		**0.735**
NCR	0.283	0.300	**0.560**	0.374	0.427	0.445	<u>0.541</u>		0.523
ET	0.352	0.342	**0.762**	0.387	0.421	0.455	<u>0.690</u>		0.716
Mean	0.433	0.386	0.633	0.458	0.453	0.500	<u>0.647</u>		**0.658**

Table 5.5: Comparison of the Dice scores obtained by models trained with images of a single modality, and to the full one, compared with the results obtained by using different ensemble techniques, in the segmentation of the three tumoral regions (ED stands for peritumoral Edematous tissue, NCR for Necrotic tumor Core, ET for Enhancing Tumor). For each row, the Dice score in bold is the higher value obtained for that specific class (or on average) while the results of the final ensemble model, the one with the higher performances, are underlined.

5.2 Stroke Lesion Segmentation analysis

5.2.1 Training of nnUNet with all but one modality

Segmentation results, expressed as Dice scores obtained by the networks in the segmentation of the ischemic stroke lesions, are showed in *Table 5.6*, in which the performances of models trained with all but one modality, are compared with the full model, trained with all available images. To compare the computational costs of these models, also the training time is represented inside the table.

The results mirror the expectations on this analysis: by alternatively removing one modality and using the remaining two to train a model, the segmentation performances are inferior with respect to the full model, trained using all available modalities. Even if the difference is not marked as for the brain tumor segmentation task, in this study it's possible to identify a modality whose discharge decreases the overall performances more with respect to the other modalities. By expressing the results as a percentage of reduction of the Dice score of the models with respect to the full model, the decrease caused by the removal of ADC and FLAIR modalities is extremely low, between 2% and 3% (precisely 2.7% and 3.3% for ADC and FLAIR respectively), while the reduction caused by using all images except DWI for training is about 10.5%. It must be specified that, even in this latter case, the decrease is really low and all models can be considered comparable with the full one, but a prior intuition can be that stroke lesions are more evident in DWI images, and this modality could thus be more efficient and could allow to extract more relevant information for the segmentation of those lesions.

To conclude, the training time of models trained by removing one modality at a time is more or less two thirds of the time employed by the full model, as expected because each model is trained with two thirds of the number of images used by the full model.

	Model trained without ADC images	Model trained without DWI images	Model trained without FLAIR images		Full model
Lesion label	0.679	0.625	0.675		0.698
Training time	1h 40 min	1h 41 min	1h 44 min		2h 38 min

Table 5.6: Comparison of the Dice scores and training times obtained by models trained by removing one modality at a time (ADC, DWI and FLAIR) with respect to the full model, trained using all available images.

5.2.2 Training of nnUNet with a single modality

The Dice scores obtained by nnUNet networks trained with a single modality, compared with the full model are showed in *Table 5.7*, together with the time employed for training each configuration. It's straightforward to notice that the results reflect the expectations for almost all modalities: the two models trained respectively only with ADC and FLAIR images obtain really low Dice scores when trying to segment the stroke lesions, with an average value which is almost one half than the performances obtained by the full model. It means that these two modalities are not able to properly identify and segment the ischemic stroke lesions when used alone, but instead benefit a lot from the conjunction with other modalities.

Moreover, analyzing more in details these results, it was also noticed that in more than one testing case, models trained only with images of ADC or FLAIR modality achieve a Dice score of 0.0, being not able to segment the lesion at all, and localizing it in a totally different position. This is probably due to the way in which the lesion is highlighted in these modalities, because the model confuses it as a part of brain anatomy and isn't able to identify it, combined with the fact that, in many of these cases, the size of the lesion is pretty small.

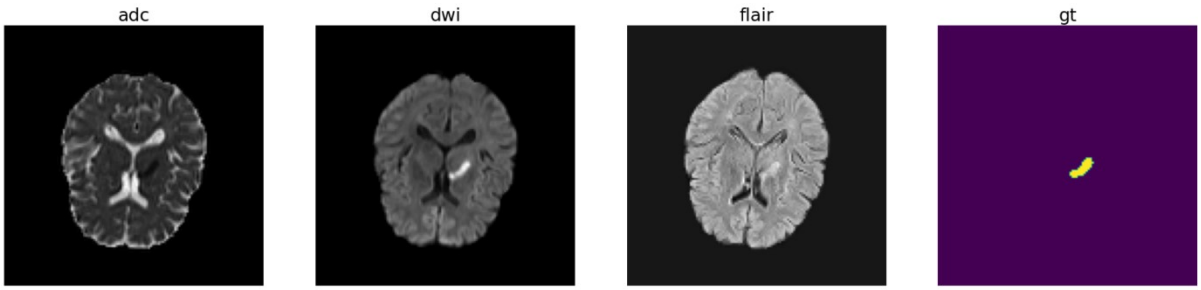


Figure 5.3: Visualization of the ADC, DWI and FLAIR scans for subject number 232 of the ISLES 2022 dataset, and corresponding segmentation map. It's possible to notice how the DWI image can highlight the lesion, while it's confused with cerebral regions in the other two scans.

While, on the other hand, the reduction of the Dice score obtained by the model trained with DWI images is significantly lower than the other two modalities, and it achieves comparable results with the full model. In *Figure 5.3* it's showed a specific testing case, in which the model trained with ADC images obtains a Dice score really close to 0 (0.06), and the model trained with FLAIR modality achieves a Dice score of 0, even if the dimension of the lesion is not so small. DWI image, on the other side, is better able to highlight the lesion and in this specific case obtains a Dice score of 0.67.

These results strengthen the idea initially developed in the analysis of models trained with all but one modality: models trained with DWI images are better able to localize ischemic stroke lesions with respect to the other modalities, because lesions are pointed out and underlined more in detail in those images.

	Model trained only with ADC images	Model trained only with DWI images	Model trained only with FLAIR images		Full model
Lesion label	0.416	**0.630**	0.396		0.698
Training time	1h 26 min	1h 31 min	1h 32 min		2h 38 min

Table 5.7: Comparison of the Dice scores and training times obtained by models trained with images of a single modality (ADC, DWI and FLAIR respectively) with respect to the full model, trained using all available images. The Dice score obtained by the network trained with DWI images is highlighted in bold because it's the only modality able to obtain comparable results with respect to the full model

5.2.2.1 Equalization of computational times

The Dice scores obtained by models trained only with ADC, DWI or FLAIR images, and by the network fed with all those three modalities, but cutting off its training to the average training time of single models, are showed in *Table 5.8.* Following the trend of the brain tumor segmentation task, also in this case the results don't add much information with respect to the case in which the training of the full model wasn't stopped. As a matter of fact, it can be noticed that the decrease of the Dice score obtained by the model whose training was stopped at 66 epochs with respect to the complete model is just of 0.05, giving credit to the hypothesis for which, even for stroke lesions segmentation, the nnUNet model stabilizes before reaching the threshold of 100 epochs.

Even if the results could seem to be useless, because the same observations carried out in the previous paragraph can be reported here, in reality this study allows a more correct comparison between models with respect to the study in which the training of the complete model wasn't cut off, because unlike the previous analysis, in this current investigation it is given to all models the same amount of time to learn from input images, equalizing then computations.

	Model trained only with ADC images	Model trained only with DWI images	Model trained only with FLAIR images		Full model equalizing training time	Full model
Lesion label	0.416	0.630	0.396		0.693	0.698

Table 5.8: Comparison of the Dice scores obtained by models trained with a single modality at a time (ADC, DWI and FLAIR) with respect to the full model, trained using all available images, with and without limiting its training time to a threshold set to be the average training time of single models.

5.2.2.2 Equalization of number of training images

The other exploited method to equalize computations between tested models was to balance the number of images used to train each network. In *Table 5.9* the Dice scores obtained by models trained with images of a single modality (ADC, DWI, FLAIR) and of the configuration trained with all these images are showed. Since the full model was trained using 70 subjects, each consisting in images of three modalities, to equalize the computations between networks,

models trained with images of a unique modality were trained tripling the input images, from 70 to 210. This study allows a better comparison between models, and giving to each architecture an input of the same dimension enables all models to learn useful features from a larger quantity of images, which could be better representative of the stroke lesions variability. Unfortunately, the performances of models trained with ADC and FLAIR images didn't improve so much, being not able to reach comparable results with respect to the full model. On the other hand, the unique network whose segmentation results became comparable with the full model was the one fed with DWI images: the model trained with those images, in fact, obtained a Dice score able to overcome the one obtained by the full model, even if only by 0.001.

