POLITECNICO DI TORINO

Master of Science
in Data Science and Engineering

Master's Thesis

Deep Neural Network for Brain Tumor Segmentation in Magnetic Resonance Imaging

Advisor
prof. Lia Morra

signature

. . . . . .

Candidate
Luca Francesco Rossi

signature

. . . . . .

Co-Advisors
prof. Fabrizio Lamberti
prof. Luca Mainardi

signatures

. . . . . .

. . . . . .

Academic Year 2021-2022

To my family

Summary

Introduction

Deep Learning (DL) has achieved cutting edge results in several medical fields, with its applicability ranging from lesion segmentation to disease relapse prediction. Neuro-oncology, being one of those, has seen important advancements, especially in automating neuroradiology tasks such as brain tumor detection and segmentation. However, even if state of the art results have been achieved by DL methods on brain tumor segmentation on pre-operative Magnetic Resonance Imaging (MRI) scans, hardly the same can be said of post-operative segmentation, where literature lacks of a more comprehensive study and the few proposed models still present strong limitations regarding the scope or the generalizability.

Goals

Gliomas account for approximately 30% of all brain and central nervous system tumors, and 80% of all malignant brain tumors. Glioblastoma multiforme (GBM) are the most commonly encountered and aggressive malignant primary tumor (15%) of the central nervous system in adults, accounting for approximately 55% of all gliomas. Nowadays, the standard of care treatment for GBM patients consists in surgical resection followed by radiotherapy and chemotherapy, leaving then the subject untreated for the succeeding four months. The aim of this study, in conjunction with the Molinette Hospital in Turin, is to try making a further step in the field of automatic and semi-automatic brain tumor segmentation from neuroimaging modalities via deep learning technologies on post-operative scans.

Method

Due to the lack of available data in clinical practice, Transfer Learning (TL) has seen a spike in popularity within the medical field, allowing to train models in absence of a large dataset by leveraging knowledge learned from other source tasks. Still, current TL techniques in medical imaging mostly implement knowledge transfer from natural imaging, usually from model trained on the ImageNet dataset.
Even if some progress is done, the knowledge transferred between the two areas can be either not sufficient to achieve promising results in the medical task or make the transfer process quite unpredictable. This work arises upon the intuition that TL between pre-

and post-operative brain tumor segmentation could lead to promising results by leveraging both the closeness of source and target domains, and the fact that the knowledge transfer process does not leave the medical field. More in detail, a nnU-Net variant will be at first trained on the BraTS 2021 challenge dataset, a collection composed of 1251 patients – each one presenting four scan modalities (FLAIR, T1, T1ce and T2), acquired with different apparati and protocols from several different institutions – with the goal of becoming the de facto benchmark for addressing the automated tumor sub-region segmentation from pre-operative multi-parametric MRI scans. The network knowledge will then be transferred on the smaller post-operative MRI dataset, made available by the Molinette Hospital, which includes 166 scans from 71 patients with imaging times ranging from immediately after the surgery (max 48h) to 12+ months after.
Following literature, the quality evaluation of the automatic brain tumor segmentation performed by the network – both pre- and post-operative – is performed through Dice score and the undirected 95th percentile Hausdorff distance. Since the private dataset includes only FLAIR and T1ce as modality, an Image Modality Transfer (IMT) method is also investigated in order to artificially synthesize the missing scans, hence trying to leverage further contextual information. In detail, a 2D U-Net architecture is trained on the BraTS 2021 dataset in order to learn the non-linear intensity mapping between the T1ce modality and the missing ones. FLAIR scans are not included as input due to the presence of non-volumetric imaging, i.e. low-res acquisitions which would eventually lead to anatomically coarse-grained outputs.

Results

Pre-operative segmentation comprises the three classes of Whole Tumor (WT), Tumor Core (TC) and Enhancing Tumor (ET), achieving an average Dice score of 91.09 (WT: 93.78, TC: 93.11, ET: 86.37) and an average Hausdorff95 distance of 8.22 (WT: 5.45, TC: 5.76, ET: 13.45), in line with the top performances indicated by the authors. On the other hand, post-operative segmentation comprises the three classes of Gross Tumor Volume (GTV) plus surrounding Edema (ED), GTV only and Resection Cavity (RC), achieving an average Dice score of 72.44 (GTV+ED: 76.51, GTV: 76.50, RC: 64.32) and an average Hausdorff distance of 23.43 (GTV+ED: 10.37, GTV: 10.30, RC: 49.63). No significant improvement in performances was observed by including the two artificially synthesized missing MRI modalities with the proposed IMT method.
Differently from other studies in literature, this work present a first omni-comprehensive tool for post-operative brain tumor segmentation, i.e. it does not focus on a single class recognition (such as the resection cavity) and it does not impose any kind of exclusion criteria for the dataset. Indeed, in previous studies, strong criteria for inclusion/exclusion of MRI scans in the final dataset are often imposed, such as, among others, newly-diagnosed GBM only, availability of all imaging modalities or resection cavity clearly present on visual inspection. These constraints undoubtedly ease the learning process but are not generalizable to the clinical practice where an incommensurable diverseness and variety is present. This translates however in dealing with scans where, for example, some classes are not present at all (brain parenchyma expanding and filling the cavity), or the same

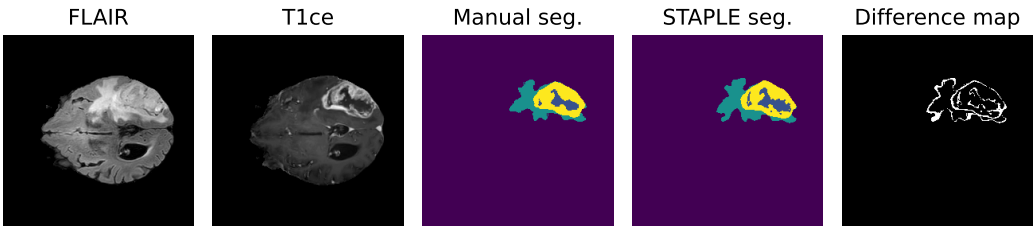


Figure 1. Example of pre-operative segmentation for a patient from the Molinette Hospital dataset (in blue: Necrosis, in yellow: ET, in turquoise: ED).

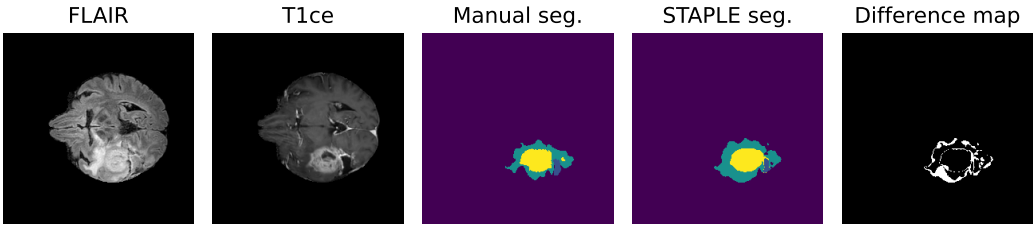


Figure 2. Example of post-operative segmentation for a patient from the Molinette Hospital dataset (in blue: CV, in yellow: ET, in turquoise: ED).

class behaves in a completely dissimilar way among different patients (hemoglobin degradation). This emerges vividly in the RC segmentation, undoubtedly the Achilles' heel of this work, with the network either oversegmenting the cavity by mistakenly including other hypointense regions or by completely misrecognizing it when a too heterogeneous pattern is present. The former issue is nontheless eventually mitigated by the STAPLE fusion and by leveraging the three overlapping prediction confidences, thus improving the model performances.
Under this light, results are therefore in line with the literature, considering its strongly heterogeneous nature and the reduced dimension of the dataset. While looking forward to a publicly available multi-site dataset like BraTS, this work presents itself as an interesting proposal and starting point in order to reduce the gap between pre- and post-operative GBM segmentation.

Acknowledgements

To my parents, for always caring, supporting and encouraging me. After a thesis on quantum mechanics and one on brain tumor segmentation, finding a way to express all my gratitude is still the hardest thing to do. So just thank you. Really.

To Federico, for constantly reminding me to enjoy life and to have priorities. I promise I'll be better.

To the rest of the family, for always being there from day one.

To my advisors, for their valuable and always on point guidance. A special thanks, in particular way, goes to Prof. Morra for her support and directions.

To the Molinette neurosurgeons and neuroradiologists, for their priceless contribution. A special thanks, in particular way, goes to Dr. Bianconi, for his availability and patience.

To the Department of Control and Computer Engineering, for providing the computational resources of the HPC@POLITO academic project.

To the San Giuseppe family, for having made me feel at home every single day.

To the Campulo family, for showing how geographical distance is nothing if you meet the right person.

To the Farminator family, for making me constantly feel homesick.

To the SoCo family, for teaching me once again how the people you work with are more important than the work itself.

To Conan, Matt and Sona, for being my pocket friends and brightening my days.

Contents

List of Tables

List of Figures

"Don't Panic."
[D. Adams, The Hitchhiker's Guide to the Galaxy]

Chapter 1

Introduction

Artificial intelligence will not replace radiologists... but radiologists who use AI will replace radiologists who don't.

- Dr. Curtis Langlotz

1.1 AI in medical imaging: a new paradigm

Medical imaging refers to those imaging techniques which try to build visual representations of either internal or external tissues of the human body by means of physical phenomena such as electromagnetic radiation, radioactivity, light, nuclear magnetic resonance or sound via both non-invasive and invasive procedures (Zhou et al. [2020a]). The most widely adopted imaging modalities in the clinical environment include computed tomography (CT), X-ray radiography, ultrasound, digital pathology and magnetic resonance imaging (MRI). There are several traits influencing the nature and sustainability of computer vision solutions in medical imaging. First, medical images not only present themselves in all those different medical modalities but are getting denser and denser in pixel resolution (as an example, MRI and CT spatial resolution has reached a sub-millimeter level) so that several complications arise then from the trade-off between pixel/voxel resolution and information density. Second, medical images are both isolated and acquired in non-standard settings. But the lack of standardization and the high heterogeneity in equipment and scanning settings are only part of the problem since, due to patient privacy and clinical data management requirements, datasets are scattered among different locations. Third, disease patterns are numerous and the labels associated with medical images are often sparse and noisy. Indeed, annotating is expensive and time-consuming, and different tasks require different forms of labeling. Because of this unavoidable inner variability inter-user and intra-user labeling consistency is low, with the establishment of standards for medical imaging annotating still being a long way down the road. Last, samples are both unbalanced and heterogeneous, and analysis tasks are often complex and diverse. In the case of tumors, for example, the number of pixels/voxels belonging to such class is usually many orders of magnitudes smaller than that of healthy tissue. Medical imaging is nonetheless a core aspect of the medical diagnosis and treatment process since,

typically, both the diagnosis and the treatment are planned by the referring physicians based on the radiologist's report. Furthermore, medical imaging is often recommended as part of the follow-up in order to check the successful status of the treatment. However, human image interpretation is limited, as anticipated, by the wide heterogeneity across interpreters. The process is often wearing and time consuming and, given the limited time that radiologists have when reviewing the cases, actual missed findings can heavily stretch the path to a more evidence-based personalized healthcare. In this framework, AI and deep learning (DL) technologies can provide support by automating and standardizing a wide spectrum of tasks, from quantification of disease extent to characterization of pathologies. Refined versions of state-of-the-art methods have been successfully adopted for tasks such as object classification, localization, detection or segmentation. Indeed, if sufficient data were present, the accuracy of such models often matched (if not even surpassed) the one of expert physicians (Esteva et al. [2021]).
In order to try to summarize this heterogeneous environment, it is possible to let all medical imaging-related deep learning tasks fall under one of the following applications:

- image reconstruction, i.e. building an image from those signals returned by medical devices such as a CT of MRI scanner;
- image enhancement, i.e. adjusting intensity in an image to make the resulting image more suitable and prone to further analysis (de-noising, bias field correction, harmonization, etc.);
- image segmentation, i.e. assigning labels to pixels/voxels in order to form segmented objects as components;
- image registration, i.e. aligning spatial coordinates of an input image in order to match a common coordinate system;
- computer aided detection (CADe), i.e. localizing and bounding an object of interest;
- computer aided diagnosis (CADx), i.e. deepening the output of CADe by further classifying the localized object of interest as either one of multiple types or as benign/malignant;
- other technologies, such as automatic report generation or landmark detection.

Any of the above can eventually be regarded as a function approximation method, with a mapping $\mathcal{F}$ that takes as input an image x and returns as output a task-specific $y = \mathcal{F}(x)$. Because of its focus on learning rather than modeling, deep learning has been widely adopted in medical imaging since its conception and represents a substantial departure from previous approaches. However, it is undeniable that this new approach suffers from lack of interpretability: the lack of evidence and interpretation for such algorithms (often used as black-box) makes it difficult for radiologists to trust the DL model's prediction, especially since it is also true that interpretability is often the source of new knowledge.

Thoracic imaging

Lung diseases are characterized by an extremely high morbidity and mortality. Indeed, among the top ten causes of death worldwide it is possible to observe lung cancer, pneumonia, tubercolosis and chronic obstructive pulmonary diseases. Moreover, pulmonary complications are common side-effects during hospitalization. The hyper-presence of these malignancies leads however also to an hyper-presence of chest radiography screenings, by far the most common radiological examination. When imaging the chest, plain radiography and CT are the two most common modalities. Automated segmentation methods from chest CT scans have been proposed and state-of-the-art results have been achieved by deep learning during the last few years (Gerard and Reinhardt [2019]). Another important area of applicability and research for AI in thoracic imaging is the segmentation of the vasculature, separating into arteries and veins, and the airway tree. Moreover, the availability of large public datasets during the last years (ChestXRay, CheXpert, MIMIC, PadChest) has driven the research towards DL-based methods for abnormalities detection in chest X-ray scans. As an example, given the overwhelming presence of patients in hospitals during the COVID-19 pandemic situation and the subsequent shortage of molecular testing, AI has been adopted to analyse chest X-ray or CT scans in order to obtain a working diagnosis.

Cardiovascular imaging

There are three main sub-fields in which deep learning impacted the field of cardiovascular imaging: cardiac chamber segmentation, cardiac motion/deformation analysis and analysis of cardiac vessels (Zhou et al. [2020a]). Cardiac image segmentation is often the first step for many clinical applications, whose main goal is usually the segmentation of the main chambers, bounding the two ventricles as well as the two atriums. Cardiac motion tracking is of core importance for analyzing the actual mechanic performance of heart chambers and combinations of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been adopted in order to estimate both temporal and spatial variations. Cardiac vessel segmentation consists in the segmentation of the vessels including both the great vessels (i.e. aorta, pulmonary arteries and veins) and the coronary arteries.

Abdominal imaging

The liver, spine and prostate are without any doubt the most accurately segmented and most actively investigated organs by deep learning techniques. Given the promising results, recent research started investigating other organs of interest such as the pancreas, lymph nodes and the bowel. Moreover, potentially diffuse disorders in the kidneys were detectable by using DL strategies (Kuo et al. [2019]). Opportunistic screening for survival prediction and cardiovascular morbidity (heart attack, stroke) has been an area of recent interest. Furthermore, deep learning methods have been applied in domains such as visceral fat assessment, muscle volume assessment, aortic atherosclerotic plaque quantification and bone mineral densitometry.

Microscopy imaging

With the development of large datasets containing digital tissue slide images, a noticeable increase in DL-based methods applications to digital pathology data has been observed. An early application to whole slide pathology images was related to the detection and segmentation of individual primitives such as lymphocytes and cancer nuclei. Several studies showed indeed the utility of DL approaches in the identification of several histologic primitives, including tubules, cancer extent and in the classification of different disease categories related to leukemia (Janowczyk and Madabhushi [2016]). Another promising field of applicability of AI in microscopy imaging is the one of disease grading. Specifically, DL has been adopted to mimic pathologist's identification of disease hallmarks – with particular focus in cancerous ones – achieving results often comparable with the ones of expert physicians. Some recent research focused also on the identification of specific mutations and their association with biological pathways. As an example, it was shown that deep learning networks are actually able to recognize several of the mutated genes in non-small cell lung adenocarcinoma (Coudray et al. [2017]). Some attempts to apply deep learning in survival and disease outcome prediction have also been made but, while the studies presented above show the rising potential of DL on image analysis and classification tasks in digital pathology, there is still some concern regarding its interpretability so that it remain to be seen how these approaches will ultimately, eventually, translate to the clinic (Zhou et al. [2020a]).

1.1.1 Oncology

Cancer diagnostics is the oncology-related field which artificial intelligence impacted the most (Luchini et al. [2022]), both widening known horizons and opening new opportunities. This is a huge achievement since cancer diagnostics represents the fundamental starting point for designing both the appropriate therapeutic approach and the clinical management. In precision oncology, this new paradigm allows to reach ground-breaking results by integrating huge amount of data derived from heterogeneous analyses with high-performing deep-learning techniques (Bhinder et al. [2021]). Furthermore, by 2030, it is estimated that 13% of the world population will fall in the elderly category, with a huge burden on the health care sector due to the higher risk in incurring into serious diseases such as brain tumors, chest cancer, lung cancer, diabetes, hypertension or heart failure (Parekh et al. [2011]), so AI can be of great help in lightening the situation. Figure 1.1 shows the study conducted by the Center for Medicare and Medicaid Services (CMS) which revealed that US expenditures in health care increased nonstop year by year. Figure 1.2 shows instead the results of the systematic review performed by the authors investigating AI-based devices that have already obtained the Federal Drug Administration (FDA) approval for entering the clinical practice in oncology (last access: 05/31/2021). Classifying instead by oncology-related specialties, the authors highlighted that the field counting the largest number of approved AI-based devices is cancer radiology (54.9%), followed by pathology (19.7%), radiation oncology (8.5%), gastroenterology (8.5%), clinical oncology (7.0%) and gynaecology (1.4%).