The improvement of performances obtained by the model fed with only DWI images is so small that can be considered meaningless, though it suggest a relevant consideration: based on the assumptions and training schedules followed in this study, training a model for ischemic stroke lesion segmentation using ADC, DWI and FLAIR images leads to get more or less the same results obtained by training a model using the same amount of DWI images, which is the most meaningful and informative modality for stroke lesions.

	Model trained tripling ADC images	Model trained tripling DWI images	Model trained tripling FLAIR images		Full model
Lesion label	0.490	**0.699**	0.505		0.698

Table 5.9: Comparison of the Dice scores of models trained with 210 images of a single modality with respect to the full one, trained with 70 subjects (210 images) in the segmentation of ischemic stroke lesions. Segmentation results obtained by single models remain lower than the full model, except for the model fed with DWI images, which is able to match the performances obtained by the full model (in bold)

5.2.2.3 Analysis of the consistency of results

The analysis of nnUNet models trained with images of a single modality (ADC, DWI and FLAIR) was repeated using a proportionally decreasing number of training images, to demonstrate the consistency of the results and verify if the relationship between models' performances was stable over the number of images used to train them.

Starting from the previously analyzed case, where single models were trained with 210 images of single modalities, while the full model used 70 subjects (each consisting of three modalities),

the following variants were trained: single models trained with 108, 54 and 27 images and full model respectively with 36, 18 and 9 subjects.

The trend of the Dice scores for the unique class of the stroke lesion are showed in *Figure 5.4*. All models, including the full one, seem to have an increasing trend over the number of training images, but the inferiority of the results obtained by the ADC and FLAIR models is evident. On the other hand, the DWI model tends to have higher performances than the full network for all the range of training images tried, even if the gap is really low, and even if the difference between them with 360 images is extremely limited, as underlined in the previous paragraph. This result strengthens the idea for which, between all models trained with images of a single modality for the stroke lesion segmentation task, the DWI is the only one that can reach comparable performances or even overcome the full one.

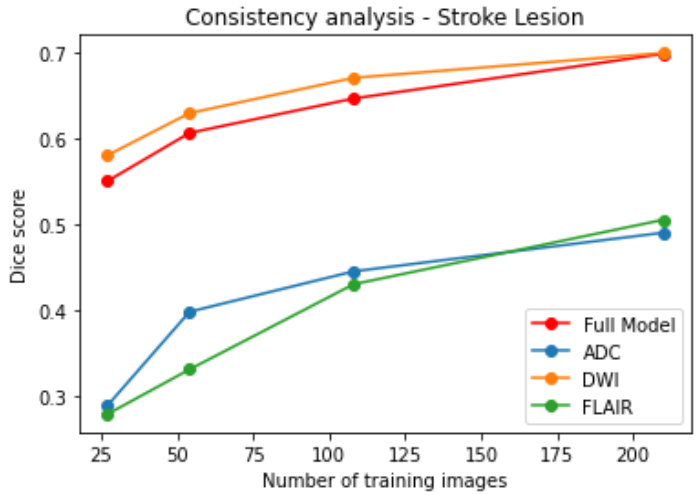


Figure 5.4: Consistency analysis for the stroke lesion class. Four different models were trained (three with images of a single modality, which can be ADC, DWI or FLAIR, and a full model using all those) with an increasing number of training images, specifically: 27, 54, 108 and 210. After that, their performances were compared and the trend of the Dice scores for the different number of training images is showed for the unique label of the ischemic stroke lesion

5.2.3 Ensemble of nnUNet models trained with a single modality

All the ensemble techniques introduced in Chapter 4 were tried also for the stroke lesion segmentation task, but, unfortunately, the results were not as good as for the brain tumor segmentation task.

Having a single class, and so dealing with a binary classification problem for the identification of the label of each voxel, the room for ensemble learning reduces drastically: as emerged previously, the best performing model also on the validation set was the one trained with DWI images. Implementing the averaging techniques, with weights or not, doesn't produce relevant results: the good segmentation results produced by the model trained with DWI images are indeed soiled by the interaction with the other single models, which inevitably reduces the performances due to their lower capabilities.

The same considerations can be done on the majority voting technique and all the different methods explained in Chapter 4. In all those cases, the contamination of the DWI network with the other single models could only lower the performances.

The negative side of the application of ensemble learning in the stroke lesion segmentation field, which blocks its application with this task, is the presence of this unique label: while in the brain tumor segmentation task, the ensemble of models techniques could exploit the capability of different models to perform better in the segmentation of one class over another, merging their different segmentation abilities, in the stroke lesion segmentation field the presence of a single label allows to identify a single model performing better than the others on that class, while the combination with the other models, with worse performances, can only reduce the results.

It follows that, with the ensemble learning strategy, the performances of the model are always worse than the full model, except in the case in which the ensemble of models follows the best performing model (trained with DWI images), and in this case its corresponding Dice score (0.699 against 0.698 of the full model) is obtained.

5.3 Dense Multi-path nnUNet

The architecture of IVD-Net, a network developed by Dolz et al. for the segmentation of the intervertebral disc (IVD) and subsequently adapted for the segmentation of 2D images of stroke lesions, was merged with nnUNet for the generation of Dense Multi-path nnUNet. Different architectures were created and trained, based on the combination of different aspects of these networks, and on their modification to increase efficiency and decrease the computational cost. The large part of these alterations was tried for brain tumor segmentation, and then the best performing ones were adapted also for ischemic stroke lesion segmentation, allowing to identify

a new baseline for both the tasks, able to overcome the performances obtained by the simple nnUNet.

5.3.1 Brain Tumor Segmentation task

Different architectures derived from the combination of IVD-Net and nnUNet were tried for brain tumor segmentation, and the average Dice scores obtained during inference for the segmentation of the three tumoral subregions (ED: peritumoral Edematous tissue, NCR: Necrotic tumor Core, ET: Enhancing Tumor), and their average, are showed in *Table 5.10*. All architectures are obviously compared between each other and with the baseline, the original nnUNet, represented in the last column.

As can be seen, by simply modifying the structure of nnUnet and disentangling the input data and feed them into four different encoders, concatenated through a bridge to a unique decoder, the performances approach the basic model, but without being able to reach it. On the other hand, when trying to complicate the model by adapting the nnUNet structure to the one of IVD-Net, the performances increase considerably, being able, for each architecture and modification proposed, to overcome the performances of the original nnUNet in the segmentation of all single classes and, of course, on average. After simply combining nnUNet with IVD-Net, it was tried to implement some modifications with the aim of increasing the efficiency, decrease the computational cost, and at the same time, increase the capability of the network to keep separated the input images acquired with different acquisition modalities.

A substantial improvement (leading to an increase of the average Dice score from 0.690 to 0.697) was obtained by modifying the dense connections between the different levels of the encoders (see *Chapter 4.6.1*) and using the tensor obtained by concatenating the outputs of the same level of the different streams, and compressing the result, as skip connection at each level, as showed in *Figure 4.74*.

The last network, obtained by using path-specific skip connection, using indeed stream-specific compressions of the concatenated tensor, as showed in *Figure 4.75*, is the one obtaining the best performances. Its training was repeated to demonstrate the consistency of the results and this model was chosen as baseline, being able to overcome all the other tried alternatives, and to largely overtake the results obtained by the original nnUNet.

	nnUNet with 4 encoders	nnUNet combined with IVD-Net	Dense Multi-path nnUNet modified skips according to *Figure 4.74*	Dense Multi-path nnUNet modified skips according to *Figure 4.75*	Last case repeated		Original nnUNet
ED	0.692	0.744	**0.751**	0.748	0.748		0.735
NCR	0.539	**0.589**	0.588	0.579	0.576		0.523
ET	0.693	0.736	0.753	**0.788**	0.774		0.716
Mean	0.641	0.690	0.697	**0.705**	0.699		0.658

Table 5.10: Average Dice scores obtained in the testing phase in the segmentation of the three tumoral subregions (ED: peritumoral Edematous tissue, NCR: Necrotic tumor Core, ET: Enhancing Tumor) and their mean, achieved by the different architectures tried combining IVD-Net and nnUNet, compared to the original nnUNet. It is evident that all networks, with the exception of the one obtained by only adapting nnUNet to have 4 encoders, are able to overcome the original nnUNet. The bold values are the best Dice scores reached for each class (or on average).

5.3.2 Stroke Lesion Segmentation task

The same analyzes performed for brain tumor segmentation, were repeated for stroke lesion segmentation, and the average Dice scores achieved in the test set for the segmentation of the ischemic stroke lesion, by the different network variants implemented merging IVD-Net and nnUNet, are shown in *Table 5.11* compared to the original nnUNet.

The first and simplest architecture, implemented by splitting the encoder into three paths, one for each input modality, isn't able to achieve comparable performances with respect to the original nnUNet.

Unfortunately, also the architecture realized by adapting the nnUNet structure to the one of IVD-Net, can't reach the Dice score obtained by the original nnUNet, even if it gets closer with respect to the previous case. The last two architectures, in which the dense connections between the different levels of the encoders were modified to increase the efficiency and decrease the computations, while the skip connections were changed in two distinct ways, as showed in *Figures 4.74* and *4.75,* are the most promising ones, being able to achieve the best performances for the brain tumor segmentation task. As a matter of fact, both these methods are able to overcome the original nnUNet, allowing to identify a new baseline, which consists in the model where the skip connections at each level are identified as the tensor obtained by concatenating

the outputs of the previous level of the encoder (*Figure 4.74*), which represents the best performing model for the stroke lesion segmentation task, and the new baseline chosen.

	nnUNet with 3 encoders	nnUNet combined with IVD-Net	Dense Multi-path nnUNet modified skips according to *Figure 4.74*	Dense Multi-path nnUNet modified skips according to *Figure 4.75*		Original nnUNet
Stroke Lesion	0.677	0.692	**0.710**	0.700		0.698

Table 5.11: Average Dice scores obtained in the testing phase in the segmentation of stroke lesions, achieved by the different architectures tried combining IVD-Net and nnUNet, compared to the original nnUNet. Only the last two networks (whose Dice scores are highlighted in bold) are able to overcome the performances of the original nnUNet.

5.4 Inter-pathology Learning

The alternative models trained for stroke lesion segmentation using the Inter-pathology Learning technique from the corresponding model trained for brain tumor segmentation, both using FLAIR images, were evaluated using a test set composed by 40 images extracted from ISLES 2022, as performed in previous analyzes. The average Dice scores produced by these networks are showed in *Table 5.12,* compared with the results obtained by the original model trained with the same FLAIR training images.

It can be observed that implementing the transfer learning technique and maintaining the same weights learnt in the encoder for brain tumor segmentation doesn't improve the capability of the model to identify and segment precisely stroke lesions, causing a decrease of the overall Dice score. This is an important result, meaning that the features learnt in the encoder of the nnUNet model trained for the brain tumor segmentation task are not useful for the correspondent model for stroke lesion segmentation, underlining the difference of the two pathologies and the inadequacy of considering the former task as a generalization of the latter one.

But these results don't mean that brain tumor and stroke lesion segmentation are two completely independent and separated tasks. Transfer Learning can still be helpful: by considering the previous annotation, and noticing that the difference of the two tasks but also the availability of

a huge amount of data, it was chosen also to apply transfer learning by considering the nnUNet structure used for brain tumor segmentation also for stroke lesion segmentation, but without freezing the pretrained weights of the first task, and instead using them just as an initialization for the second task. In this way, allowing the network to learn features related to stroke lesions, starting from a structure and weights learnt from brain tumors, the performance of the network improves, as underlined in *Table 5.12*.
This result highlights the connection between these two tasks, and the capability to increase the performances of a model trained for stroke lesion segmentation, if a model for brain tumor segmentation is available and was already trained with the same type of images.

	nnUNet using brain tumor segmentation pretrained weights	nnUNet using brain tumor segmentation pretrained weights and freezing encoder		nnUNet trained with FLAIR images
Stroke Lesion	**0.440**	0.366		0.396

Table 5.12: Average Dice scores obtained in the testing phase in the segmentation of stroke lesions, achieved by models obtained implementing inter-pathology learning from the model trained with FLAIR images for brain tumor segmentation, trying to freeze or not the encoder's weights, compared to the original nnUNet trained with FLAIR images. Only the first network, where weights were not frozen (whose Dice scores are highlighted in bold) is able to overcome the performances of the original nnUNet.

5.5 Final models

In this section are reported the results of the best performing models, between the previously tested ones, employing improved training datasets and increasing also the training epochs.

5.5.1 Brain Tumor Segmentation

Table 5.13 summarizes the Dice scores obtained by the original nnUNet, compared to the Dense Multi-path nnUNet, on the three tumoral subregions (ED, NCR and ET) and on average, on the

test set obtained from the FeTS 2022 dataset. The training times and the number of trainable parameters for the two models are also reported.

By letting nnUNet learn from a large training set and increasing the number of training epochs in which the network can optimize itself, good performances can be obtained, reaching an average Dice score of 0.868. On the other hand, the large computational cost of Dense Multi-path nnUNet has been decreased for this task, not to stumble in an out of memory problem, by reducing the number of channels for each level: in this way, its number of trainable parameters is still larger than nnUNet, as can be seen in *Table 5.13*, but the difference is not so huge, javing also that the training time of Dense Multi-path nnUNet is lower.

Under all these circumstances, Dense Multi-path nnUNet is still able to overcome the performances of nnUNet on all classes (with the exception of the first one in which they equalize) and on average.

These results underline the strength and the potential of Dense Multi-path nnUNet, and the possibility to continue to increase its performances by increasing again the number of channels and restore the original values automatically set by nnUNet, if the available GPU memory allows it.

	nnUNet	Dense Multi-path nnUNet
ED	0.886	0.886
NCR	0.816	**0.823**
ET	0.902	**0.903**
Mean	0.868	**0.871**
Training time	38h 35min	37h 27 min
Number of parameters	31˙198˙976	70˙468˙080

Table 5.13: Average Dice scores obtained by nnUNet and Dense Multi-path nnUNet in the segmentation of the three tumoral subregions (ED: peritumoral edematous/invaded tissue, NCR: necrotic tumor core, ET: enhancing tumor), and on average, in a test set extracted from the FeTS 2022 dataset. The training times and the number of trainable parameters of the two models are also report

The segmentation maps produced by nnUNet and Dense Multi-path nnUNet, compared with the corresponding ground truths, of some random subjects extracted from the test set are showed in *Figure 5.5.* The related T1ce images are also reported, being the ones which are more capable of highlighting the brain tumoral subregions.