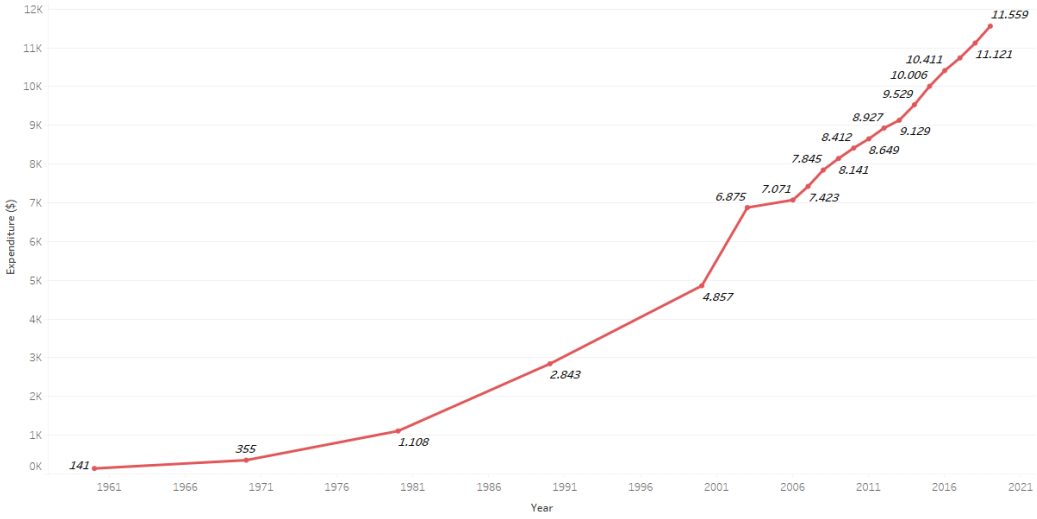


Figure 1.1. CMS data for national healthcare expenditure per capita in the US.

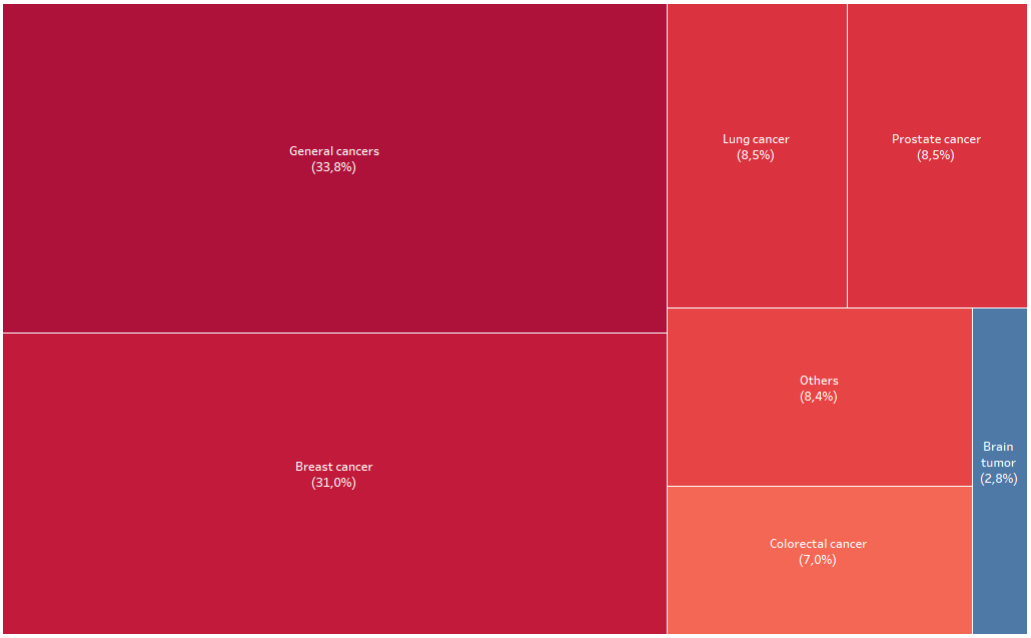


Figure 1.2. AI-based devices, FDA-approved, expressed by tumor type (Luchini et al. [2022]).

There are three main clinical radiology tasks that specifically concern oncology: abnormality detection, characterization and monitoring of change. Both medical and technical skills are required when dealing with such tasks, since knowledge of both disease diagnosis and radiographic image processing are necessary. Human manual detection relies on the radiologists' perceptive skills to identify abnormalities and on their cognitive skills

to either validate or not the findings. Automated CADe methods for processing and identification have been proposed and utilized in the clinic (Castellino [2005]). However, such algorithms are often task-specific and do not generalize well. Moreover, some of this methods have still sub-human performances and little to no help is brought if the radiologist needs to review the output from scratches. Characterization is a wide sphere including segmentation, diagnosis and staging of a disease. This is likely the sub-field in which artificial intelligence shines the most since, contrary to humans, computers are able to account for more than a handful of qualitative features. Disease monitoring is essential as it is a fundamental step in order to evaluate the goodness of treatment response. The workflow consists in co-registering the diseased tissue across different scans and comparing its temporal evolution (as an example, tumor size in oncology).

1.1.2 Neuro-oncology

Tumors of the central nervous system (CNS) account for a small proportion of all invasive cancers (around 1%). Still, their heterogeneity is astonishing considering that both malignant (29.7%) and non-malignant (70.3%) brain tumors present over a hundred different sub-types, each one of those with their own clinical and biological behaviors (Miller et al. [2021]). Even if they are rare tumors – with an annual incidence of around 0.8% – these malignancies bring with them a very high mortality rate, permanent disability and high public health costs (Di Nunno et al. [2022]). Hence, researchers and clinicians in the field of neuro-oncology have to face important challenges: among those, the prediction of disease progression, an improved and non-invasive tumor diagnosis and the research of new effective treatments. During the last years, deep learning has seen a noticeable rise in popularity within the neuroimaging community. Indeed, it offers fascinating prospects for improving both generalizability and accuracy while at the same time reducing the need for complex processing and inference time. Convolutional neural networks (CNNs) have especially proven themselves effective in analysis and high-level prediction tasks, reducing the need for specific and high-level domain-related prior knowledge.

Neuroradiology

Integrating artificial intelligence (AI) into the clinical workflow as an aid to assist physicians can lead to more accurate and reproducible radiology assessments (Hosny et al. [2018]). Indeed, an integrated AI component within the medical imaging workflow would conduct towards better efficacy, better efficiency, reducing errors and achieving objectives with minimal human intervention. Given the high amount of data provided by examinations and its non-invasive nature, neuroradiology is considered to be a thriving field for AI and DL applicability. Accurate segmentation is indeed a fundamental starting point for any analytic or predictive tasks done in neuroimaging. The required tasks are various and range from brain tissue segmentation and volume estimation of white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF), to brain lesion detection and segmentation, from cortical and subcortical segmentation, to brain extraction and deformable image registration (Chen et al. [2017a]), i.e. co-registering different images to a given reference system in order to analyze them across different modalities and time points. The advent of

generative adversarial networks (GANs) and their variants has been of great importance in neuroimaging since complex synthesis tasks – not even achievable by machine learning methods – became feasible, and have been used ever since for resolution upscaling, cross-modality synthesis, motion artifact reduction, and more. Still, even if DL in neuroimaging has opened many interesting perspectives worth of investigation, certain fields still lack a rigorous understanding, hence the upcoming years will be crucial for further investigation.

Molecular-pathological diagnosis

Even if deep learning is able to provide valuable insights to radiologists from MRI data, a pathological assessment is still necessary before planning either surgical or post-surgical treatment. Currently, the histomorphological diagnostic process is entirely based on the pathologist's expertise in the required field. This morphological and histological evaluation is needed because it is necessary ((Di Nunno et al. [2022])):

- to verify that the sample is representative of the disease;
- to confirm the histotype and evaluate the morphological heterogeneity of the tumor;
- for the triage of the available material for molecular analysis;
- to select the most appropriate area of investigation;
- to guide the choice of the correct method.

In neuro-oncology, for example, AI algorithms have been adopted with the goal of classifying glioma subtypes, which could be essential for the correct prognosis estimation since the complexity of these histopathological analyses of brain tumor have greatly increased with the increasing number of immunohistochemical studies that are needed to be interpreted in parallel.

Surgery, radiation therapy and systemic treatment

Artificial intelligence could improve both the performance and the outcome of brain tumor resection. Indeed, an adequate assessment of extension and burden is fundamental during the pre-surgical planning stages. More generally, AI has been often applied with the aim of automatizing and optimizing complex tasks while also increasing quality, standardization and accuracy. Radiation therapy processes are often time-consuming and associated with large user variability, so that accurate segmentation and optimization of radiotherapy treatment planning are crucial steps that could be supported by AI and DL-based technologies. However, despite different studies dealing with AI in radiation oncology, none of such technique has still been approved for clinical practice. Few are instead the studies investigating the potential role of AI in systemic treatment for CNS malignancies. Both the identification of promising drugs for pre-clinical testing and of the optimal dosage and protocol of systemic treatment are tasks that are possible applications and some steps in this direction have been made. Finally, another interesting and promising use of AI is the one of predictive response factors identification, as an aid to evaluate the goodness of the treatment status.

1.2 Goals

Gliomas account for approximately 30% of all brain and central nervous system tumors, and 80% of all malignant brain tumors (Goodenberger and Jenkins [2012]). A brain tumor consists in an uncontrolled growth and multiplication of cells giving birth to an abnormal mass of tissue. They are further classified by the World Health Organization (WHO) as astrocytoma, oligodendroglioma, mixed oliogastrocyntoma, and ependymoma (Louis et al. [2021]). Low-grade astrocytomas (WHO Grade II), differently from pilocytic astrocytomas (WHO Grade I), often degenerate to higher grade tumors. Conventionally, both the histological subtype and the grade are connected to the malignant potential and the survival. Patient with Grade II astrocytomas have a survival rate at 5-year of arount 50% (Wu et al. [2010]).
Apart from the WHO histologic classification, they can be more generally grouped into primary and secondary (metastatic) tumors, based on their origin place. More specifically, primary tumors arise in the brain itself and can be either benign or malignant. Beningn tumors, differently from malignant ones, grow slowly and do not spread to surrounding tissues. However, their growth can still put pressure and compromise the brain function. On the other side, secondary brain tumors originate from another part of the body, usually due to cancer cells spreading to the brain (Biratu et al. [2021]).

Glioblastoma multiforme

Glioblastoma multiforme (GBM) are the most commonly encountered and aggressive malignant primary tumor (15%) of the central nervous system in adults, accounting for approximately 55% of all gliomas. As WHO Grade IV astrocytoma, GBM present themselves with a vast intrinsic heterogeneity in both appearance and shape. Nowadays, the standard of care treatment for GBM patients consists in surgical resection followed by radiotherapy and chemotherapy, leaving then the subject untreated for the succeeding four months (Baid et al. [2021]). Unfortunately, even if several experimental treatments have been proposed and discussed in the last twenty years, no substantial improvement – or difference – in patient prognosis has taken place. For brain tumor diagnosis, the gold standard consisted in biopsy, which includes resection and pathological examination via histologic analyses. However, such procedure is invasive and may lead to bleeding or injury, with possible complications as problematic as functional loss. As a results, modern neuroimaging moved towards non-invasive brain tumor diagnosis using Magnetic Resonance Imaging (MRI) in order to characterize structural, cellular and functional properties of the tumor (Roberts et al. [2020]). Normal brain tissues in MRI are characterized by cerebrospinal fluid (CSF), white matter (WM) and gray matter (GM). Tumorous brain scan present instead necrosis, tumor core and edema. The necrosis is a mass of dead cells located inside the core tumor, whereas the edema is a swelling situated near active tumor borders, present due to trapped fluids around a tumor. In gliomas, the edema is said to be infiltrative since it invades WM tracts of a brain.
The detailed segmentation and identification of glioma sub-regions boundaries in MRI data has therefore high importance in many clinical situations, from monitoring to treatment planning. Yet, the manual delineation of tumor sub-regions is time-consuming and

highly dependent on the radiologist(s) implementing it, hence becoming impractical when dealing with numerous subjects. The aim of this study, in conjunction with the Molinette Hospital in Turin, is to make a further step in the field of automatic and semi-automatic brain tumor segmentation from neuroimaging modalities via deep learning technologies on post-operative scans.

Chapter 2

Theoretical background

> *nnU-Net is much more than the architecture. It is a holistic framework for automatically generating/configuring segmentation pipelines. The spirit of nnU-Net is cross-dataset compatibility and out-of-the-box performance without the need to fiddle with hyperparameters.*
>
> - Fabian Isensee

2.1 U-Net: the state of the art

Deep convolutional neural networks have outperformed what was the state of the art in several visual recognition tasks. However, their success was strongly linked to the number of available training samples and the size of the network itself and, sadly, thousands of image samples are often beyond reach in the medical field. Moreover, the quintessential use of convolutional neural networks is classification, where an image is taken as input and a single class label is given back as output. Unfortunately, when dealing with biomedical image processing, this is not often the case, where instead the desired output should include localization, meaning that each image pixel should be classified with a class label. A first solution to these problems was proposed by Ciresan et al. [2012], where the network is trained with a sliding-window to predict the class of each pixel by providing as input a fixed patch around that given pixel. Even though astonishing results were obtained, the proposed model had two drawbacks:

1. slowness, due to the fact that the network has to be run separately for each patch,
2. a tradeoff between contextual information and localization accuracy, i.e. larger patches allow an higher information gain but reduce localization accuracy, and viceversa.

Ronneberger et al. [2015] built a new solution upon a more refined architecture: the fully convolutional network, whose main idea behind is to append to the usual contracting network a serie of layers where pooling is substituted by upsampling, hence increasing the resolution of the output. Localization turns out then to be facilitated by combining features extracted by the contracting path with the output of the following upsampling.

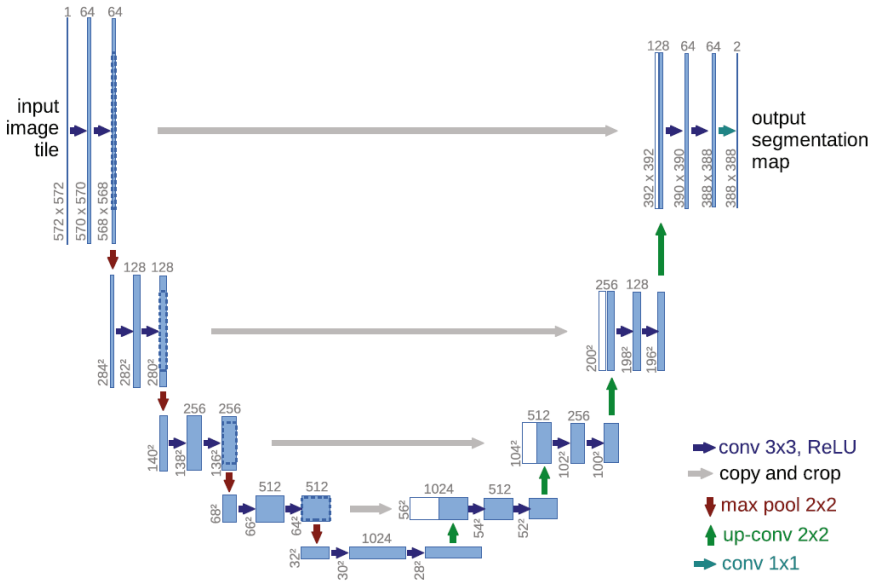


Figure 2.1. Base U-Net as depicted in the original paper by Ronneberger et al. [2015].

Figure 2.1 shows the architecture proposed (here an example for 32×32 pixels in the lowest resolution) – called U-Net because of the "U"-shape derived from the nearly-symmetric encoder-decoder structure – where it is observable the behaviour presented above. The two informations combined are then assembled through a convolution layer in order to extract a more precise and informative output.

2.1.1 U-Net variants for medical image segmentation

The fact that U-Net is able to provide solid results in image segmentation using a scarce amount of data is the main reason behind its extensive adoption within the medical imaging community. Starting from the skip connections structure between the downsampling and upsampling paths which allows for contextual information to be preserved, several advancements and adjustments in U-Net architecture have been implemented, giving birth to several variants (Siddique et al. [2021]).

3D U-Net

The first advancement consisted was the enabling of 3D volumetric segmentation (Çiçek et al. [2016]). The structure of the architecture was the same as the 2D base version, however, all of 2D operations were replaced by 3D ones, namely 3D pooling and 3D convolution, thus allowing for three dimensional segmentation.

Attention U-Net

A worthwhile trait in neural networks for image processing is the possibility of focusing on specific regions that are considered important while ignoring areas regarded as not worth of notice. Attention U-Net is an important step towards this ability, achieved by the adoption of the attention gate (Oktay et al. [2018]).

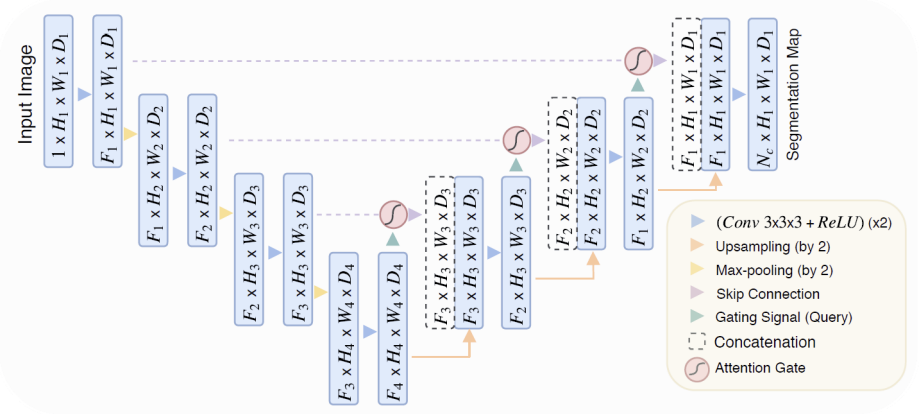


Figure 2.2. Attention U-Net as depicted in the original paper by Oktay et al. [2018].