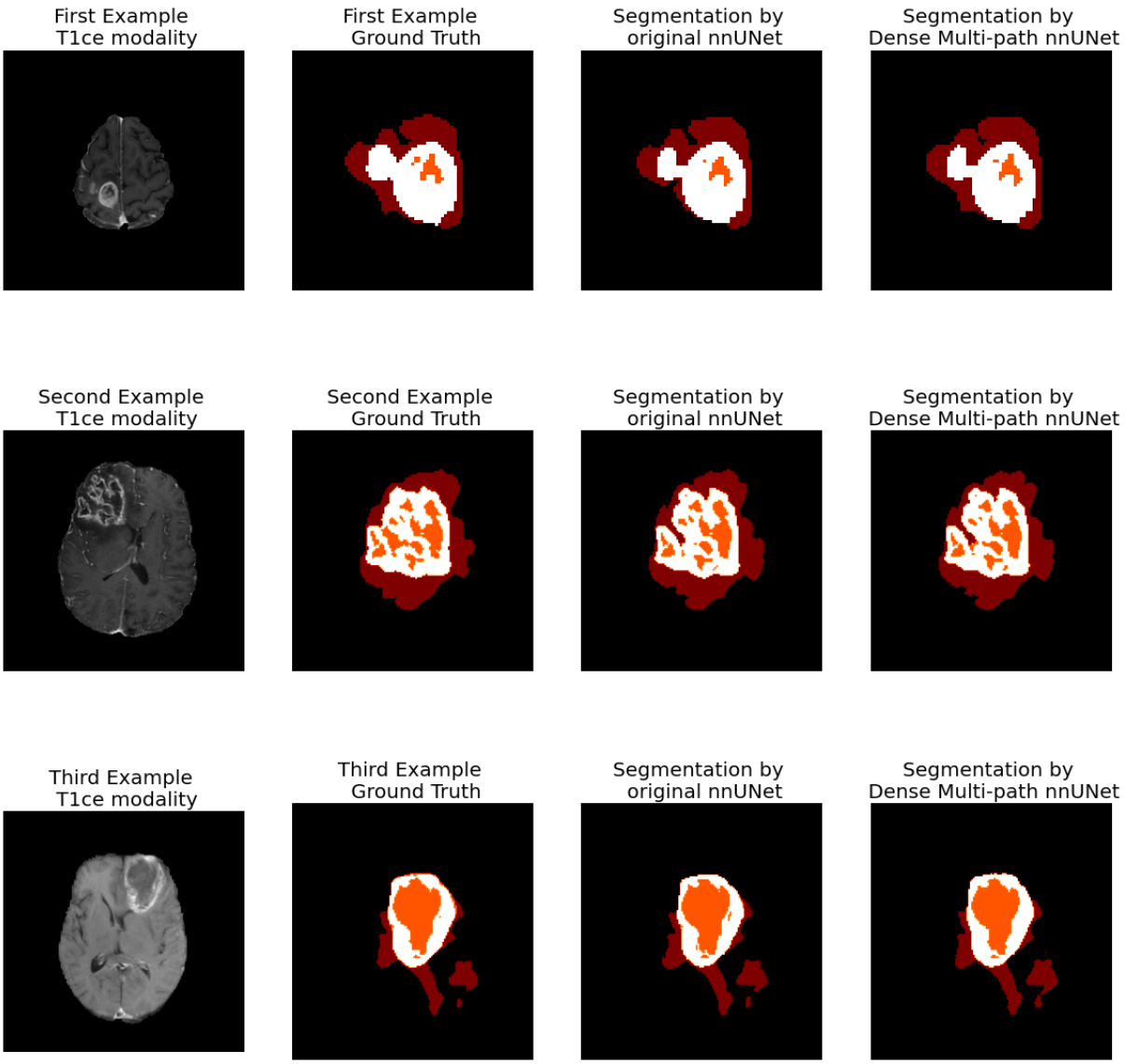


Figure 5.5: Comparison of the ground truth segmentations with the segmentation maps produced by the original nnUNet and Dense Multi-path nnUNet for three random cases extracted from the test set. For each segmentation map a zoom of the tumoral region is shown, to better visualize the differences between the segmented subregions. The corresponding T1ce images are also showed, because brain tumor subregions are more visible in this modality. Moreover, the "gist_heat" matplotlib color map has been used to better highlight the different regions: red represents the peritumoral edematous/invaded tissue (ED), white the enhancing tumor (ET) while orange the necrotic tumor core (NCR).

5.5.2 Stroke Lesion Segmentation

The average Dice scores obtained by the original nnUNet and Dense Multi-path nnUNet in the segmentation of stroke lesions on a test set derived from the ISLES 2022 dataset is reported in *Table 5.14*. The training times are also showed for the two networks, remembering the original

nnUNet model for stroke lesion segmentation was the first trained model for these final analyses, and was thus trained with 1000 epochs. After that, it was chosen to set the number of training epochs to 600, given that the validation loss didn't change considerably in the last 400 epochs. Anyway, even if the number of epochs for the Dense Multi-path nnUNet is almost halved, its training time is superior to nnUNet, underlining a large computational cost of this last network, justified by its huge number of trainable parameters, as can be seen in *Table 5.14,* with resepct to nnUNet.

The results are consistent with the previous analyses, with the Dense Multi-path nnUNet being able to beat the performances of the original nnUNet.

It must be specified that the Dice scores are inferior to the ones obtained with a smaller training dataset and with just 100 training epochs because, in that case, training and test images were picked sequentially, while in this last case they were selected randomly, removing any possible correlation between consecutive images that could influence the results.

	nnUNet	Dense Multi-path nnUNet
Stroke Lesion	0.635	**0.660**
Training time	35h 13min (1000 epochs)	36h 47min
Number of parameters	16˙548˙832	294˙755˙520

Table 5.14: Average Dice scores obtained by nnUNet and Dense Multi-path nnUNet in the segmentation of stroke lesions in a test set extracted from the ISLES 2022 dataset. The training times and the number of trainable parameters of the two models are also reported.

Figure 5.6 also reports a comparison of the segmentation maps produced by the original nnUNet and by Dense Multi-path nnUNet, with the ground truth segmentation, for three random cases extracted from the test set.

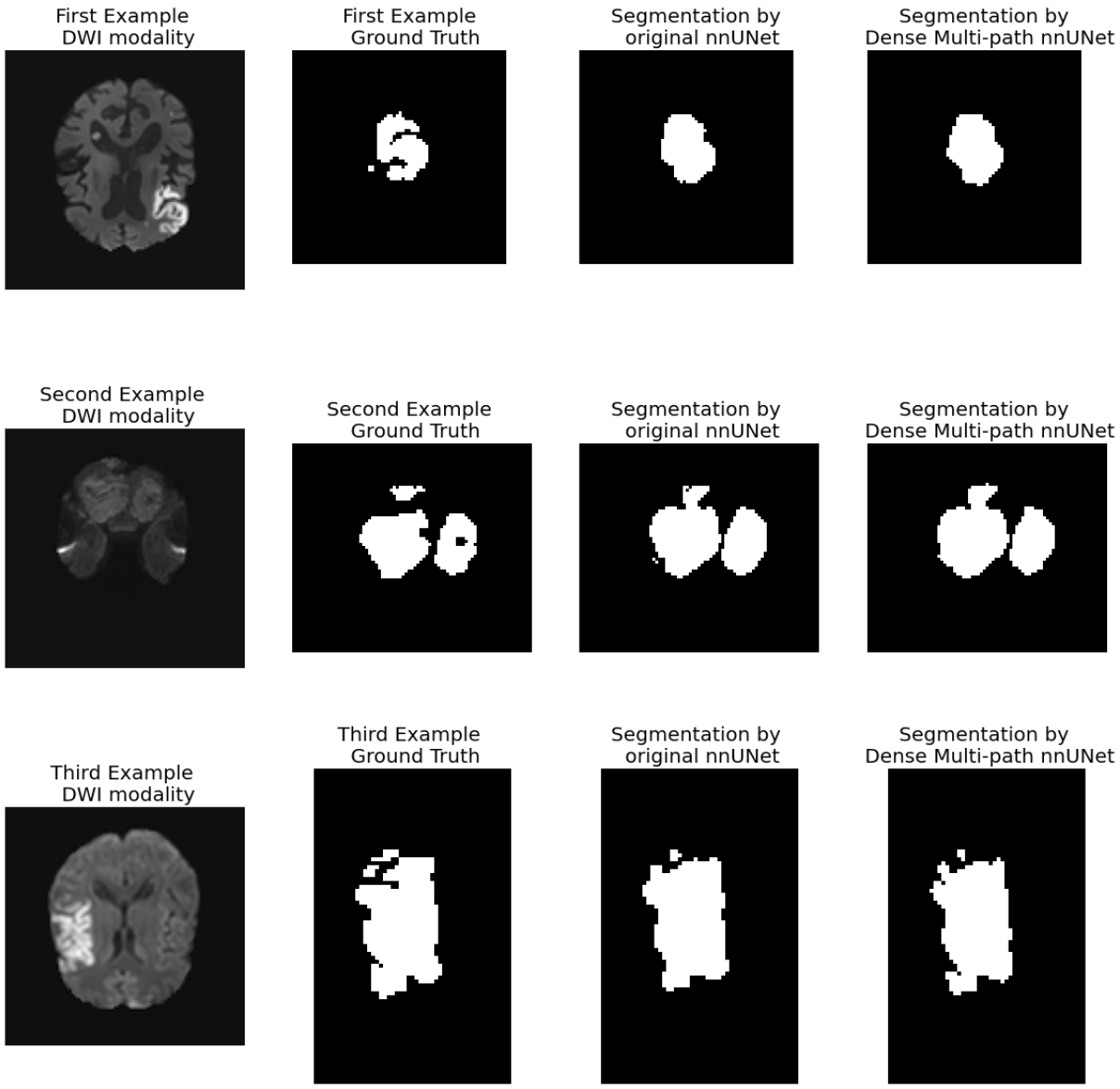


Figure 5.6: Comparison of the ground truth segmentations with the segmentation maps produced by the original nnUNet and Dense Multi-path nnUNet for three random cases extracted from the test set. For each segmentation map a zoom of the stroke lesion is shown, to better visualize the differences between the segmented regions. The corresponding DWI images are also showed, because ischemic strokes are more visible in this modality.

5.5.3 Inter-pathology Learning

The comparison between the average Dice scores obtained in the segmentation of stroke lesions by the original nnUNet trained with FLAIR images, and the nnUNet network trained with the Inter-pathology Learning technique transferring the knowledge learnt from the model trained for brain tumor segmentation with FLAIR images, are visible in *Table 5.15*.

The results are consistent with the previous tests, with the second model being able to overcome the performances of the first one. They also strengthen the assumption for which, if a nnUNet model is available and was already trained for brain tumor segmentation with FLAIR images, it can be used as a basis applying an Inter-pathology Learning technique with the corresponding stroke lesion segmentation model trained with FLAIR images, increasing its performances with respect to its traditional training.

	nnUNet trained with FLAIR images	nnUNet trained with FLAIR images with Inter-pathology Learning
Stroke Lesion	0.509	**0.583**

Table 5.15: Average Dice scores obtained by the basic nnUNet model and the network obtained applying the Inter-pathology Learning strategy, both trained with FLAIR images, in the segmentation of stroke lesions in a test set extracted from the ISLES 2022 dataset.