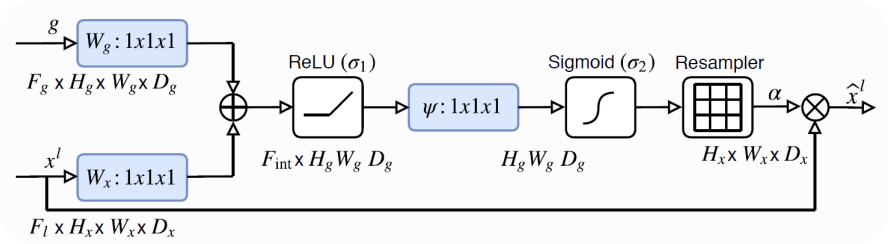


Figure 2.3. Schematic representation of the adopted attention gate (Oktay et al. [2018]).

Encoder-decoder networks such as U-Net benefit from the attention unit since it provides more accurate and informative localized information. It can be shown that, in the specific case of U-Net, this allows for different parts of the architecture to focus on segmenting distinct different items.

Inception U-Net

Neural networks for image processing tend to be set with fixed-size filters for convolutions. Fine tuning the algorithm to find the ad hoc filter size can be burdensome, if not computationally unfeasible. In addition, fixed-size filters are a good choice only for processing images whose salient parts do not differ in size. In medical analysis, sadly, this is not the case, with large variations in both size and shape in the regions of interest.
A solution to this problem, called the inception network, uses several filters with different size on the same layer in the network, whose outputs are then concatenated and forwarded to the next layer. The architecture is therefore able in this way to analyze images with different regions of interest in quite an effective manner.

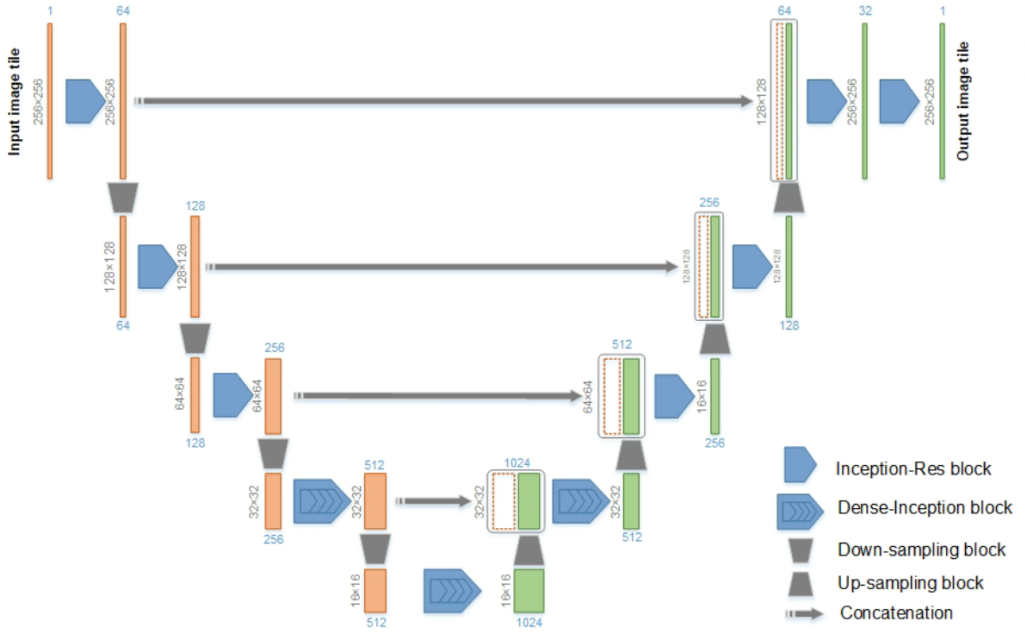


Figure 2.4. Dense-Inception U-Net as depicted in the original paper by Zhang et al. [2020b].

Inception components of different configurations, after the astounding results obtained in the ILSVRC14 competition by GoogLeNet, have been adopted – applied on architectures derived from U-Net – for medical imaging tasks.
Several medical fields, such as brain tissue mapping, cardiac segmentation, human embryo segmentation, lung nodule detection and ultrasound nerve segmentation have been benefiting from the adoption of U-Net-like inception modules. Figure 2.4, as an example, shows an architecture presented for brain tumor detection (Zhang et al. [2020b]).

Residual U-Net

It is a well known fact that the more layers are present in a deep neural network, the faster will be its convergence. That said, experimental results highlight that this can eventually lead also to saturation, possibly causing further degradation in performance (He et al. [2015]). This variant of U-Net is based on the ResNet architecture. The advantage of ResNet is the adoption of skip connections which forward the feature map from one layer to another one located more deeply in the architecture. This enable the network to preserve information and avoid unwanted behaviors such as the one presented above.
In the residual U-Net, the input to the first convolutional layers is added to the output of the second convolutional layer using, at each block, a skip connection. Each residual unit can be described by the following equation:

$$x_{l+1} = f\big(x_l + \mathcal{F}(x_l, \mathcal{W}_l)\big), \tag{2.1}$$

where x_l and x_{l+1} represent the input and the output of the residual unit respectively, $\mathcal{F}(\cdot)$ is the residual fuction and $f(\cdot)$ is the activation function.

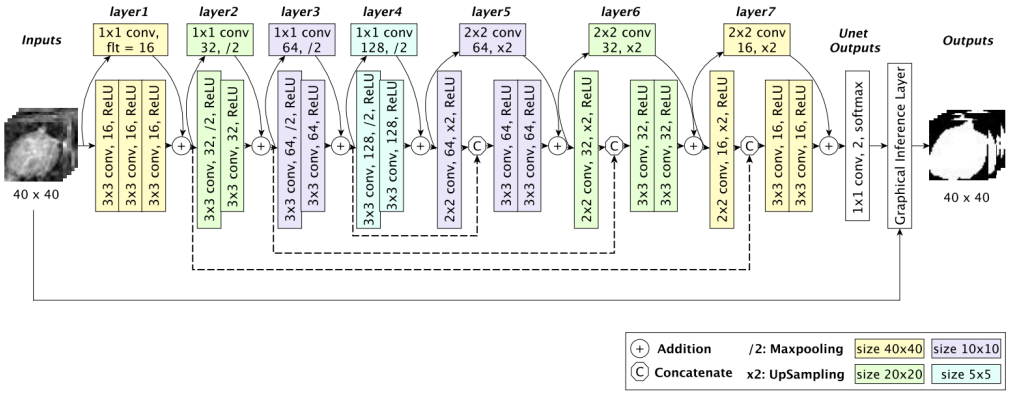


Figure 2.5. CRU-Net as depicted in the original paper by Li et al. [2018].

Due to the advantage derived from residual connections alleviating the vanishing gradient problem, Residual U-Net are the ideal tool for complex image analysis. Several examples can be found in literature, with applications in several different fields ranging from brain structure mapping to retinal vessel segmentation, from prostate cancer to bone structure analysis. As an example, Figure 2.5 shows the deep Conditional Residual U-Net (CRU-Net) architecture proposed by Li et al. [2018] for breast cancer, where skip residual connections are added in each of the architecture's seven layers.

Dense U-Net

Even though Residual U-Net alleviates the problem of vanishing gradients, it does not eliminate it. While it allows for deeper neural networks, its performance will eventually degrade if the number of layers keeps increasing.
DenseNet (Huang et al. [2017]) is a deep learning architecture proposed on top of ResNet with two main changes. First, each layer receives now the feature maps from all of its preceding layers. Secondly, the identity maps that forward the feature maps from one layer to the others are combined via channel-wise concatenation, so that results are dependent on all previous layers and gradient propagation is significantly promoted.
The output for each layer in a dense block can therefore be described by the following equation:

$$x_{l+1} = \mathcal{H}_l([x_0, x_1, x_2, \ldots, x_{l-1}, x_l]), \tag{2.2}$$

where $\mathcal{H}_l(\cdot)$ represents the dense mapping function at layer l (usually including batch normalization, ReLU and a convolution), while $[\cdot]$ denotes the channel-wise concatenation. In the specific case of U-Net, each standard U-Net block is substituted by a dense block of two (or more) convolutional layers.

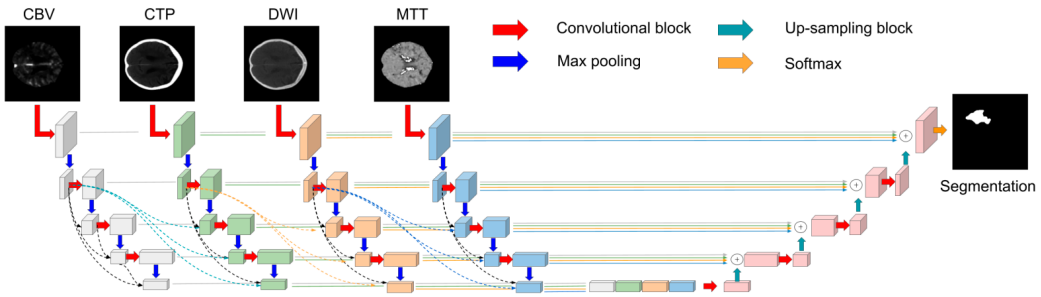


Figure 2.6. Multi-path Dense U-Net as depicted in the original paper by Dolz et al. [2018].

Since dense blocks eliminate the problem of vanishing gradients, Dense U-Net can be modeled even deeper, hence enabling the network to segment objects with greater distinction. For this reason, applications of Dense U-Net can be observed in several medical imaging fields, including retinal blood vessel segmentation, cerebral blood vessel segmentation, melanoma, lung cancer, liver cancer, multi-organ segmentation and brain tumors. This feature of Dense U-Net is considered extremely valuable in medical image analysis, since objects are often extremely close and overlapping.
Figure 2.6 presents a multi-path Dense U-Net architecture proposed by Dolz et al. [2018] for ischemic stroke lesion segmentation in multiple image modalities.

U-Net++

U-Net++ is a further U-Net-based architecture, proposed as a variant on top of the Dense U-Net structure. Differently from its ancestor, it implements also a dense network of residual skip connection between the contracting and the expansive paths, enabling the network to propagate more semantic information between the two paths leading to more accurate image segmentation (Zhou et al. [2018]). The difference is therefore that each level is now equivalent to a dense block, so feature maps of the contracting path are no more directly concatenated only onto the layer situated at the same level in the expansive path. If i and j are the index down the contracting path and across the residual skip connection respectively, and x represents the feature map, it is possible to describe the operation of such skip connection unit as follow:

$$x^{i,j} = \begin{cases} \mathcal{H}(x^{i-1,j}), & \text{if } j = 0 \\ \mathcal{H}([[x^{i,k}]_{k=0}^{j-1}, \mathcal{U}(x^{i+1,j-1})]) & \text{if } j > 0 \end{cases}, \tag{2.3}$$

where $\mathcal{H}(\cdot)$ represents the dense mapping function, $\mathcal{U}(\cdot)$ the upsampling operation and $[\cdot]$ denotes channel-wise concatentation. It is worth noticing that the number of intermediary connection units decreases linearly going down the contracting path, thus depending on the layer number.

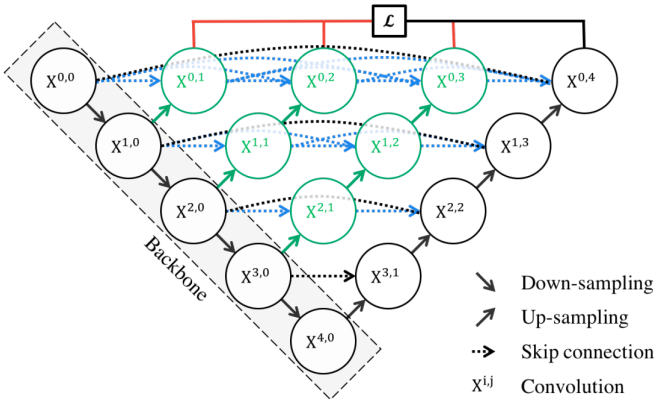


Figure 2.7. U-Net++ as depicted in the original paper by Zhou et al. [2018].

Because of its features and improved performance, applications of U-Net++ include extremely specific objects of study such as cancer tissue, cardiac structures, vessels, pelvic organs or cell nuclei. In Figure 2.7 it is possible to observed the U-Net++ architecture as originally proposed by Zhou et al. [2018], where black indicates the original U-Net backbone, while green and blue show dense convolution blocks on the skip pathways.

Adversarial U-Net

The adversarial structure consists in two network – respectively called generator and discriminator – "competing" against each other in order to symbiotically improve their performance. The goal of the discriminator is to identify whether a given input image is either derived from the data set or is crafted by the generator. Images produced by the generator are periodically forwarded to the discriminator, and its gradient function is computed as a function of the one of the discriminator in order to adjust the weights based on the output of the last. It can be shown that, if enough time is provided, adversarial networks are able to reach the optimal state in which the discriminator always outputs a probability of 1/2 independently of whether its input image is derived from the data set or produced by the generator, meaning that it is no longer able to distinguish real images from synthetic ones (Goodfellow et al. [2014]).
The Adversarial U-Net is a type of generative adversarial network (GAN) called conditional GAN, whose main advantage is the production of limited band of synthetic images by controlling its labels and input images (Mirza and Osindero [2014]). The purpose of the generator is therefore to synthesize transformed images rather than new ones.

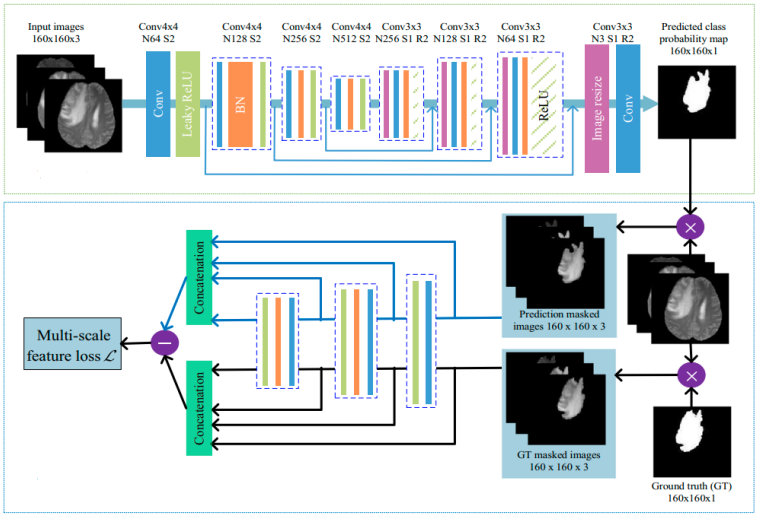


Figure 2.8. SegAN as depicted in the original paper by Xue et al. [2018].

Given the theoretical result of GAN convergence, Adversarial U-Net spectrum of applications is extremely wide, ranging from the segmentation of cardiac structures to breast cancer. As an example, Figure 2.8 shows SegAN, an architecture based on the Adversarial U-Net proposed by Xue et al. [2018] for brain tumors.

2.1.2 nnU-Net: one net to rule them all

Even though several U-Net variants have been proposed in medical imaging literature, it is undeniable that the design of specialized solutions is highly dependent on both hardware conditions and dataset properties. Indeed, despite the promising results obtained by deep learning-based methods, their translation (generalization) to task-specific problems is often problematic and eventually leading to large drops in performance (Litjens et al. [2017]). Especially in 3D imaging, where successful configurations from one dataset rarely translates to another, high levels of expertise and experience are often required. The numerous decisions needed for either designing a well-performing algorithm or adapting one for other tasks range from the architecture definition to data augmentation, from training schedule to post-processing. Moreover, each one of these features is further controlled by parameters such as batch size or learning rate, and a supplementary layer of complexity is posed by the available hardware. AutoML (Hutter et al. [2019]) consists in a purely empirical optimization of all those design choices in a high-dimensional space which determines – based on the dataset – the top-performing configuration. This comes however at the expenses of needing a higher (by orders of magnitude) number of both training cases and computer resources, often becoming practically unfeasible in medical imaging. In addition, the high-dimensional search space has to be anyway defined a priori, thus still needing a lot of field-specific knowledge and expertise.
nnU-Net (Isensee et al. [2021a]) arises as a path in-between the status quo of expert-driven approaches and the primarily data-driven AutoML methodologies. The core result is that nnU-Net dramatically reduces the search space by generalizing the search process itself, thus being as more task-agnostic as possible. This is possible through the following three key points:

1. those design choices that do not depend on the dataset (i.e. do not require adaptation), are collected together to form a common robust configuration, the so-called "fixed parameters";

2. for as many as possible of the remaining decisions, explicit dependencies between the design choices, called "pipeline fingerprint", and the dataset specific properties, called "dataset fingerprint", are formulated as heuristic rules, thus enabling almost-instant configuring based on the application;

3. the few remaining design choices, here called "empirical parameters", are learned empirically from the dataset.

Therefore, nnU-Net presents itself as a self-configuring deep learning-based segmentation method for any new task, from pre-processing to training and post-processing. This holistic property differs it from existing research methods since it enables nnU-Net to cover the whole segmentation pipeline without any manual decision. Moreover, the configuration of nnU-Net is fast when compared to other existing methods since only few empirical choices have to be made. Third, the proposed model is also data efficient since the encoding of design choices based on heterogeneous data pools lead to a strong inductive bias when applied to limited data.

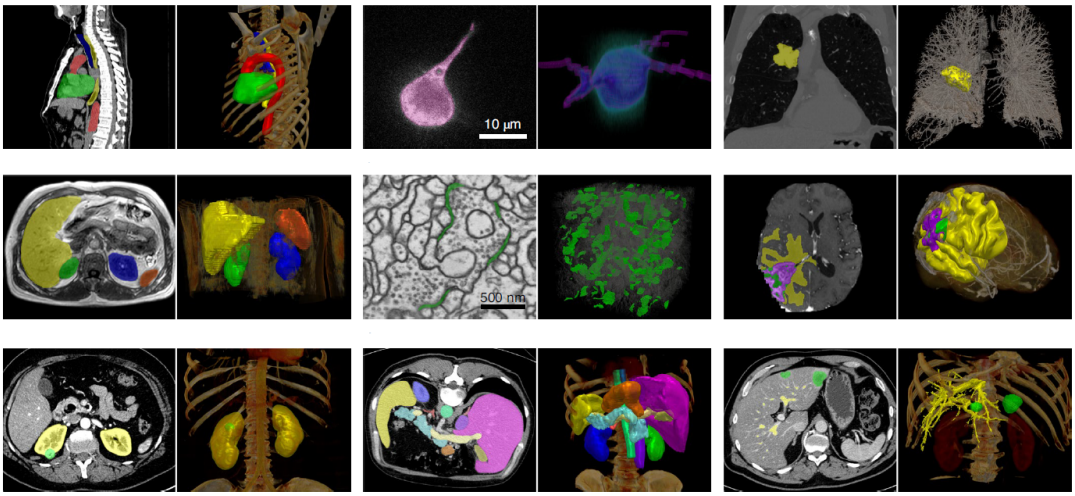
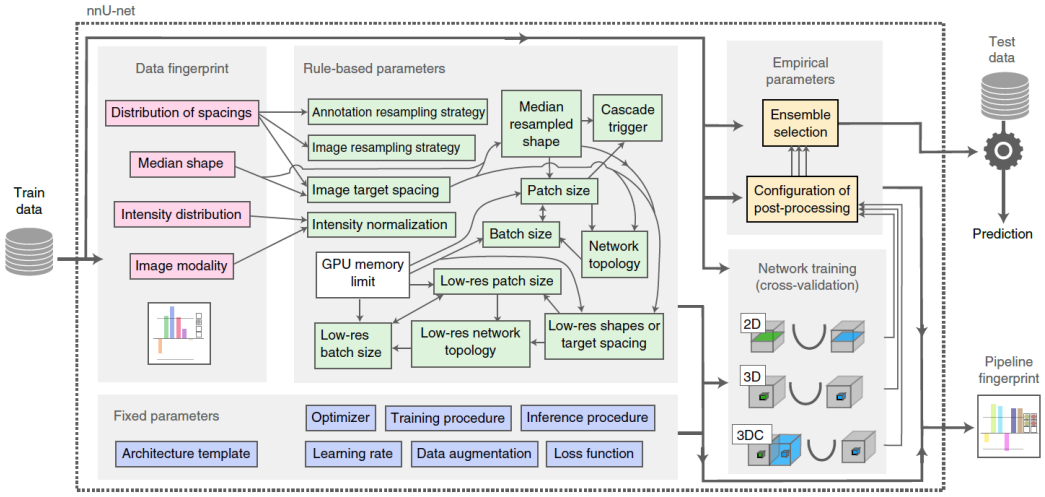


Figure 2.9. Examples of nnU-Net application to a set of international segmentation challenges as shown in the original paper by Isensee et al. [2021a].

The heterogeneous applicability of nnU-Net is demonstrated by the authors by its application on 13 different datasets – resulting in a total of 53 segmentation tasks – thus covering a wide range of image types, properties and structures, as it can be seen in Figure 2.9. In each one of these, nnU-Net automatically configures itself without human intervention, so that (with the exception of few empirical choices) no computational cost is required in addition to the one necessary for training. More in detail, the automated configuration of nnU-Net start with the generation of three different U-Net architectures: a base two-dimensional (2D) U-Net, a 3D U-Net working at full image resolution and a 3D U-Net cascade where the first U-Net operates on a downsampled version of the input image while the second refines at full resolution the results returned by the former. Finally, nnU-Net chooses the top-scoring configuration or ensemble after cross-validation against the dataset.
An in-depth representation as presented by Isensee et al. [2021a] of all the configuring steps taken by nnU-Net in order to reach its final form are presented in Figure 2.10. Given a novel segmentation task, "dataset fingerprints" (in pink) are extracted. Thin arrows represent the set of heuristic rules adopted to model parameter interdependencies. This rules then operate on the fingerprints to extract the "rule-based parameters" (in green) of the pipeline. Meanwhile, "fixed parameters" (in blue) – which do not need any kind of adaptation – are already predefined. A five-fold cross-validation is then performed to train the three configurations and the optimal ensemble is automatically selected while



Design choice	Required input	Automated (fixed, rule-based or empirical) configuration derived by distilling expert knowledge (more details in online methods)
Learning rate	–	Poly learning rate schedule (initial, 0.01)
Loss function	–	Dice and cross-entropy
Architecture template	–	Encoder–decoder with skip-connection ('U-Net-like') and instance normalization, leaky ReLU, deep supervision (topology-adapted in inferred parameters)
Optimizer	–	SGD with Nesterov momentum ($\mu = 0.99$)
Data augmentation	–	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring
Training procedure	–	1,000 epochs × 250 minibatches, foreground oversampling
Inference procedure	–	Sliding window with half-patch size overlap, Gaussian patch center weighting
Intensity normalization	Modality, intensity distribution	If CT, global dataset percentile clipping & z score with global foreground mean and s.d. Otherwise, z score with per image mean and s.d.
Image resampling strategy	Distribution of spacings	If anisotropic, in-plane with third-order spline, out-of-plane with nearest neighbor Otherwise, third-order spline
Annotation resampling strategy	Distribution of spacings	Convert to one-hot encoding → If anisotropic, in-plane with linear interpolation, out-of-plane with nearest neighbor Otherwise, linear interpolation

Image target spacing	Distribution of spacings	If anisotropic, lowest resolution axis tenth percentile, other axes median. Otherwise, median spacing for each axis. (computed based on spacings found in training cases)
Network topology, patch size, batch size	Median resampled shape, target spacing, GPU memory limit	Initialize the patch size to median image shape and iteratively reduce it while adapting the network topology accordingly until the network can be trained with a batch size of at least 2 given GPU memory constraints. for details see online methods.
Trigger of 3D U-Net cascade	Median resampled image size, patch size	Yes, if patch size of the 3D full resolution U-Net covers less than 12.5% of the median resampled image shape
Configuration of low-resolution 3D U-Net	Low-res target spacing or image shapes, GPU memory limit	Iteratively increase target spacing while reconfiguring patch size, network topology and batch size (as described above) until the configured patch size covers 25% of the median image shape. For details, see online methods.
Configuration of post-processing	Full set of training data and annotations	Treating all foreground classes as one; does all-but-largest-component-suppression increase cross-validation performance? Yes, apply; reiterate for individual classes No, do not apply; reiterate for individual foreground classes
Ensemble selection	Full set of training data and annotations	From 2D U-Net, 3D U-Net or 3D cascade, choose the best model (or combination of two) according to cross-validation performance

Figure 2.10. Proposed nnU-Net automatic configuration for biomedical image segmentation as shown in the original paper by Isensee et al. [2021a].

"empirical parameters" (in yellow) are settled to determine whether any kind of post-processing is needed. In such a way, nnU-Net can be adopted as an out-of-the-box deep

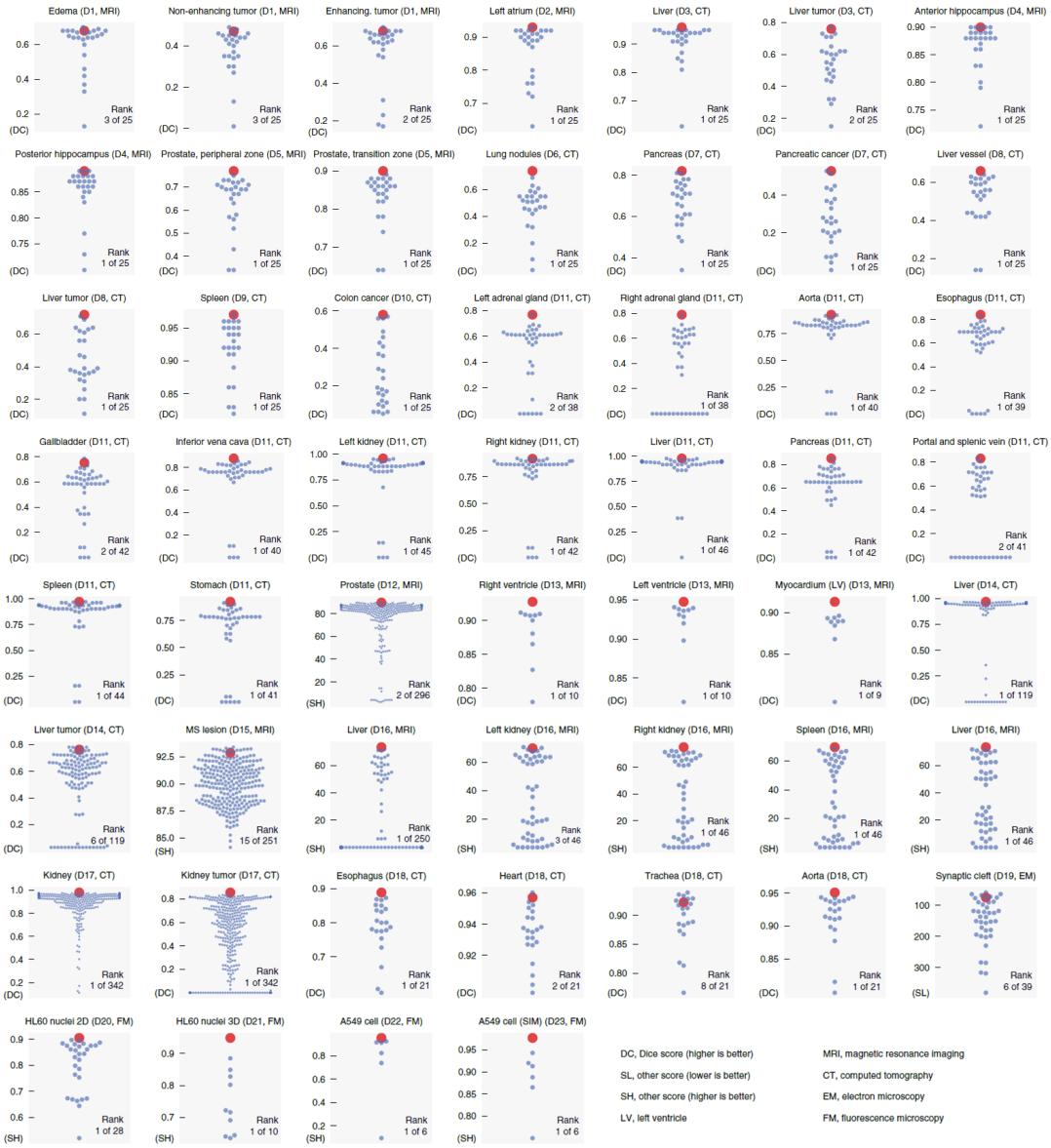


Figure 2.11. nnU-Net performances on 53 different segmentation tasks as shown in the original paper by Isensee et al. [2021a].

learning-based segmentation tool capable of handling a wide variety of both target structures and image properties: the international challenges to which nnU-Net has been applied comprise a variety of tumors, organs substructures, lesions and cellular structures

Reprinted by permission from Springer Nature: Springer Nature, Nature Methods, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Fabian Isensee et al., ©2020, The Author(s), under exclusive licence to Springer Nature America, Inc. (2020)

presented in two- as well as three-dimensional images acquired by magnetic resonance imaging (MRI), electron microscopy (EM), fluorescence microscopy (FM) and computed tomography (CT). The authors provided also an overview of the results obtained by nnU-Net against its competitors across 53 different segmentation tasks – observable in Figure 2.11 – which highlights how, despite its generic nature, it outperformed most existing and task-specific solutions. Specifically, nnU-Net set new state-of-the-art performances in 33 out of the 53 tasks, and performing close to top leaderboard entries in the remaining cases. The take-home message is therefore that the strong performances of nnU-Net are not reached by network architecture, loss function or training scheme fine tuning (hence the name nnU-Net, "no new net"), but by "systematizing the complex process of manual method configuration" (Isensee et al. [2021a]), which was previously addressed usually by purely empirical approaches such as trial and error.

2.2 Transfer learning

It is a well known fact that deep learning algorithms – CNNs included – require a huge amount of data for training. However, as already anticipated, the medical field often presents the data scarcity problem, with few scans available for analysis mainly due to the cost of expert-annotation (Kim et al. [2022]). A promising solution to this problem is the one of transfer learning (TL), i.e. by leveraging knowledge learned from other source tasks (Pan and Yang [2010]). TL derives from the cognitive conception that humans are able to solve similar tasks by exploiting previously-learned knowledge, with such knowledge being therefore transferred across similar tasks to improve performances on a new one. Formally, the definition of transfer learning was provided by Pan and Yang [2010]: "a domain consists of a feature space $\mathcal{X}$ and a marginal probability distribution $\mathbb{P}(X)$, where $X = \{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$. Given a specific domain denoted by $\mathcal{D} = \{\mathcal{X}, \mathbb{P}(\mathbb{X})\}$, a task is denoted by $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where $\mathcal{Y}$ is a label space and $f(\cdot)$ is an objective predictive function. A task is learned from the pair (x_i, y_i) where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ by using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$". In deep learning, transfer learning is the idea that knowledge can be transferred at the parametric level. Current transfer learning techniques in medical imaging implement knowledge transfer from natural imaging, usually from model trained on the ImageNet dataset[1]. Overall, two main paths to apply transfer learning have historically been delineated: feature extractor and fine-tuning. The main difference lies in the fact that the first freezes the convolutional layers whereas the latter updates parameters during model fitting. Still, even if some progress is done, the knowledge transferred between the two areas can be either not sufficient to achieve promising results in the medical task or make the transfer process quite unpredictable (Raghu et al. [2019]). Hence, given the lack of post-operative brain scans, the idea behind this study is to apply transfer learning from the pre-operative brain tumor segmentation

[1] "ImageNet large scale visual recognition challenge; ILSVRC"

task (a research area dealt with more in depth in literature) by fine-tuning the model in the end for the desired goal. This intuition builds itself upon two main consideration. First, the domains are both related to brain MRI segmentation, therefore the transfer process does not flow outside the medical area. Second, a "parallelism" between pre- and post-operative brain MRI scan labelling can be identified, as it will be discussed more in detail later, allowing the fine-tuning process to be more predictable.

2.3 BraTS

In the past, brain tumor datasets have remained relatively small, especially the ones presenting subjects imaged at a single institution. The Brain Tumor Segmentation Challenge (BraTS) provides the largest publicly available fully-annotated dataset for model development in the field of brain tumor medical imaging segmentation, establishing itself during the years as the benchmark competition for objective comparison of the several segmentation algorithms. Recent top-performing entries in the challenge are exclusively built upon deep neural networks, specifically on encoder-decoder structures with skip connections, a motif initially introduced by U-Net (see Chapter **??**). Architectural improvements upon U-Net have been proposed in the field of brain tumor segmentation, with auxiliary tools such as residual connections, attention mechanisms and densely connected layers, and it is worth noticing that winning contributions of 2018 (Myronenko [2019]) and 2019 (Jiang et al. [2020]) are both based on the introduction of a second decoder for regularization purposes. The RSNA ASNR MICCAI Brain Tumor Segmentation (BraTS) challenge (Menze et al. [2015], Bakas et al. [2017]) is a project started in 2012 in conjunction with the MICCAI conference with the goal of becoming the de facto benchmark for addressing the automated tumor sub-region segmentation from pre-operative multi-parametric Magnetic Resonance Imaging (mpMRI) scans. The BraTS dataset consists in a retroactive collection of brain tumor scans acquired with different apparati and protocols from several different institutions, thus presenting a highly heterogeneous image quality. The mpMRI scans present in the dataset fall under one of the following four MRI modalities: native (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2) or T2 Fluid Attenuated Inversion Recovery (FLAIR). Standardized pre-processing has been applied to all the scans, specifically:

1. conversion of the DICOM files to the NIfTI format (Cox et al. [2004]),
2. co-registration to the same anatomical template (Rohlfing et al. [2009]),
3. resampling to a uniform isotropic resolution ($1mm^3$),
4. skull-stripping.

The annotated tumor sub-regions in the BraTS scans (Baid et al. [2021]) comprise the Gd-enhancing tumor (ET), the peritumoral edematous/invaded tissue (ED), and the necrotic tumor core (NCR).