The segmentation maps produced by the two models, compared with the ground truths, are also reported in *Figure 5.7* for three subjects sampled from the test set. The corresponding FLAIR images are showed too, being the only acquisition modality used in this study.
In the first case it can be observed that, even if the segmentation map produced by the last model isn't optimal, it's still much better than the one generated by the original nnUNet. While in the last subject, the nnUNet isn't able to identify the stroke lesion at all, obtaining an average Dice score of 0.0, while the transfer Learning technique allows an accurate localization of the lesion, with an average Dice score of 0.622.

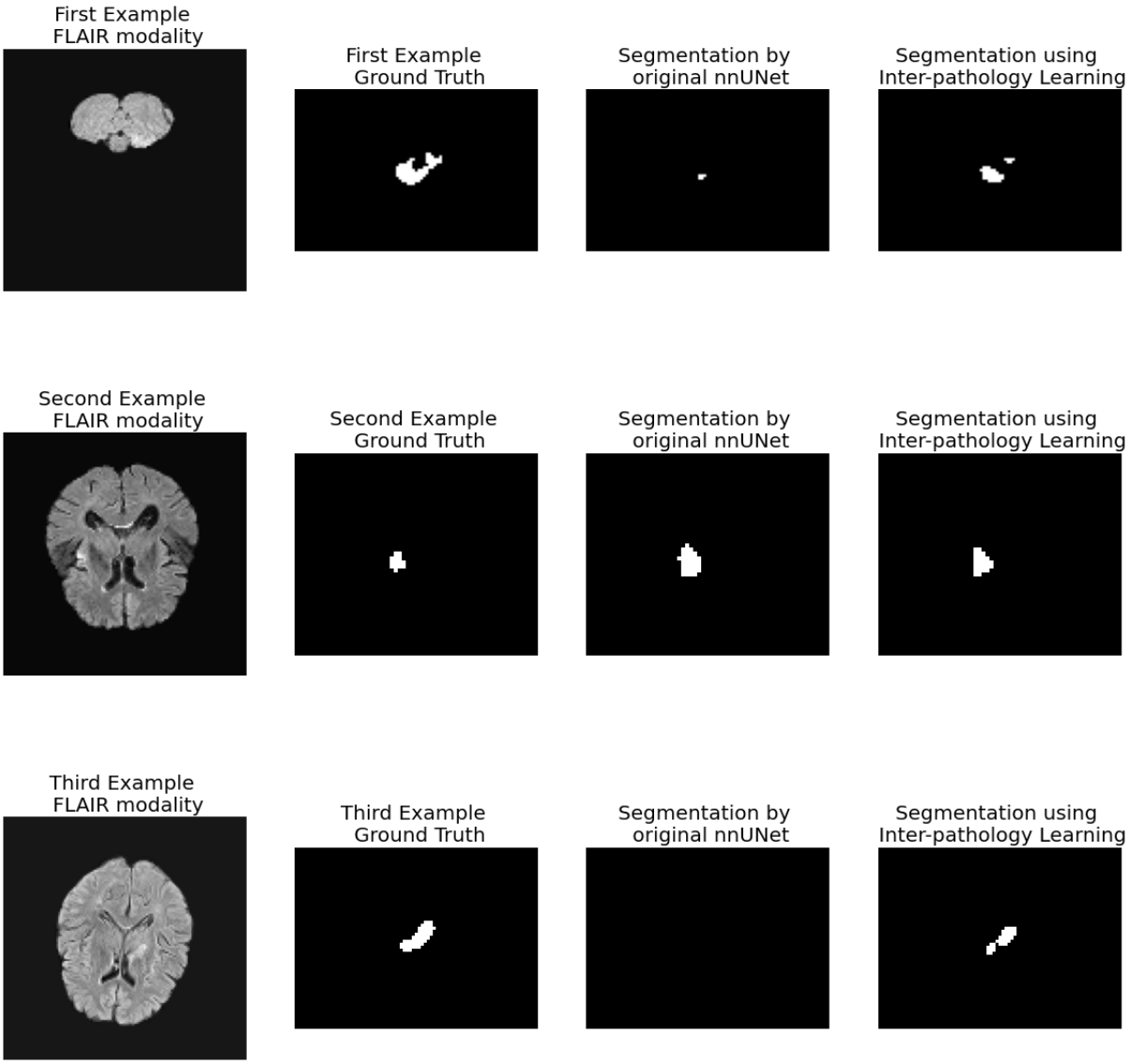


Figure 5.7: Comparison of the ground truth segmentations with the segmentation maps produced by the original nnUNet model and the nnUNet trained using the Inter-pathology Learning technique from brain tumor segmentation, both trained with FLAIR images, for three cases extracted from the test set. For each segmentation map a zoom of the stroke lesion is shown, to better visualize the differences between the segmented regions. The corresponding FLAIR images are also showed.

6. Discussion

Accurate delineation and segmentation of brain tumors and ischemic stroke lesions represent a crucial aspect for diagnosis, treatment planning and subsequent evaluations of entity and consequences of brain lesions or tumors. Since manual depiction of medical images is time consuming and laborious, many automatic methods have been developed for those tasks, achieving important results for both of them.

After a meticulous analysis of the most recent and advanced techniques in these fields, an optimal baseline for the subsequent research was identified in nnUNet ("no-new-UNet"), given its capability to win the latest versions of the Brain Tumor Segmentation challenge (2020, 2021) and setting a new state-of-the-art in the segmentation of many other medical images. Its great performances are due to its capacity to adapt the architecture of the model to the dataset to which it's applied, identifying the best possible UNet configuration for the specific task under examination.

All the analyses were performed using common training and test sets for the two tasks of brain tumor and stroke lesion segmentation, extracted from BraTS 2020, FeTS 2022 and ISLES 2022 datasets.

nnUNet was submitted to the BraTS 2020 challenge with some minor modifications, some of which were introduced with the only aim of increasing the segmentation results based on the specific BraTS evaluation metrics, while others targeting a better adaptation of nnUNet to the brain tumor segmentation task, including a more aggressive data augmentation, the substitution of instance normalization with batch normalization and the introduction of the Batch Dice loss. These modifications were implemented and evaluated, pointing out their incapability to increase much the overall performances if used alone. During the investigations about the application of nnUNet for brain tumor segmentation, it was observed an increase of the average Dice score obtained in the test set from 0.613 to 0.659 when coupling these BraTS specific modifications with the *Wassertian Dice Loss*, a peculiar Dice Loss introduced by Fidon et al. (Fidon et al., 2021) which takes advantage of the hierarchical structure of BraTS labels.

These results show the ability of the Wassertian Dice Loss to improve the segmentation results of nnUNet but only when coupled with BraTS specific settings and applied for brain tumor segmentation, identifying a new alternative for the classic Dice Loss.

However, the main goal of this thesis was the detachment of the input images into the different available acquisition modalities before feeding them inside the model, allowing the network to extract specific features from the modalities' images while keeping them separated. This study was performed due to the big differences between the highlighting of stroke lesions or brain tumoral subregions between the available modalities inside the used datasets: nnUNet, as many other networks, treats images of different modalities as different input channels, merging their information at early stages. In this way, features extracted from images in which lesions are highlighted in different, or even opposite ways, are combined and mixed.

With this aim, the most relevant acquisition modalities for the considered tasks were identified by training nnUNet models for brain tumor or stroke lesion segmentation with all the available modalities but one, and with just images of one modality at a time. In particular, for the brain tumor segmentation task, T1ce was pointed out as the most useful modality, given by comparing the results of a nnUNet model trained with all available modalities (FLAIR, T1, T1ce and T2) and a nnUNet network trained with an equal number of T1ce images, the latter one obtained the best performances on the segmentation of the Tumor Core (Necrotic Tumor Core + Enhancing Tumor) and on average. This analysis was repeated for an increasing number of training images, validating the results for the full tested range. These results underline the superiority of the T1ce modality in the delineation of the brain tumor subregions with respect to the other modalities, and the possible contamination of its performances when combining T1ce images with the other modalities (especially T1 and T2), decreasing the overall performances. The general message is that, if a large amount of T1ce images is annotated and therefore available for brain tumor segmentation, it is preferable to use only those images instead of combining them with the corresponding images of other modalities.

On the other side, for stroke lesion segmentation the most meaningful modality was found to be DWI. Also in this case, by training a model with all available modalities (ADC, DWI and FLAIR) and training another nnUNet network with the same number of images, but only DWI, the second one was able to overcome the performances of the full model, for the whole range of training images tried. DWI has thus the same role for stroke lesion segmentation, that T1ce has for brain tumor segmentation, leading to point out the DWI modality as the most effective one in the highlighting of stroke lesions.