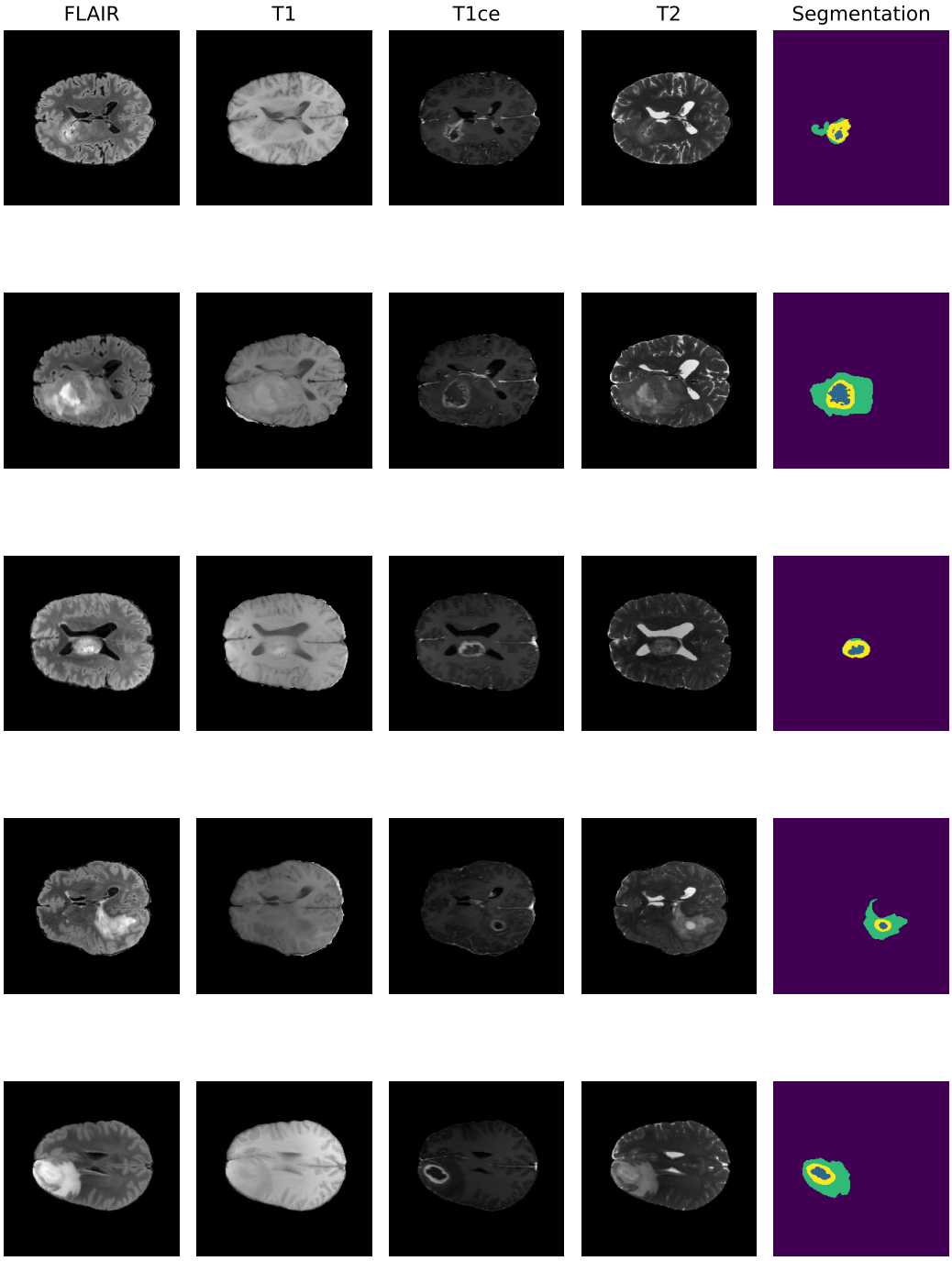


Figure 2.12. Examples of brain slices from five BraTS 2021 patients.

The ET is defined as the enhancing portion of the tumor which visually emerges on T1ce MRI. An opposite behaviour in T1ce MRI is exhibited instead by the NCR, i.e. the necrotic core of the tumor, which manifests an hypointense presence. The ED is the peritumoral edematous and infiltrated tissue, it presents an hyperintense signal on the FLAIR volumes and includes both the vasogenic edema and the non-enhancing tumor. Figure 4.1 shows the same brain slice depicted in the four different MRI modalities for five different patients retrieved from the BraTS 2021 dataset. The last column indicates instead the ground truth segmentation of the brain tumor, with the necrotic core (NCR) highlighted in blue, the enhancing tumor (ET) in yellow, and the edema (ED) in turquoise.

Year	Total Data	Training Data	Validation Data	Testing Data	Timepoint
2012	50	35	n/a	15	Pre-operative
2013	60	35	n/a	25	Pre-operative
2014	238	200	n/a	38	Longitudinal
2015	253	200	n/a	53	Longitudinal
2016	391	200	n/a	191	Longitudinal
2017	477	285	46	146	Pre-operative
2018	542	285	66	191	Pre-operative
2019	626	335	125	166	Pre-operative
2020	660	369	125	166	Pre-operative
2021	2040	1251	219	570	Pre-operative

Table 2.1. Evolution of the BraTS challenge dataset since its inception.

There are currently two guidelines regarding post-operative target definition in GBM: one from the European Organization for Research and Treatment of Cancer (EORTC) and the other from the Radiation Therapy Oncology Group (RTOG).
Both of them define the gross tumor volume (GTV) as the resection cavity (RC) plus the residual enhancing tumor. According to the RTOG guideline, surrounding edema should be included as well (Niyazi et al. [2016]). The goal of this work will be therefore to address the problem of post-operative brain tumor segmentation in a more comprehensive way, trying to achieve promising results in segmenting all the classes presented in the RTOG guideline.

2.3.1 Evaluation metrics

Following literature, the quality evaluation of the automatic brain implemented through Dice score and Hausdorff distance. Dice score is defined as

$$\text{DICE} = \frac{2|S_m \cap S_a|}{|S_m| + |S_a|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{2.4}$$

where S_m is the set of manually segmented tumor voxels while S_a the one of automatically segmented voxels, TP is the number of true positive voxels, FP of false-positives and FN of false-negatives.
The previously presented in Section 2.1.2 Hausdorff metric is instead defined as the maximum distance between the borders of two sets and is defined as

$$d_H = \max\left\{ \min_{p_a \in \mathcal{P}_a} d(p_a, p_m), \min_{p_m \in \mathcal{P}_m} d(p_a, p_m) \right\}, \tag{2.5}$$

where $\mathcal{P}_m$ is the set of vertices describing the border of the segmentation performed manually by the radiologist, $\mathcal{P}_a$ the one of vertices describing the border of the automatic segmentation, and $d(p_a, p_m)$ is the distance between two vertices (if not specified, the euclidean distance is adopted). That said, it emerges clearly that such definition is strongly sensitive to outliers. For such reason, the undirected 95th percentile Hausdorff distance is often taken as a more robust and informative metric. It is defined as

$$d_{H_{95}} = \max\left\{ 95\ \underset{p_m \in \mathcal{P}_m}{\textit{percentile}} \min_{p_a \in \mathcal{P}_a} d(p_a, p_m),\ 95\ \underset{p_a \in \mathcal{P}_a}{\textit{percentile}} \min_{p_m \in \mathcal{P}_m} d(p_a, p_m) \right\}. \tag{2.6}$$

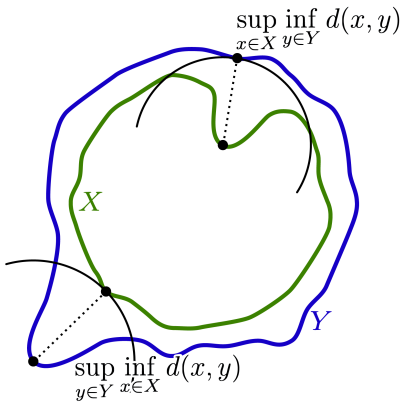


Figure 2.13. Qualitative visualization of the Hausdorff distance.

2.4 Automated glioma segmentation

As previously presented in Subection 2.1.2, nnU-Net has delineated itself as the state of the art and top-ranking model in several medical challenges. Among those, the RSNA ASNR MICCAI Brain Tumor Segmentation challenge. For this reason, the model intended to be trained in this study on pre-operative scans, whose knowledge will then be transferred to post-operative ones, is derived from the nnU-Net framework. Indeed, the winners of both

the validation and test phases for the BraTS 2020 challenge – as well as the BraTS 2021 one – are modifications obtained from vanilla applications of nnU-Net.
Subsections 2.4.1 and 4.1.1 present respectively the top-ranking model in the BraTS 2020 and 2021 challenges.

2.4.1 nnU-Net in BraTS 2020

Isensee et al. [2021b] applied the nnU-Net segmentation model to the task of the BraTS 2020 challenge. Even though the untouched nnU-Net baseline achieved good results, by implementing task-specific modifications regarding training, data augmentation and post-processing, performances improved substantially, enabling nnU-Net to take the first place in the competition. nnU-Net base preprocessing consists in normalizing the brain voxels of each image by subtracting their mean and dividing by their standard deviation (non-brain voxels remain untouched at 0). The proposed architecture follows a standard 3D U-Net pattern, with downsampling performed as strided convolutions while upsampling as transposed convolutions. The input patch size is selected to be 128×128×128 with a batch size of 2. In total, five downsampling are performed, leading to a bottleneck with a feature size of 4×4×4. The initial number of convolutions is fixed to 32, doubling at each downsampling step up to a maximum of 320. The decoder mirrors the encoder, and so do all the corresponding hyperparameters (see Figure 2.14). Leaky ReLU is adopted as nonlinearity and instance normalization is used in feature map normalization. The training loss is obtained as the sum of Dice and cross-entropy ones, operating on the three classes of edema, necrotic tumor core and enhancing tumor. Stochastic gradient descent with a starting learning rate of 0.01 and a Nesterov momentum of 0.99 is implemented. Training run for 1000 epochs, with the learning rate decaying following a polynomial schedule. Moreover, as anticipated, nnU-Net was further developed implementing BraTS-specific optimizations. More in detail: first, even though the provided labels are necrotic tumor core, edema and enhancing tumor, it has been shown (Myronenko [2019], Jiang et al. [2020]) that performances improve when segmentation is performed on the three partially overlapping regions (i.e. whole tumor, tumor core and enhancing tumor). Second, batch size is increased from 2 to 5 in an attempt to improve model accuracy by fitting more accurately the training data. Third, a more aggressive data augmentation strategy to the one proposed by default by nnU-Net is adopted on the fly, with the aim of increasing the robustness of the model. The following changes are therefore made:

- the probability of applying both rotation and scaling is increased from 0.2 to 0.3,
- the scale range increases as well from (0.85, 1.25) to (0.65, 1.6),
- the scaling factor is set individually for each axis,
- elastic deformation is applied with a probability of 0.3,
- additive brightness augmentation is applied with a probability of 0.3,
- the aggressiveness of the Gamma augmentation is increased.

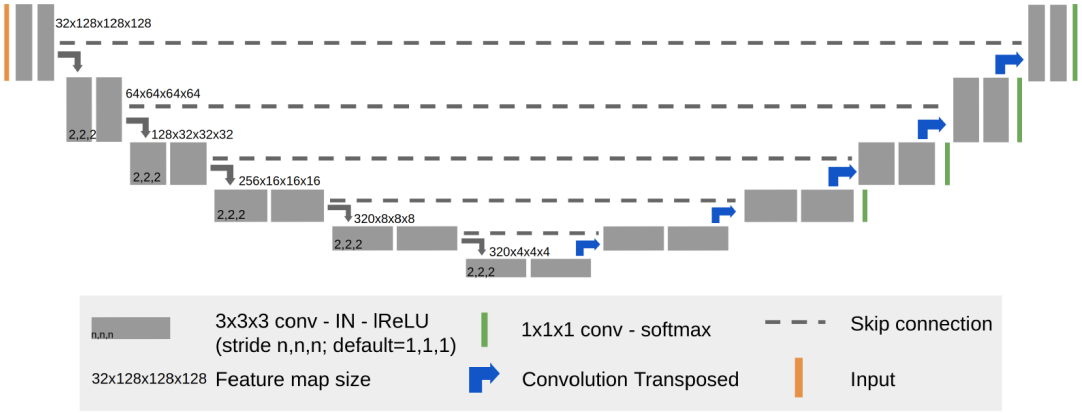


Figure 2.14. Network architecture generated by nnU-Net for BraTS 2020 as shown in the original paper by Isensee et al. [2021b].

Model	Whole	Core	Enh.	Mean
BL	91.60	87.23	80.83	86.55
BL*	91.85	86.24	80.18	86.09
BL*+R	91.75	87.24	**82.21**	86.73
BL*+R+DA	**91.87**	**87.97**	81.37	**87.07**
BL*+R+DA+BN	91.57	87.59	81.29	86.82
BL*+R+DA+BD	91.76	87.67	80.94	86.79
BL*+R+DA+BN+BD	91.70	87.21	81.70	86.87
BL*+R+DA*+BN	91.60	87.51	80.94	86.68
BL*+R+DA*+BN+BD	91.47	87.13	81.33	86.64

Table 2.2. Dice scores obtained by various nnU-Net versions on the BraTS 2020 training set segmentation task with a 5-fold CV as presented by Isensee et al. [2021b].

Fourth, instead of computing the Dice score for each sample in a batch and then averaging those values (sample Dice), the metric is computed over all samples just as if it were a single large sample (batch Dice). This comes in handy in the case of imperfect reference segmentation with accurate model predictions in avoiding misleading large gradients. Tables 2.2, 2.3 and 2.4 present some segmentation metrics obtained by various nnU-Net versions on both the training and validation set of the BraTS 2020 challenge. The different models are identified by the following abbreviations:

Model	Whole	Core	Enh.	Mean
BL	90.60	84.26	77.67	84.18
BL*	90.93	83.70	76.64	83.76
BL*+R	90.96	83.76	77.65	84.13
BL*+R+DA	90.90	84.61	78.67	84.73
BL*+R+DA+BN	**91.24**	85.04	79.32	85.20
BL*+R+DA+BD	90.97	83.91	77.48	84.12
BL*+R+DA+BN+BD	91.15	84.19	**79.99**	85.11
BL*+R+DA*+BN	91.18	**85.71**	79.85	**85.58**
BL*+R+DA*+BN+BD	91.19	85.24	79.45	85.29

Table 2.3. Dice scores obtained by various nnU-Net versions on the BraTS 2020 validation set using the five models from the training CV as ensemble as presented by Isensee et al. [2021b].

Model	Whole	Core	Enh.	Mean
BL	4.89	5.91	35.10	15.30
BL*	4.23	6.01	41.06	17.10
BL*+R	4.41	8.80	29.82	14.34
BL*+R+DA	4.70	5.62	29.50	13.28
BL*+R+DA+BN	3.97	**5.17**	29.25	12.80
BL*+R+DA+BD	4.11	8.60	38.06	16.93
BL*+R+DA+BN+BD	**3.72**	7.97	**26.28**	12.66
BL*+R+DA*+BN	3.73	5.64	26.41	**11.93**
BL*+R+DA*+BN+BD	3.79	7.77	29.23	13.60

Table 2.4. Hausdorff95 distances obtained by various nnU-Net versions on the BraTS 2020 validation set using the five models from the training CV as ensemble as presented by Isensee et al. [2021b].

- BL/BL*: baseline nnU-Net (* indicates a batch size of 5 instead of 2),
- R: training with overlapping regions,
- DA/DA*: more aggressive data augmentation (* means that brightness augmentation is applied for each modality with a probability of 0.5),
- BN: batch normalization is used, as opposed to instance normalization,

- BD: batch Dice is adopted, as opposed to sample Dice.

Each configuration is trained on the training set with a 5-fold cross-validation, and those five models are then applied as an ensemble to the validation set. Inflated values can be observed in the Hausdorff95 distance values for the enhancing tumors, which is due to the fact that during post-processing this section of tumor is entirely removed if the volume predicted is less than a given threshold. This strategy leads to the removal of some true positive predictions, but in practice the net gain in robustness out-weights the losses. Moreover, wrong enhancing tumor predictions (false positive, i.e. reference segmentation does not present such class) are hard coded to receive a Hausdorff95 distance of 373.13, and those outliers strongly dominate the mean aggregation.
Following the results on the validation set, the authors decided to build as final submission an ensemble of 20 models – 5 models configured as BL*+R+DA*+BN+BD, 5 models configured as BL*+R+DA+BN+BD and 15 models configured as BL*+R+DA*+BN –, which lead to a final Dice scores of 91.24, 85.06 and 79.89, and Hausdorff95 distances of 3.69, 7.82 and 23.50 for the whole tumor, tumor core and enhancing tumor (on the validation set) respectively, allowing them to win the subsequent test phase and achieve the first place in the competition.

Chapter 3

State of the art

As any doctor can tell you, the most crucial step toward healing is having the right diagnosis. If the disease is precisely identified, a good resolution is far more likely. Conversely, a bad diagnosis usually means a bad outcome, no matter how skilled the physician.

- Dr. Andrew Weil

3.1 Pre-operative

A lot of effort on (semi-)automatic brain tumor segmentation and classification from MR imaging scans has been put since the late nineties, given its non-invasive nature and the challenging and time-consuming features of manual segmentation (Biratu et al. [2021]). This led to an ample literature focusing on segmentation, ranging from traditional machine learning to deep learning techniques. For comparison purposes, here will be presented only those methods operating on BraTS-MICCAI datasets (see Section 2.3) and introducing the Dice Similarity Coefficient (see Subsection 2.3.1) as performance.

3.1.1 Machine learning

Tang et al. [2015] achieved a DSC of 96.5% on the BraTS 2012 dataset via a random forest (RF) with intensity, similarity and blobness as features. Pre-processing included registration and normalization, while post-processing consisted in an independent connected component analysis. Chen et al. [2017b] proposed instead a support vector machine (SVM) having gray-level co-occurrence matrix (GLCM) and gray statistical as features, together with simple linear iterative clustering as pre-processing, which achieved a DSC of 86.1% on the BraTS 2013 dataset. Bougacha et al. [2018] illustrated two methods: one based on an artificial neural network (ANN) and another on a SVM, both working on intensity and texture as features. The former achieved – on the BraTS 2015 dataset – a DSC of 90.8%, while the latter a DSC of 88.7%, with neither pre- nor post-processing. Ma et al. [2018] expanded the BraTS 2015 dataset with some patient MRs from The Cancer Genome Atlas

Publication	BraTS	Model	DSC
Tang et al. [2015]	2012	RF	96.5
Chen et al. [2017b]	2013	SVM	86.1
Bougacha et al. [2018]	2015	ANN	90.8
Bougacha et al. [2018]	2015	SVM	88.7
Ma et al. [2018]	2015	ccRF	89.0
Hatami et al. [2019]	2015	RF	98.4
Fulop et al. [2020]	2015	RF	85.5

Table 3.1. Summary of most relevant studies regarding GBM segmentation with ML techniques.

Glioblastoma (TCGA-GBM) collection and suggested a concatenated and connected random forest (ccRF) with multiscale patch driven active contour as post-processing which achieved a DSC of 89.0%. Hatami et al. [2019] proposed another random forest with noise removal and morphological filtering which scored a DSC of 98.4% on the BraTS 2015 dataset. Fulop et al. [2020] presented a further RF model with histogram enhancement and morphological filtering which achieved, on the BraTS 2015 dataset as well, a DSC of 85.5%. Table 3.1 gives a summarized overview of these methods.

3.1.2 Deep learning

The advent of DL drastically reduced the number of ML methods applied for brain tumor segmentation given the astonishing results and the minor need for handcrafted features. It is therefore noticeable an unilateral shift from one field to the other starting from 2015. Pereira et al. [2016] proposed a custom Convolutional Neural Network (CNN) and a pre-processing pipeline including bias field correction, intensity and patch normalization, achieving a DSC of 88.0% on the BraTS 2015 dataset. Ding et al. [2019b] presented instead a deep residual dilate network with middle supervision which scored a DSC of 86.0% on z-normalized data from the BraTS 2015 dataset. A stack multi-connection simple reducing net (SMCSRNet) was proposed by Ding et al. [2019a], which achieved on the same configuration a DSC of 83.4%. Razzak et al. [2019] introduced a model called Two-Pathway-Group CNN to leverage both local and global contextual features, which achieved a DSC of 89.2% on the BraTS 2015 dataset with bias correction and intensity normalization as pre-processing. Wang et al. [2018] proposed image specific fine tuning after a custom CNN, scoring a DSC of 86.3% on the BraTS 2015 dataset. Xu et al. [2019] experimented with a LSTM multi-modal U-Net on the BraTS 2015 dataset, achieving a DSC of 73.1%. Zhou et al. [2020b] suggested a model comprising cascaded 3D U-Nets, which achieved a DSC of 89.4% on the BraTS 2017 dataset pre-processed with intensity normalization, resizing and bias field correction. Ye et al. [2021] introduced a parallel pathway dense neural network with weighted fusion structure able to score a DSC of 88.4% and 88.7% respectively for the BraTS 2015 and BraTS 2017 datasets. Li et al. [2019] extended an Inception U-Net with a cascaded training strategy achieving a DSC of 89.0% on z-normalized data from the BraTS 2017 dataset.

Publication	BraTS	Model	DSC
Pereira et al. [2016]	2015	CNN	88.0
Ding et al. [2019b]	2015	RDM-Net	86.0
Ding et al. [2019a]	2015	SMCSRNet	83.4
Razzak et al. [2019]	2015	2PG-CNN	89.2
Wang et al. [2018]	2015	BIFSeg	86.3
Xu et al. [2019]	2015	LSTM U-Net	73.1
Ye et al. [2021]	2015	Dense CNN	88.4
Ye et al. [2021]	2017	Dense CNN	88.7
Zhou et al. [2020b]	2017	3D U-Nets	89.4
Li et al. [2019]	2017	Inception U-Net	89.0
Ben Naceur et al. [2020]	2018	Attention CNN	86.2
Sun et al. [2021]	2018	3D FCN	90.0
Myronenko [2019]	2018	U-Net + VAE	82.2
Aboelenein et al. [2020]	2018	HTTU-Net	86.5
Sun et al. [2021]	2019	3D FCN	89.0
Ali et al. [2020]	2019	3D-CNN + U-Net	90.6
Jiang et al. [2020]	2019	Cascaded U-Net	85.3
Henry et al. [2021]	2020	U-Net + DS	87.0
Cirillo et al. [2020]	2020	3D-GAN	82.3
Yuan [2021]	2020	Attention U-Net	84.8
Jia et al. [2021]	2020	H^2NF-Net	85.2
Isensee et al. [2021b]	2020	nnU-Net	85.4
Futrega et al. [2021]	2021	nnU-Net	91.6
Luu and Park [2021]	2021	Attention U-Net	91.5
Siddiquee and Myronenko [2022]	2021	Train Mod. U-Net	89.1

Table 3.2. Summary of most relevant studies regarding GBM segmentation with DL techniques.

Ben Naceur et al. [2020] proposed another Deep Convolutional Neural Network (DCNN) implementing a selective attention technique inspired by the Occipito-Temporal pathway achieving a DSC of 86.2% on the intensity-normalized BraTS 2018 dataset. Sun et al. [2021] applied a 3D Fully Convolutional Network (FCN) on z-normalized MRI scans from both BraTS 2018 and BraTS 2019 datasets, achieving a DSC of 90.0% and 89.0% respectively. Aboelenein et al. [2020] presented HTTU-Net, a hybrid two track U-Net which scored a DSC of 86.5% on the BraTS 2018 dataset. Ali et al. [2020] proposed instead an ensemble of a 3D-CNN and U-Net able to achieve a DSC of 90.6% on the BraTS 2019 dataset. Myronenko [2019] proposed an encode-decoder structure with an auxiliary variational auto-encoder branch which won the 1st place in the BraTS 2018 challenge scoring a DSC of 82.2% on the testing dataset. Jiang et al. [2020] devised a novel two-stage cascaded U-Net which won the BraTS 2019 challenge achieving a DSC of 85.3% on the

testing dataset. Henry et al. [2021] proposed an U-Net-like architecture with deep supervision and stochastic weight averaging which ranked among the top-10 entries in the BraTS 2020 challenge with a DSC of 87.0%. Cirillo et al. [2020] introduced Vox2Vox, a 3D-GAN for brain tumor segmentation which achieved a DSC of 82.3% on the BraTS 2020 dataset. Yuan [2021] obtained the 3rd place in the BraTS 2020 challenge by devising SA-Net, an U-Net-like model with scale attention blocks, achieving a DSC of 84.8%. Jia et al. [2021] ranked 2nd on the BraTS 2020 leaderboard by proposing H^{2}NF-Net, a hybrid high-resolution and non-local feature network which scored a DSC of 85.2%. The 1st place was instead obtained – as presented in Subsection 2.4.1 – by Isensee et al. [2021b] with some ad hoc modifications from the vanilla 3D U-Net returned by the nnU-Net pipeline, with a DSC of 85.4%. Similarly, another optimized version of the nnU-Net output (presented more in detail in Subsection 4.1.1) proposed by Futrega et al. [2021] ranked 1st on the BraTS 2021 validation leaderboard, with a DSC of 91.6%. Luu and Park [2021] extended nnU-Net as well, achieving a DSC of 91.5% on the BraTS 2021 dataset by adding axial attention in the decoding path. Siddiquee and Myronenko [2022] supported instead a U-Net-like architecture with some modifications of the network training process in order to minimize redundancy under perturbations, achieving a DSC of 89.1% on the BraTS 2021 dataset. Table 3.2 gives a summarized overview of these method,

3.2 Post-operative

While many efforts have been put in order to improve the segmentation of pre-operative scans, no comparable result has still been achieved for post-operative ones. Literature regarding post-operative brain tumor segmentation is still scarce, and the few studies available either focus only on the resection cavity segmentation or achieve strongly sub-human performances due to the general lack of data and the mixed training between pre- and post-operative scans. This is mainly due to the two following factors: first, the absence of a wide dataset like BraTS requires researchers to deploy their model on small private collections, hence reducing comparability and generalizability. Second, post-operative segmentation is way more tricky than pre-operative one, with possible presence/absence of some structures, such as the enhancing part of the tumor. For this reason, part of what is defined as post-operative segmentation focuses on segmenting pre-operative classes in post-operative tumor recurrence.
Zeng et al. [2017], for example, proposed a hybrid generative-discriminative model which achieved a DSC of 59.3% for whole tumor (71.0%), tumor core (56.0%) and enhancing tumor (51.0%) on 32 post-operative patients from the BraTS 2016 dataset. The drop is noticeable when considering that their application on the BraTS 2016 HGG pre-operative patients returned a DSC of 85.3%. Ghaffari et al. [2022], in a process similar to the one implemented in this work, studied the effect of transfer learning from pre-operative scans to post-operative ones with a Dense U-Net, achieving an average DSC of 73.3% for whole tumor (83.0%), tumor core (77.0%) and enhancing tumor (60.0%) on 15 patients treated with post-operative radiation therapy. Here is noticeable as well the gap between the performances reached on pre-operative cases and post-operative ones, with the model scoring a DSC of 83.7% on the BraTS 2020 validation dataset for whole tumor (DSC

90.0%), tumor core (83.0%) and enhancing tumor (78.0%) subregions. As anticipated, some studies focused also on segmenting the resection cavity only. Jungo et al. [2018] introduced a FCNN with uncertainty estimates able to segment the cavity in 30 post-operative patients with a DSC of 79.2%. Ermis et al. [2020] devised a DenseNet-like architecture, observing an agreement with three raters (mean DSC 85.0%, 84.0%, 86.0%) of DSC 83.0%, 81.0% and 81.0% respectively on 30 post-operative GBM patients. Lotan et al. [2022] built a dataset including both pre- and post-operative scans for a total of 432 cases, achieving a DSC of 79.7% for whole tumor (83.0%), tumor core (84.0%) and enhancing tumor (72.0%). Helland et al. [2022], in what is likely the widest study on post-operative GBM segmentation, achieved a DSC of 88.5% in segmenting residual tumor tissue in early post-operative MRI scans from 645 patients from 13 hospitals across Europe and the US.

Chapter 4

Experimental setup and methods

If you trust software, you tend to believe it
- Enrico Coiera

4.1 Segmentation architecture

4.1.1 nnU-Net in BraTS 2021

Futrega et al. [2021], following the performances of U-Net and nnU-Net-based models in the BraTS 2020 challenge, proposed an optimized architecture for the brain tumor segmentation task of the BraTS 2021 challenge which won the validation phase and ended in third place on the final leaderboard after the test phase[1]. Given the huge boost in the number of patients provided by the challenge in comparison to the previous year – BraTS 2021 presents 1251 patients in the training set, juxtaposed to "only" 369 in the BraTS 2020 one – the authors decided to run extensive ablation studies in order to select the optimal U-Net variant first (see Section 2.1.1), and the best training schedule then.
As a first step, all four modalities (FLAIR, T1, T1ce, T2) were stacked in order to get each sample shaped as (4, 240, 240, 155), where input tensor follow the (C, H, W, D) layout. Then, cropping was performed: zero-value voxels on the borders (redundant background) were removed as they do not convey useful information to the network. All volumes have been normalized afterwards, by subtracting the mean and dividing by the standard deviation, both computed separately for each channel within the non-zero region. This allows background voxels to keep their value at zero. As a supplement, in order to distinguish between those voxels with value close to zero and actual background ones, an additional input channel was created with one-hot encoding for those foreground voxels

[1] "NVIDIA Data Scientists Take Top Spots in MICCAI 2021 Brain Tumor Segmentation Challenge"

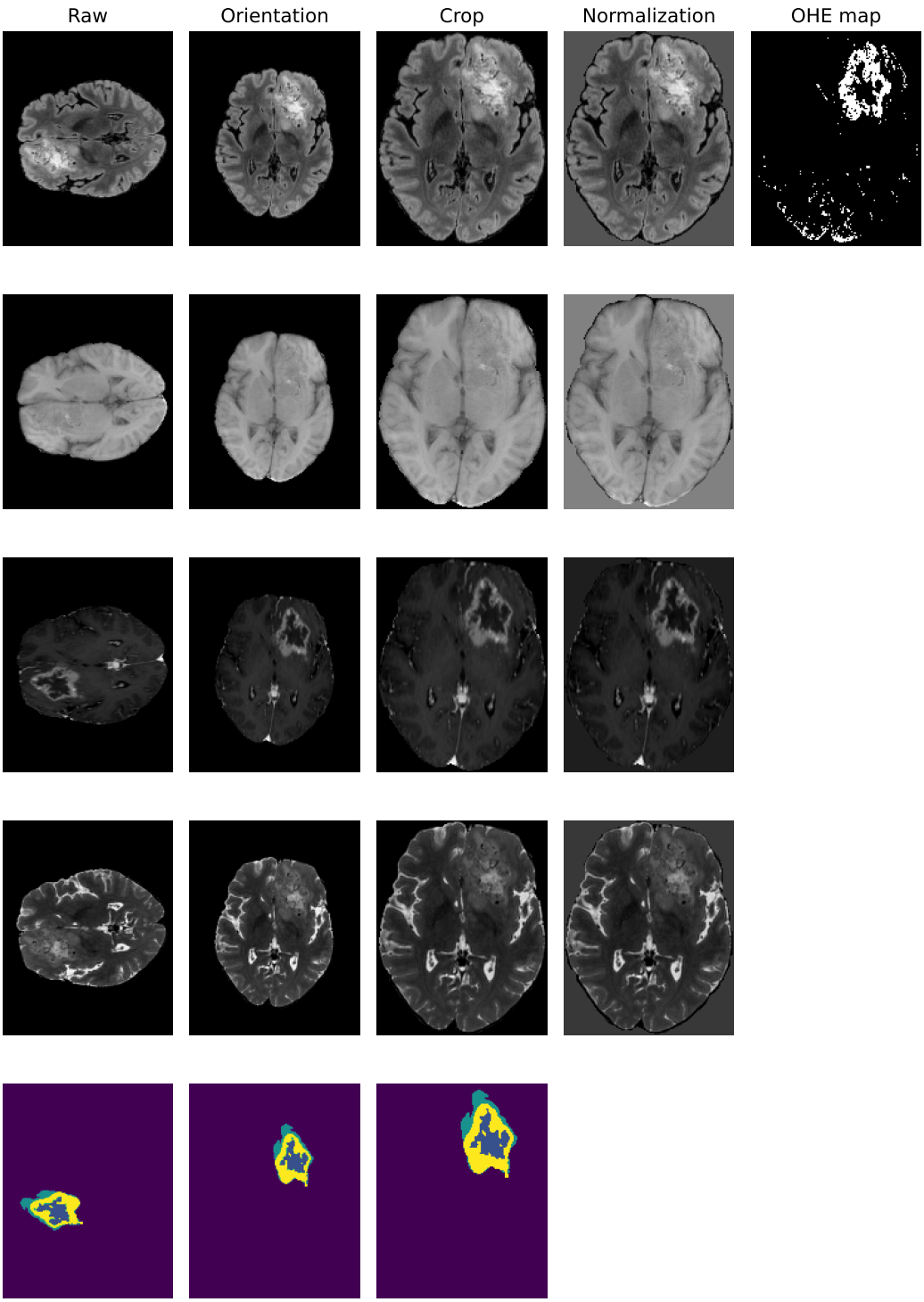


Figure 4.1. Preprocessing applied to each modality (FLAIR, T1, T1ce and T2 respectively).

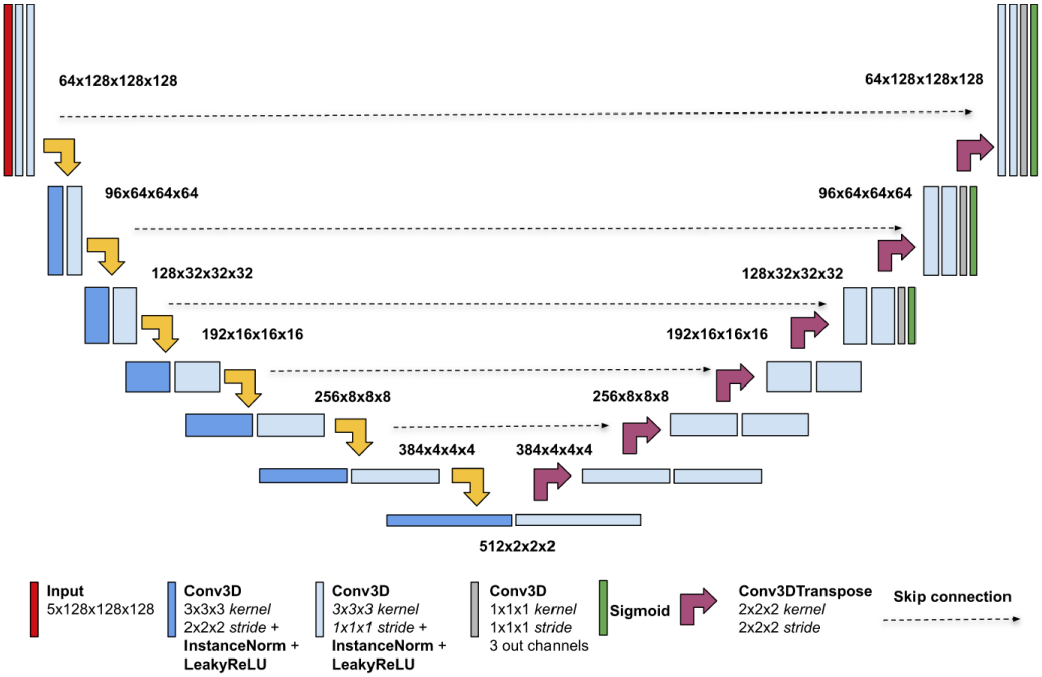


Figure 4.2. Final architecture proposed for the BraTS 2021 challenge as described in the original paper by Futrega et al. [2021].

and was stacked with input data. With the aim of alleviating the overfitting problem, data augmentation have been applied for this challenge as well. Specifically:

- a patch of dimensions (5, 128, 128, 128) was cropped randomly from the input volume. Moreover, with probability of 0.4, it was guaranteed that some foreground voxels with positive class in the ground truth were actually present in the cropped region,
- a random value is sampled uniformly from (1.0, 1.4) with probability of 0.15 and the image is resized to such sampled value times its original size with cubic interpolation, whereas the ground truth is zoomed with nearest neighbour interpolation,
- for each axis independently, the input was flipped along that axis with probability of 0.5,
- random Gaussian noise with zero mean and standard deviation uniformly sampled from (0, 0.33) is sampled and added to the input with probability of 0.15,

- Gaussian blurring with Gaussian kernel standard deviation uniformly sampled from (0.5, 1.5) is applied to the input with probability of 0.15,
- brightness is modified with probability of 0.15 by sampling uniformly from (0.7, 1.3) a random value and by multiplying for such value all input voxels,
- contrast is modified with probability of 0.15 by sampling uniformly from (0.65, 1.5) a random value, by multiplying for such value all input voxels and by clipping them to their original value range.