The contamination of the most performing acquisition modalities with the others, is probably due to the way in which the extracted features are mixed in the early stages of the network. The fusion of information extracted from images of different modalities can still be useful, but different ways of combining it were explored.

The first idea was to develop an ensemble of models, each trained with images of single modalities, and then combined at inference level. This road was walkable only for brain tumor segmentation, where it was possible to combine models performing better for one class over the others; while for stroke lesion segmentation, the ensemble of models wasn't able to overcome the performances of the best model trained with images of a single modality (DWI) so didn't produce significant results. On the other hand, for brain tumor segmentation, after having tried different ensemble techniques, a complex majority voting method was able to produce good results, being able to overcome the model trained only with T1ce images on average and on the segmentation of the first class, but not for the segmentation of the tumor core, and also without being able to reach the Dice scores of the full nnUNet model (average Dice score of 0.647 against 0.658). However, these results highlight that there is still room for improving the Ensemble technique, trying to approach the performances of the T1ce model for the second and third class and, as a consequence, overcome the full model.

The most important model developed in this thesis is the *Dense Multi-path nnUNet*, obtained by combining the architectures of nnUNet and IVD-Net, a UNet developed by Dolz et al. in which the encoder is split in N streams, in which N corresponds to the number of input modalities, which flow into a bridge that finally convey the information to a unique decoder. This structure, with the presence of dense connections between paths, allows to keep the input modalities separated, and however combine the information extracted from them. After the introduction of some modifications to increase the efficiency and decrease the computational cost, this model was able to obtain excellent results.

On the final evaluation, increasing the number of training epochs and of training images, the Dense Multi-path nnUNet was able to overcome the performances of the simple nnUNet obtaining a Dice score of 0.660 against 0.635 for stroke lesion segmentation, and an average Dice score of 0.871 against 0.868 for brain tumor segmentation.

This last model demonstrates that, keeping the input modalities separated but combining the extracted features to better model their relationships, allows to increase the performances of an optimal network such as nnUNet.

Finally, the presence of FLAIR images both for brain tumors and stroke lesions segmentation, suggested the implementation of an *Inter-pathology Learning* strategy between the two tasks, with the following result: by training a model for brain tumor segmentation with FLAIR images and applying transfer learning without freezing any weight for the training of the corresponding model for stroke lesion segmentation using only FLAIR images, the Dice score increased from 0.509 (obtained by the same model without using transfer learning) to 0.583.

Possible future improvements could be the identification of other acquisition modalities useful for the segmentation of both brain tumors and stroke lesions, or even the usage of modalities with similar characteristics (with comparable intensities for the same regions) for the application of the Inter-pathology Learning method.

7. Conclusion

The principal objective of this thesis was the identification of alternative ways in which images of different acquisition modalities could be fed inside a network, by separating them and without fusing at early stages features extracted from images in which stroke lesions or tumoral subregions could be highlighted in different or even opposite ways. After the analysis of the best performing and most recently released methods for Brain Tumor and Stroke Lesion segmentation, nnUNet was selected as baseline architecture, representing the state-of-the-art for brain tumor segmentation but also in many other medical images segmentation fields. The combination of nnUNet with IVD-Net, a UNet with a number of streams (encoders) equal to the number of input modalities, with the introduction of modifications in skip and dense connections between paths, lead to the implementation of *Dense Multi-path nnUNet*. Comparing its performances with the original nnUNet using a test set extracted from FeTS 2022 dataset for brain tumor segmentation, and from ISLES 2022 dataset for ischemic stroke lesion segmentation, it was observed an increase of the performances for both tasks.

Dense Multi-path nnUNet therefore represents a very promising architecture, and possible future improvements could be the further reduction of its computational cost, the automatic adaptation of the number of streams from the number of input modalities, and the combination of the Wassertian Dice loss and BraTS specific settings in the segmentation of brain tumors.

Moreover, to exploit the knowledge learnt for Brain Tumor Segmentation also for Stroke Lesion Segmentation, an Inter-pathology Learning technique was developed between the two tasks, taking advantage of the availment of FLAIR images in both fields. This method increased the overall segmentation performances compared to the original case, leading to the result that, if a model is available and was already trained for brain tumor segmentation, transferring and retraining it for stroke lesion segmentation is better than training a model from scratch for this task.

References

Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C. C. le, Elhalawani, H., Jreige, M., Prior, J. O., Vallières, M., Visvikis, D., Hatt, M., & Depeursinge, A. (2022). *Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images*. http://arxiv.org/abs/2201.04138

Andrearczyk, V., Oreiller, V., Jreige, M., Vallì Eres, M., Castelli, J., Elhalawani, H., Boughdad, S., Prior, J. O., & Depeursinge, A. (n.d.). *Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT*. https://portal.fli-iam.irisa.fr/petseg-challenge/overview#_ftn1,

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., Prevedello, L. M., Rudie, J. D., Sako, C., Shinohara, R. T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., … Bakas, S. (2021). *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*. http://arxiv.org/abs/2107.02314

Baur, C., Wiestler, B., Albarqouni, S., & Navab, N. (2018). *Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images*. https://doi.org/10.1007/978-3-030-11723-8_16

Ben Naceur, M., Akil, M., Saouli, R., & Kachouri, R. (2020). Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. *Medical Image Analysis*, *63*. https://doi.org/10.1016/j.media.2020.101692

Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.-W., Han, X., Heng, P.-A., Hesser, J., Kadoury, S., Konopczynski, T., Le, M., Li, C., Li, X., Lipkovà, J., Lowengrub, J., Meine, H., Moltz, J. H., … Menze, B. H. (2019). *The Liver Tumor Segmentation Benchmark (LiTS)*. http://arxiv.org/abs/1901.04056

Chen, H., Qin, Z., Ding, Y., Tian, L., & Qin, Z. (2020). Brain tumor segmentation with deep convolutional symmetric neural network. *Neurocomputing*, *392*, 305–313. https://doi.org/10.1016/j.neucom.2019.01.111

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016). *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. http://arxiv.org/abs/1606.00915

Chen, S., Ding, C., & Liu, M. (2019). Dual-force convolutional neural networks for accurate brain tumor segmentation. *Pattern Recognition*, *88*, 90–100. https://doi.org/10.1016/j.patcog.2018.11.009

Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., & Lladó, X. (2018). *Acute and sub-acute stroke lesion segmentation from multimodal MRI*. http://arxiv.org/abs/1810.13304

Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., & Lladó, X. (2019). Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Computers in Biology and Medicine*, *115*. https://doi.org/10.1016/j.compbiomed.2019.103487

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., & Ye, C. (2019). *Semi-Supervised Brain Lesion Segmentation with an Adapted Mean Teacher Model*. http://arxiv.org/abs/1903.01248

Dolz, J., Ayed, I. ben, & Desrosiers, C. (2018). *Dense Multi-path U-Net for Ischemic Stroke Lesion Segmentation in Multiple Image Modalities*. http://arxiv.org/abs/1810.07003

Dolz, J., Desrosiers, C., & Ayed, I. ben. (2018). *IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet*. http://arxiv.org/abs/1811.08305

Donahue, J., Krähenbühl, P., & Darrell, T. (2016). *Adversarial Feature Learning*. http://arxiv.org/abs/1605.09782

Duong, M. T., Rudie, J. D., Wang, J., Xie, L., Mohan, S., Gee, J. C., & Rauschecker, A. M. (2019). Convolutional neural network for automated flair lesion segmentation on clinical brain MR imaging. *American Journal of Neuroradiology*, *40*(8), 1282–1290. https://doi.org/10.3174/ajnr.A6138

Fidon, L., Li, W., Garcia-Peraza-Herrera, L. C., Ekanayake, J., Kitchen, N., Ourselin, S., & Vercauteren, T. (2017). *Generalised Wasserstein Dice Score for Imbalanced Multi-class Segmentation using Holistic Convolutional Networks*. https://doi.org/10.1007/978-3-319-75238-9_6

Fidon, L., Shit, S., Ezhov, I., Paetzold, J. C., Ourselin, S., & Vercauteren, T. (2021). *Generalized Wasserstein Dice Loss, Test-time Augmentation, and Transformers for the BraTS 2021 challenge*. http://arxiv.org/abs/2112.13054

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., & Poggio, T. (n.d.). *Learning with a Wasserstein Loss*. http://cbcl.mit.edu/wasserstein.

Futrega, M., Milesi, A., Marcinkiewicz, M., & Ribalta, P. (2021). *Optimized U-Net for Brain Tumor Segmentation*. http://arxiv.org/abs/2110.03352

Gordillo, N., Montseny, E., & Sobrevilla, P. (2013). State of the art survey on MRI brain tumor segmentation. In *Magnetic Resonance Imaging* (Vol. 31, Issue 8, pp. 1426–1438). https://doi.org/10.1016/j.mri.2013.05.002

Gu, Y., Ruan, R., Yan, Y., Zhao, J., Sheng, W., Liang, L., & Huang, B. (2022). Glomerulus Semantic Segmentation Using Ensemble of Deep Learning Models. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-022-06608-9

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M. C., Dickie, D. A., Wardlaw, J., & Rueckert, D. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, *17*, 918–934. https://doi.org/10.1016/j.nicl.2017.12.022

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*. http://arxiv.org/abs/2201.01266

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2015). *Brain Tumor Segmentation with Deep Neural Networks*. https://doi.org/10.1016/j.media.2016.05.004

He, K., Zhang, X., Ren, S., & Sun, J. (n.d.). *Deep Residual Learning for Image Recognition*. http://image-net.org/challenges/LSVRC/2015/

Ho, J., Kalchbrenner, N., Weissenborn, D., & Salimans, T. (2019). *Axial Attention in Multidimensional Transformers*. http://arxiv.org/abs/1912.12180

Hu, K., Gan, Q., Zhang, Y., Deng, S., Xiao, F., Huang, W., Cao, C., & Gao, X. (2019). Brain Tumor Segmentation Using Multi-Cascaded Convolutional Neural Networks and Conditional Random Field. *IEEE Access*, *7*, 92615–92629. https://doi.org/10.1109/ACCESS.2019.2927433

Hu, X., Huang, W., Guo, S., & Scott, M. R. (n.d.). *StrokeNet: 3D Local Refinement Network for Ischemic Stroke Lesion Segmentation*.

Hu, X., Luo, W., Hu, J., Guo, S., Huang, W., Scott, M. R., Wiest, R., Dahlweid, M., & Reyes, M. (2020). *Brain SegNet: 3D local refinement network for brain lesion segmentation*. https://doi.org/s12880-020-0409-2

Hui, H., Zhang, X., Li, F., Mei, X., & Guo, Y. (2020). A Partitioning-Stacking Prediction Fusion Network Based on an Improved Attention U-Net for Stroke Lesion Segmentation. *IEEE Access*, *8*, 47419–47432. https://doi.org/10.1109/ACCESS.2020.2977946

Isensee, F., Jaeger, P. F., Full, P. M., Vollmuth, P., & Maier-Hein, K. H. (2020a). *nnU-Net for Brain Tumor Segmentation*. http://arxiv.org/abs/2011.00848

Isensee, F., Jaeger, P. F., Full, P. M., Vollmuth, P., & Maier-Hein, K. H. (2020b). *nnU-Net for Brain Tumor Segmentation*. http://arxiv.org/abs/2011.00848

Isensee, F., Jäger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2019). *Automated Design of Deep Learning Methods for Biomedical Image Segmentation*. https://doi.org/10.1038/s41592-020-01008-z

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., & Maier-Hein, K. H. (2018). *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*. http://arxiv.org/abs/1809.10486

Jia, H., Bai, C., Cai, W., Huang, H., & Xia, Y. (2022). *HNF-Netv2 for Brain Tumor Segmentation using multi-modal MR Imaging*. http://arxiv.org/abs/2202.05268

Jiang, Z., Ding, C., Liu, M., & Tao, D. (2020). Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11992 LNCS*, 231–241. https://doi.org/10.1007/978-3-030-46640-4_22

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., & Glocker, B. (2017). *Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation*. http://arxiv.org/abs/1711.01468

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A., Criminisi, A., Rueckert, D., & Glocker, B. (n.d.). *DeepMedic for Brain Tumor Segmentation*. https://github.com/Kamnitsask/deepmedic

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78. https://doi.org/10.1016/j.media.2016.10.004

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I. ben, Ca, I. B., & Montreal, ´ Ets. (2019). Boundary loss for highly unbalanced segmentation. In *Proceedings of Machine Learning Research* (Vol. 102). https://github.com/LIVIAETS/surface-loss

Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2017). An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 31–40. https://doi.org/10.1109/JBHI.2016.2635663

Li, H., Li, A., & Wang, M. (2019). A novel end-to-end brain tumor segmentation method using improved fully convolutional networks. *Computers in Biology and Medicine*, *108*, 150–160. https://doi.org/10.1016/j.compbiomed.2019.03.014

Liew, S.-L., Lo, B., Donnelly, M. R., Zavaliangos-Petropulu, A., Jeong, J. N., Barisano, G., Hutton, A., Simon, J. P., Juliano, J. M., Suri, A., Ard, T., Banaj, N., Borich, M. R., Boyd, L. A., Brodtmann, A., Buetefisch, C. M., Cao, L., Cassidy, J. M., Ciullo, V., … Yu, C. (n.d.). *A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms*. https://doi.org/10.1101/2021.12.09.21267554

Luu, H. M., & Park, S.-H. (2021). *Extending nn-UNet for brain tumor segmentation*. http://arxiv.org/abs/2112.04653

Magadza, T., & Viriri, S. (2021). Deep learning for brain tumor segmentation: A survey of state-of-the-art. In *Journal of Imaging* (Vol. 7, Issue 2). MDPI AG. https://doi.org/10.3390/jimaging7020019

Maier, O., Menze, B. H., von der Gablentz, J., Häni, L., Heinrich, M. P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H. L., Havaei, M., … Reyes, M. (2017). ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, *35*, 250–269. https://doi.org/10.1016/j.media.2016.07.009

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., & Stilla, U. (2016). SEMANTIC SEGMENTATION OF AERIAL IMAGES WITH AN ENSEMBLE OF CNNS. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *III–3*, 473–480. https://doi.org/10.5194/isprsannals-iii-3-473-2016

MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation Proceedings of MICCAI-BRATS 2012 October 1 st , Nice, France. (n.d.).

MICCAI_BraTS_2017_proceedings_shortPapers. (n.d.).

Moon, W. K., Lee, Y. W., Ke, H. H., Lee, S. H., Huang, C. S., & Chang, R. F. (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *190*. https://doi.org/10.1016/j.cmpb.2020.105361

Myronenko, A. (2018). *3D MRI brain tumor segmentation using autoencoder regularization*. http://arxiv.org/abs/1810.11654

Nvidia, A. H., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., Daguang, N., & Nvidia, X. (n.d.). *UNETR: Transformers for 3D Medical Image Segmentation*. https://monai.io/research/unetr

Oktay, O., Schlemper, J., Folgoc, L. le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). *Attention U-Net: Learning Where to Look for the Pancreas*. http://arxiv.org/abs/1804.03999

Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., Pekmezci, M., Schwartzbaum, J. A., Turner, M. C., Walsh, K. M., Wrensch, M. R., & Barnholtz-Sloan, J. S. (2014). The epidemiology of glioma in adults: A state of the science review. In *Neuro-Oncology* (Vol. 16, Issue 7, pp. 896–913). Oxford University Press. https://doi.org/10.1093/neuonc/nou087

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Xamla, A. K., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*.

Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M., Reina, G. A., Foley, P., Gruzdev, A., Martin, J., Albarqouni, S., Chen, Y., Shinohara, R. T., Reinke, A., Zimmerer, D., Freymann, J. B., Kirby, J. S., Davatzikos, C., Colen, R. R., Kotrotsou, A., … Bakas, S. (2021). *The Federated Tumor Segmentation (FeTS) Challenge*. http://arxiv.org/abs/2105.05874

Petzsche, M. R. H., de la Rosa, E., Hanning, U., Wiest, R., Pinilla, W. E. V., Reyes, M., Meyer, M. I., Liew, S.-L., Kofler, F., Ezhov, I., Robben, D., Hutton, A., Friedrich, T., Zarth, T., Bürkle, J., Baran, T. A., Menze, B., Broocks, G., Meyer, L., … Kirschke, J. S. (2022). *ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset*. http://arxiv.org/abs/2206.06694

Praveen, G. B., Agrawal, A., Sundaram, P., & Sardesai, S. (2018). Ischemic stroke lesion segmentation using stacked sparse autoencoder. *Computers in Biology and Medicine*, *99*, 38–52. https://doi.org/10.1016/j.compbiomed.2018.05.027

Quek, F., Williams, G., Bryll, R., & Gutierrez-Osuna, R. (2003). Attribute bagging:improving accuracy of classi#er ensembles by using random feature subsets Related papers At t ribut e bagging: improving accuracy of classifier ensembles by using random feat ure subs… Classifying Very-High-Dimensional Dat a wit h Random Forest s of Oblique Decision Trees Philippe Lenca Classifying Very High-Dimensional Dat a wit h Random Forest s Built from Small Subspaces Attribute bagging: improving accuracy of classier ensembles by using random feature subsets. In *Pattern Recognition* (Vol. 36). www.elsevier.com/locate/patcog

Sewell, M., & Tat, R. (2011). *Ensemble learning Related papers Ensemble Met hods Mart in Sewell T he Superiorit y of t he Ensemble Classificat ion Met hods: A Comprehensive Review Nzuva M Silas Classifier Combinat ion for In Vivo Magnet ic Resonance Spect ra of Brain Tumours Ensemble Learning Ensemble Learning*.