Based on the authors experiments, the top-scoring architectural U-Net variant turned out to be the standard 3D one returned by the nnU-Net framework (see Subsection 2.4.1). Such model was therefore selected for further exploration. Figure 4.2 shows the final architecture proposed for the BraTS 2021 challenge. As a start, experiments showed that increasing the depth of the encoder from 6 to 7 and changing the number of channels from [32, 64, 128, 256, 320, 320] to [64, 96, 128, 192, 256, 386, 512] further improves its performance. Following Isensee et al. [2021b], labels were translated into the three overlapping regions and the loss was computed separately for each region as the sum of binary cross-entropy and Dice score (in its batch Dice form). Focal loss with gamma parameter set to 2 was experimented in place of binary cross-entropy but it led to worse performances. Finally, deep supervision (Zhu et al. [2017]) was implemented by adding two heads (in green, in Figure 4.2) at the end of the second- and third-last depth levels in the decoding path. Computing the loss on different levels improves a network performance by leveraging a better gradient flow. Hence, if p_i is the prediction for label y_i, with i being the depth level, the loss function is computed as

$$\mathcal{L}(y_1, y_2, y_3, p_1, p_2, p_3) = \mathcal{L}(y_1, p_1) + \frac{1}{2}\mathcal{L}(y_2, p_2) + \frac{1}{4}\mathcal{L}(y_3, p_3). \tag{4.1}$$

Model	Standard U-Net	Attention U-Net	Residual U-Net
Fold 0	**0.9087**	0.9044	0.9086
Fold 1	**0.9100**	0.8976	0.9090
Fold 2	**0.9162**	0.9051	0.9140
Fold 3	**0.9238**	0.9111	0.9219
Fold 4	**0.9061**	0.8971	0.9053
Mean Dice	**0.9130**	0.9031	0.9118

Table 4.1. Experimental results as presented in Futrega et al. [2021] showing that 3D baseline U-Net is the highest-scoring variant in the 5-fold comparison (although residual variants have similar scores, training times were noticeably longer).

Model	Standard	Attention	Residual	DS	Focal
Fold 0	0.9087	0.9044	0.9086	**0.9111**	0.9094
Fold 1	0.9100	0.9110	0.9103	**0.9115**	0.9026
Fold 2	0.9162	0.9157	**0.9175**	**0.9175**	0.9146
Fold 3	0.9238	0.9234	0.9233	**0.9268**	0.9229
Fold 4	0.9061	0.9061	0.9070	**0.9074**	0.9072
Mean Dice	0.9130	0.9130	0.9134	**0.9149**	0.9133

Table 4.2. Experimental results as presented in Futrega et al. [2021] showing that the only extension significantly improving the 5-fold average Dice score over the standard U-Net was the implementation of deep supervision (DS).

Model	DS	Deeper	Channels	One-hot	D+C+O
Fold 0	0.9111	**0.9118**	0.9107	0.9109	**0.9118**
Fold 1	0.9115	0.9140	0.9135	0.9132	**0.9141**
Fold 2	0.9175	0.9170	0.9173	0.9174	**0.9176**
Fold 3	**0.9268**	0.9256	0.9265	0.9263	**0.9268**
Fold 4	0.9074	**0.9079**	0.9072	0.9075	0.9076
Mean Dice	0.9149	0.9152	0.9150	0.9050	**0.9156**

Table 4.3. Experimental results as presented in Futrega et al. [2021] showing that performing all modifications (deep supervision, deeper encoder, different number of channels and additional one-hot input) further improves the network score.

In order to match the patch dimension and the output of the lower levels, nearest neighbor interpolation was adopted. Tables 4.1, 4.2 and 4.3 try to summarize all these decisions by highlighting the average Dice scores recorded across several 5-fold comparison experiments. Inference is performed with the aid of a sliding window having the same size as the training patch, i.e. (128, 128, 128), with two consecutive windows overlapping by half such a patch. Following once again Isensee et al. [2021b], Gaussian importance weighting is applied to average the overlapping regions prediction so that center voxels have higher importance. Furthermore, test time augmentation is performed by creating eight different versions of the input volume through each possible flip along the x, y and z axis combination. Inference probabilities are then averaged across all these eight results to form one final prediction. Since optimization dealt with the three overlapping regions of enhancing tumor, tumor core and whole tumor, it is necessary to convert each class back to the original classes of necrosis, enhancing tumor and edema. The following strategy have been proposed: if the probability for a voxel of being whole tumor is less than 0.45, it is

considered as background (label 0). Otherwise, if the probability for such voxel of being classified as tumor core, the voxel is considered as edema (label 2). Finally, depending on whether the probability of being considered enhancing tumor is less than 0.45, necrosis (label 1) or enhancing tumor (class 4) are assigned as class respectively. Furthermore, to avoid the case in which the model predicts some voxels as enhancing tumor while none is present in the ground truth, a the following post-processing strategy was applied: once detected an enhancing tumor connected component, if its size is smaller than 16 voxels and the average probability is less than 0.9, it is replaced by necrosis voxels (so that it is still considered as tumor core). Enhancing tumor voxels are converted to necrosis also in the case in which there are less than 73 voxels classified as ET with an average probability lower than 0.9. All such hyperparameters have been selected via a grid-search method to maximize the average score on a 5-fold cross-validation.
Each experiment followed a 1000-epochs training using the Adam optimizer with a learning rate equal to 0.0003 and a weight decay of 0.0001. A linear warm-up of the learning rate was used during the first 250 iterations and then decreased with a cosine annealing scheduler. For evaluation, a 5-fold cross-validation was adopted and the average highest scores were compared.

4.2 Dataset and harmonization

The Molinette Hospital has made available a retrospectively-collected dataset comprising 71 patients suffering from Grade IV Glioma that eventually underwent surgery at the institute. All MRI scans acquired in situ were accessible by default on the hospital SYNAPSE® Mobility PACS system. In the eventuality that some scans were imaged at a different centre, such MRs have been routinely uploaded on the BRAINLAB© neuronavigation system available to the neurosurgery department. At the time of hospitalization, written consensus for personal, biological and radiological data processing for scientific purposes was explicitly asked and registered on the InterSystem TrakCare© information system. For the sake of this study, all data were anonymized beforehand as instructed by the GDPR through the apposite function available in the HOROS© DICOM image viewer. Some patients have been excluded due to one or more of the following: underage subject, absence of either T1ce or FLAIR sequences, post-operative complications (such as hemorrages or abscesses) that could possibly invalidate the segmentation, or GBM histologically not-yet confirmed. Each patient presents one or more scans, with post-operative acquisition times ranging from immediately after the surgery (max 48h) to 12+ months later. Both pre- and post-operative segmentations have been performed semi-automatically through the SmartBrush feature of the Cranial Planning workflow inside the BRAINLAB© neuronavigation system (Build 3.3.1.404). The volumetric representation is reconstructed by the software by combining the semi-automatic segmentations in the axial, coronal and sagittal planes. The axial view is then extracted and, if necessary, manually adjusted. The process was carried on individually by 4 neurosurgeons, 1 neuroradiologist and 1 medical student, all supervised by a further neurosurgeon.

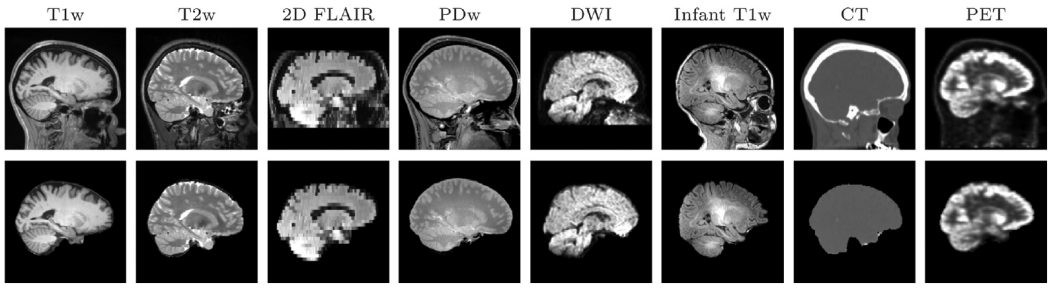


Figure 4.3. Examples of SynthStrip brain extraction from a wide variety of acquisition modalities as presented by Hoopes et al. [2022].

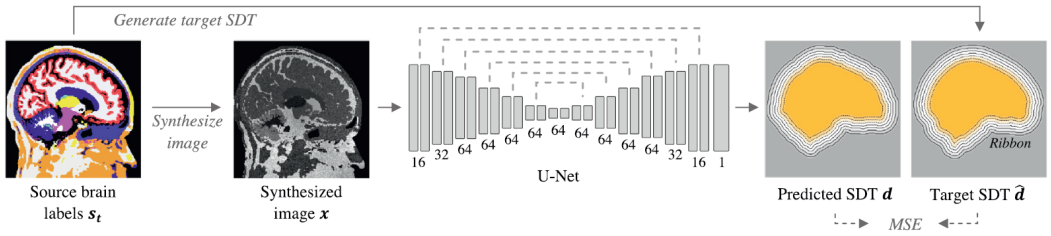


Figure 4.4. Synthstrip training framework as presented by Hoopes et al. [2022].

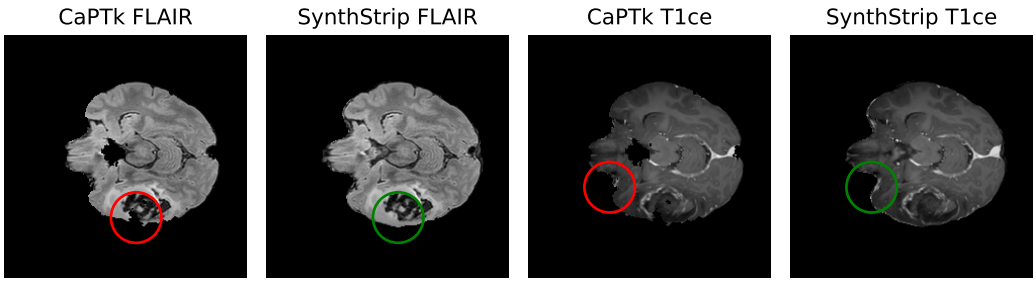


Figure 4.5. Qualitative comparison of the CaPTk skull-strip and SynthStrip algorithms.

Reprinted by permission under Creative Commons Attribution 4.0 International, SynthStrip: skull-stripping for any brain image, Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca et al., ©CC BY 4.0 (2022) – no changes were made –

The segmented classes consist in necrosis, core tumor and whole tumor – similarly to BraTS – for the pre-operative case, while resection cavity, enhancing tumor and whole tumor (enhancing tumor + edema) for the post-operative one. Whole tumor segmentation was performed on the FLAIR modality, whereas the remaining two classes on the T1ce one, both for the pre- and post-operative case.
Ground truth segmentations are exported from the BRAINLAB© neuronavigation system in their "burned-in" form, meaning that the classes overlap with the corresponding MRI scan and emerge visually by being identifiable by those voxels with the highest intensity. Since the model will be trained on the BraTS dataset, all scans received from the Molinette Hospital need to be harmonized in a "BraTS-like" manner. There are two main steps in order to reach such goal: co-registration to the SRI-24 template (Rohlfing et al. [2009]) with uniform isotropic resolution (1mm^3) and skull-stripping. The whole pipeline is available through the Cancer Imaging Phenomics Toolkit (Davatzikos et al. [2018], Pati et al. [2020]). Yet, the skull-stripping performed by CaPTk is achieved via a DeepMedic network (Kamnitsas et al. [2016], Kamnitsas et al. [2015]) trained on the BraTS 2017 training data whose outcomes are often unsatisfactory.
For this reason, the CaPTk pipeline will be limited to the SRI-24 atlas co-registration. Skull-stripping will instead be performed through SynthStrip, a novel and more robust skull-stripping method recently proposed by Hoopes et al. [2022]. Figure 4.5 presents a qualitative comparison of the two algorithms for a patient from Molinette Hospital, highlighting how CaPTk could eventually lead also to a loss of information regarding the tumor itself.

Skull-stripping

Skull-stripping – known also as brain extraction – consists in the act of removing non-brain signal from MRI data. It is a useful step in further anonymizing the brain scans since it mitigates potential facial recognition/reconstruction of the patient and a core component of many neuroimaging pipelines. Classical skull-stripping methods are well-explored but are unfortunately strongly tailored to images with specific acquisition properties. The robustness of SynthStrip derives from the fact that the underlying U-Net is exposed to a deliberately unrealistic range of contrasts, artifacts and anatomies, thus making the model agnostic to acquisition specifics since it never samples from real data during training. Figure 4.4 summarizes the training framework developed for SynthStrip: at each step, a gray-scale image x with arbitrary contrast is synthesized from a randomly transformed brain segmentation s_t. The 3D U-Net (see Subsection 2.1.1) receives then x as input and predicts the thresholded signed distance transform (SDT) d representing the distance between the skull boundaries and each voxel. The adopted loss between predictions and ground truth is a variant of the MSE, defined as

$$\mathcal{L}_{\text{SDT}} = \frac{\sum_{i\in\mathcal{P}} \omega_i (d_i - \hat{d}_i)^2}{\sum_{i\in\mathcal{P}} \omega_i}, \qquad \omega_i = \begin{cases} b & \text{if } |\hat{d}_i| > t, \\ 1 & \text{otherwise,} \end{cases} \tag{4.2}$$

where i represents a voxel in the image domain $\mathcal{P}$, $t = 5\text{mm}$ and $b = 0.1$, as optimally determined via a grid search by the authors. Comparison between other baseline methods

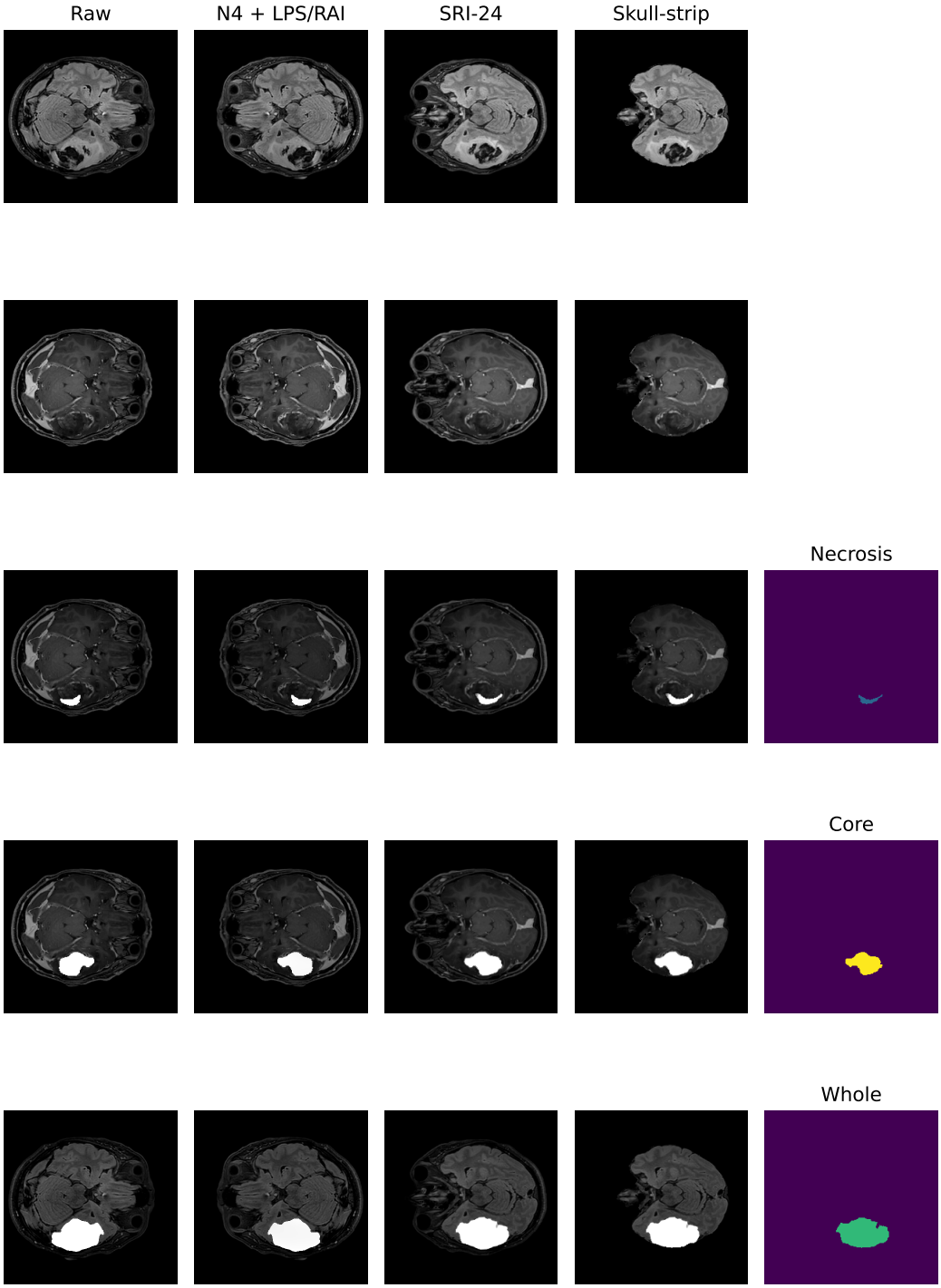


Figure 4.6. BraTS harmonization pipeline for a pre-operative Molinette Hospital patient.