Shaga Devan, K., Kestler, H. A., Read, C., & Walther, P. (2022). Weighted average ensemble-based semantic segmentation in biological electron microscopy images. *Histochemistry and Cell Biology*. https://doi.org/10.1007/s00418-022-02148-3

Song, T. (n.d.). *3D Multi-scale U-Net with Atrous Convolution for Ischemic Stroke Lesion Segmentation*.

Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (n.d.). *Rethinking the Inception Architecture for Computer Vision*.

Tarvainen, A., & Valpola, H. (n.d.). *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*.

The Federated Tumor Segmentation (FeTS) Challenge 2022: Structured description of the challenge design mission. (n.d.). www.fets.ai

Tiwari, A., Srivastava, S., & Pant, M. (2020). Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019. *Pattern Recognition Letters*, *131*, 244–260. https://doi.org/10.1016/j.patrec.2019.11.020

Tomita, N., Jiang, S., Maeder, M. E., & Hassanpour, S. (2020). Automatic post-stroke lesion segmentation on MR images using 3D residual convolutional neural network. *NeuroImage: Clinical*, *27*. https://doi.org/10.1016/j.nicl.2020.102276

Wadhwa, A., Bhardwaj, A., & Singh Verma, V. (2019). A review on brain tumor segmentation of MRI images. In *Magnetic Resonance Imaging* (Vol. 61, pp. 247–259). Elsevier Inc. https://doi.org/10.1016/j.mri.2019.05.043

Wan, S., & Yang, H. (2013). Comparison among methods of ensemble learning. *Proceedings - 2013 International Symposium on Biometrics and Security Technologies, ISBAST 2013*, 286–290. https://doi.org/10.1109/ISBAST.2013.50

Wang, G., Li, W., Ourselin, S., & Vercauteren, T. (2017). *Automatic Brain Tumor Segmentation using Cascaded Anisotropic Convolutional Neural Networks*. https://doi.org/10.1007/978-3-319-75238-9_16

Wang, G., Li, W., Ourselin, S., & Vercauteren, T. (2019). Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation. *Frontiers in Computational Neuroscience*, *13*. https://doi.org/10.3389/fncom.2019.00056

Wang, S., Chen, Z., Yu, W., & Lei, B. (2020). *Brain Stroke Lesion Segmentation Using Consistent Perception Generative Adversarial Network*. http://arxiv.org/abs/2008.13109

Wang, W., Chen, C., Ding, M., Li, J., Yu, H., & Zha, S. (2021). *TransBTS: Multimodal Brain Tumor Segmentation Using Transformer*. http://arxiv.org/abs/2103.04430

Winzeck, S., Hakim, A., McKinley, R., Pinto, J. A. A. D. S. R., Alves, V., Silva, C., Pisov, M., Krivov, E., Belyaev, M., Monteiro, M., Oliveira, A., Choi, Y., Paik, M. C., Kwon, Y., Lee, H., Kim, B. J., Won, J. H., Islam, M., Ren, H., … Reyes, M. (2018). ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on

multispectral MRI. *Frontiers in Neurology*, *9*(SEP). https://doi.org/10.3389/fneur.2018.00679

Wu, Y., & He, K. (n.d.). *Group Normalization*.

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

Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., & Yu, Y. (2022). *Cross-Modality Deep Feature Learning for Brain Tumor Segmentation*. https://doi.org/10.1016/j.patcog.2020.107562

Zhang, R., Zhao, L., Lou, W., Abrigo, J. M., Mok, V. C. T., Chu, W. C. W., Wang, D., & Shi, L. (2018). Automatic Segmentation of Acute Ischemic Stroke From DWI Using 3-D Fully Convolutional DenseNets. *IEEE Transactions on Medical Imaging*, *37*(9), 2149–2160. https://doi.org/10.1109/TMI.2018.2821244

Zhang, Y., Liu, S., Li, C., & Wang, J. (2022). Application of Deep Learning Method on Ischemic Stroke Lesion Segmentation. In *Journal of Shanghai Jiaotong University (Science)* (Vol. 27, Issue 1, pp. 99–111). Shanghai Jiaotong University. https://doi.org/10.1007/s12204-021-2273-9

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., & Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Medical Image Analysis*, *43*, 98–111. https://doi.org/10.1016/j.media.2017.10.002

Zhou, C., Ding, C., Wang, X., Lu, Z., & Tao, D. (2019). *One-pass Multi-task Networks with Cross-task Guided Attention for Brain Tumor Segmentation*. https://doi.org/10.1109/TIP.2020.2973510

Zhou, Y., Huang, W., Dong, P., Xia, Y., & Wang, S. (2019). *D-UNet: a dimension-fusion U shape network for chronic stroke lesion segmentation*. https://doi.org/10.1109/TCBB.2019.2939522.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11045 LNCS*, 3–11. https://doi.org/10.1007/978-3-030-00889-5_1

Zhou, Z.-H., & Tang, W. (n.d.). *Selective Ensemble of Decision Trees*.

Ringraziamenti

A conclusione di questo elaborato, mi è doveroso dedicare qualche riga alle persone che hanno contribuito, con il loro instancabile supporto, alla realizzazione dello stesso.
In primis, un ringraziamento speciale va al mio relatore, il Prof. Atzori, per avermi guidato durante tutto il percorso di tesi, per la sua immensa pazienza, disponibilità, cortesia e per avermi trasmesso consigli e conoscenze indispensabili per la mia crescita universitaria e personale.
Ringrazio altrettanto caldamente il mio correlatore, il Dott. Tshimanga, che mi ha seguito e formato su tutti gli aspetti pratici della ricerca, supportandomi e aiutandomi a risolvere numerosi problemi con sconfinata pazienza e conoscenza.

Ai miei genitori, pietre miliari della mia vita, senza cui non sarei la persona che sono: Mamma Carmen, un esempio di forza, coraggio e amore, e Papà/Presidente Claudio, l'uomo più determinato e impavido che io conosca, e che spero non mi chiuda presto i rubinetti. Grazie per esserci sempre stati, per aver sempre supportato ogni mia decisione e per avermi donato ogni parte di voi stessi. Questa tesi la dedico a voi.

A mia sorella Sara, con cui ho condiviso ogni momento della mia vita, e che ringrazio per gli schiaffi in testa e per avermi riso dietro mentre mi chiudevo in casa, ma anche per il sostegno che mi ha sempre dato e per l'affetto che, anche se a modo suo, mi dimostra sempre.
Ringrazio anche Michele a cui, nonostante sia un arbitro, va tutta la mia stima per la sua immensa pazienza nel sopportare Sara.

Ringrazio i miei cugini, Lisa e Luca, con cui ho vissuto i ricordi più belli della mia infanzia. Un sentito ringraziamento anche a Nonna Luciana e Nonno Guido, agli Zii Chiara, Adriano, Mauro e Carmen e a Ivano e Luigi.
A Nonna Angela e Nonno Bepi, spero siate fieri di me da lassù.

Ai miei amici, Carlo, Davide, Daniele e Andrea, con i quali ho condiviso i momenti più belli degli ultimi anni, che ci sono sempre stati nel momento del bisogno e senza cui la mia vita sarebbe molto più grigia.
Perché con voi si sta bene anche distesi per terra.

Ai miei coinquilini, in primis Niccolò e Carlo, con cui ho scoperto l'interno 2 di Via Gattamelata 74, ma anche Andrea, Ciano e Alvise. Grazie per aver sopportato i miei riposini, i miei pasti improponibili e il "Vibe in sessione", siete stati la mia seconda casa.

Grazie a tutti i miei colleghi universitari, specialmente a Caterina, con cui ho condiviso l'intero percorso universitario e che è diventata la nostra quinta coinquilina.

A tutte le persone che ho perso durante il percorso ma che hanno comunque contribuito alla mia crescita personale.
Alle persone entrate recentemente nella mia vita, portando una ventata di solarità e felicità inaspettata.

A tutti voi grazie, questo successo è per voi.