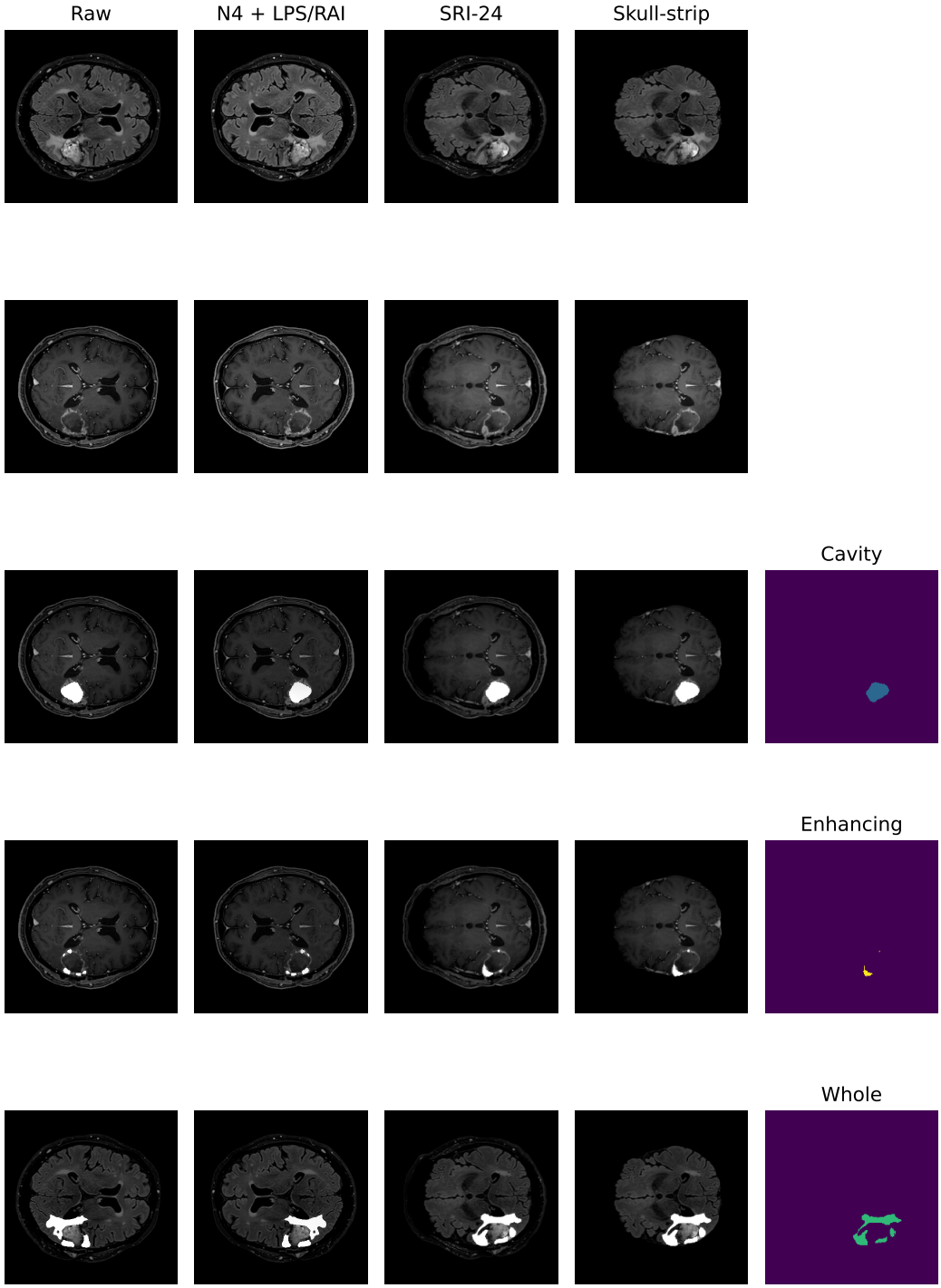


Figure 4.7. BraTS harmonization pipeline for a post-operative Molinette Hospital patient.

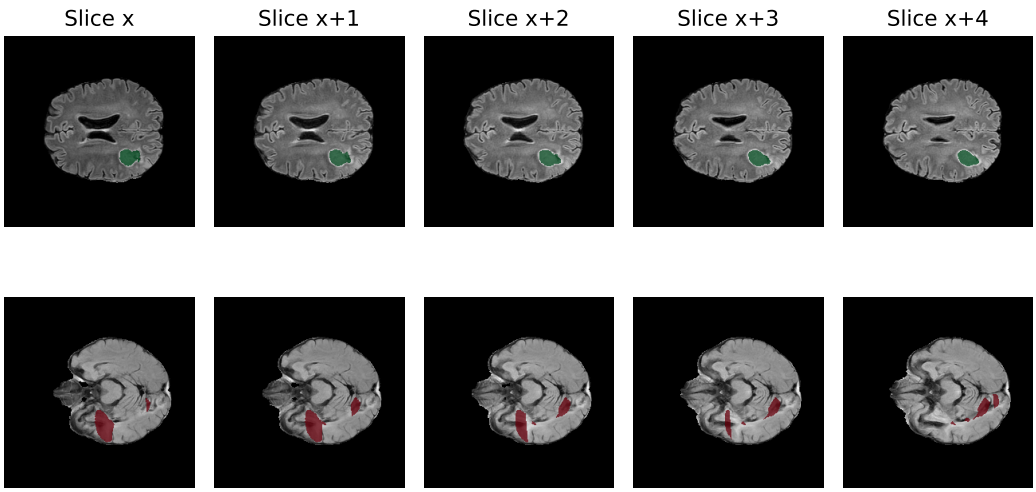


Figure 4.8. Qualitative comparison of processed FLAIR from volumetric raw input (above) and non-volumetric one (below).

shows that SynthStrip significantly outperforms their accuracy for nearly every cohort test, and when this does not happen, SynthStrip matches baseline performances (Hoopes et al. [2022]). Figures 4.6 and 4.7 show the final BraTS harmonization pipeline for the patients in the Molinette Hospital dataset, comprising SRI-24 co-registration performed with CaPTk and skull-stripping by SynthStrip. Since the CaPTk pipeline required the presence of all four scan modalities (FLAIR, T1, T1ce, T2) but for the Molinette dataset only FLAIR and T1ce were present, it was decided to pass the burned-in segmentations as T1 (burned-in T1ce) and T2 (burned-in FLAIR) as it empirically seemed not to loose any information. SynthStrip was then applied to extract the brain mask from the end-of-pipeline original T1ce scan, with such mask being then adopted for all remaining scans. Once both raw and burned-in data have been registered to the SRI-24 atlas and skull-stripped, all classes were extracted by retrieving high-intensity regions and the complete starting segmentation was built by joining the necrosis/cavity (respectively for pre/post-operative scans), the enhancing tumor and the peripheral edema.
Table 4.4 presents an overview on the post-operative part of the dataset made available by the Molinette Hospital, which includes 166 MRI scans from 71 patients. It is worth noticing the peculiar behaviour of vanishing ground truth classes emerging quite frequently: indeed, in some scans, either the resection cavity or the enhancing tumor may not be present at all. This might be due, for example, in the case of resection cavity, to either cerebral parenchyma expanding and filling it or tumor relapse. This might lead eventually to some false positive classification by the network, especially regarding enhancing tumor since it is already trained in recognizing it but its presence in a post-operative scan is way less preponderant than in a pre-operative one. Moreover, another problem arising is the

Number of scans	Number of patients	MRI scan type	Total
1	22	NC: 5 | NE: 7 | C: 10	22 (NV: 4)
2	25	NC: 3 | NE: 14 | C: 33	50 (NV: 9)
3	9	NC: 8 | NE: 8 | C: 11	27 (NV: 12)
4+	15	NC: 12 | NE: 22 | C: 33	67 (NV: 21)
Total	**71**	**NC: 28 | NE: 51 | C: 87**	**166 (NV: 46)**

Table 4.4. Quantitative description of the post-operative dataset made available by the Molinette Hospital (NC: no cavity, NE: no enhancing, C: complete, NV: non-volumetric).

Fold	Train. scans	Train. patients	Val. scans	Val. patients
0	135 (NV: 46)	58 (NV: 26)	26	9
1	146 (NV: 46)	59 (NV: 26)	15	8
2	145 (NV: 46)	59 (NV: 26)	16	8
3	143 (NV: 46)	59 (NV: 26)	18	8
4	147 (NV: 46)	59 (NV: 26)	14	8

Table 4.5. Quantitative description of the 5-fold patient-based split for the Molinette Hospital dataset.

presence of non-volumetric FLAIR scans. These scans present an extremely low resolution – ∼20-30 slices – and the anatomical information obtained from the "upsampling" procedure to interpolate the raw MR into the 155 slices required by the SRI-24 template is not satisfactory, eventually leading to imprecise edema ground truth extrapolation. Figure 4.8 shows qualitatively the depth evolution of a processed FLAIR scan obtained from a volumetric raw input and of a processed FLAIR scan obtained from a non-volumetric one, with the latter being way noisier and less precise.
Non-volumetric scans are often due to old protocols or obsolete acquisitions but their presence in the clinical practice is far from being irrelevant (more than 25% in the Molinette dataset), especially when considering that these patients are likely to be the ones with a larger number of follow-up acquisitions. For such reason, excluding entirely these data from the dataset would not be correct, clinically speaking. The middle ground chosen at last is to include these data only for training, while excluding them from validation. In this way, validation performances are not subjected to this phenomenon while information can hopefully still be extrapolated during training.
With the aim of trying to limit the problem of false positive predictions, being those strongly penalized by the network, segmentation will be performed on three overlapping classes, similarly to what has been done in the pre-operative case. More in detail, the network will segment the union of cavity and whole tumor (i.e. cavity + enhancing tumor + edema), gross tumor volume (cavity + enhancing tumor) and resection cavity, with the latter being therefore the "trickiest" to segment for the reasons presented above.

A 5-fold cross-validation will be performed on the Molinette Hospital dataset in order to test the model performance since it is known that a k-fold cross-validation, maximizing the use of available data, provides a more reliable estimation than a simple holdout method (Li and Doi [2007]). Furthermore, it is good practice (and highly recommended) to split a medical dataset at patient level, i.e. without the eventuality of having the set of scans belonging to a single patient fractionated between training and validation set, in order to understand how well the model generalizes on never seen before individuals. With the current policy on non-volumetric scans, this translates in always keeping patients with at least one such MRI scan in the training set. Having randomly excluded 4 patients for a final visual inspection (NC: 1, NE: 1, C: 3), the 5-fold cross-validation will be performed on those 41 patients not presenting non-volumetric imaging, adding then in the training set those 26 patients who actually do present them. Table 4.5 quantitatively summarizes the patients/scans distribution for the analysis included in this work.

4.3 IMT architecture

The proposed method (see Subsection 4.1.1) requires the presence of all four brain MRI modalities, i.e. FLAIR, T1, T1ce and T2. Indeed, each one of those brings some information which is then exploited by the network to segment the brain tumor. Unfortunately, in clinical practice, missing sequences is a likely occurrence. Eijgelaar et al. [2020] proposed a sparsified training strategy which improved the model performance on incomplete clinical datasets but, even with such adjustment, it was shown that the best performance was achieved only if all sequences were available. Due to the advent of deep learning, intramodality (for example, MRI $\rightarrow$ CT) and intermodality (for example, FLAIR MRI $\rightarrow$ T1ce MRI) image synthesis, i.e. artificially reconstruction of the missing sequence by receiving as input the available ones, is a promising and active field of research in both radiation oncology and radiology fields.. Various network architectures have been proposed for such task in medical imaging during the last years, but three main backbone models established themselves since were found to achieve the best results: autoencoder, U-Net and GAN, with the first starting to lose pace with respect to the others (Wang et al. [2020]). Some studies tackled the problem of brain MRI intramodality synthesis. Yang et al. [2020] proposed a method to perform image modality translation (IMT) by leveraging conditional generative adversarial networks (cGANs), whose generator follows the U-Net shape by adding skip connections between mirrored layers in the encoder-decoder network and whose discriminator is derived from a PatchGAN classifier (see Figure 4.9). Osman and Tamam [2022] implemented instead a U-Net model aimed at learning the non-linear mapping between a source image contrast to a target image contrast. The U-Net architecture implemented by the authors was adapted from the standard architecture (see Section 2.1). The main differences proposed consist in: replacing the sigmoid activation function in the last layer with tanh, adding dropout in order to prevent overfitting, implementing zero-padding before each convolution with the goal of keeping the image size constant and substituting the up-convolutions with simple upsampling layers. Figure 4.10 shows the final architecture as proposed by the autors. BraTS 2018 was used as imaging dataset in order to train the network. Each MR was preprocessed by removing null slices

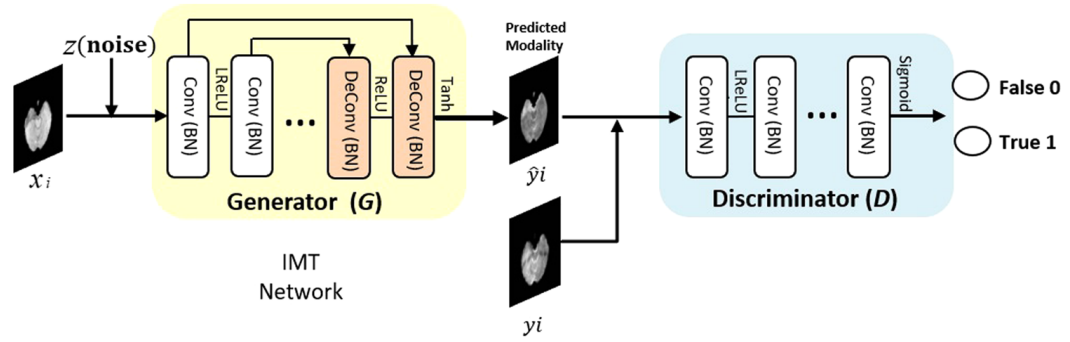


Figure 4.9. cGAN model for IMT as presented in the original paper by Yang et al. [2020].

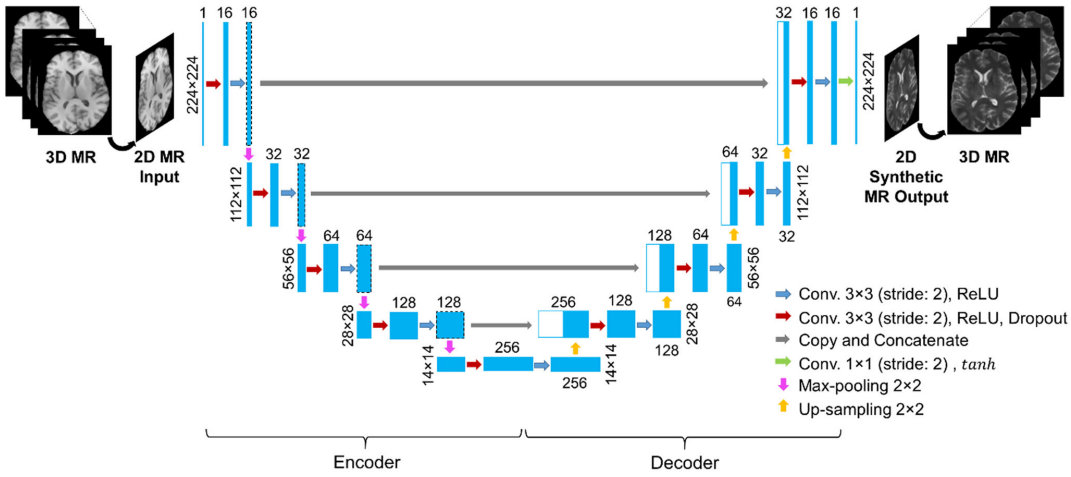


Figure 4.10. U-Net model for IMT as presented in the original paper by Osman and Tamam [2022].

(i.e. empty ones) since not all the slices provided contain anatomical information. Then, MRI intensities were standardized to reduce the negative impact of signal variation across different institution scanners and protocols, and normalized to have zero mean and unit standard deviation. Intensities are then scaled within a [0, 1] range since it was shown to improve the results (Reinhold et al. [2019]). Finally, images were resized to match the U-Net 2D input size of 224×224 pixels by means of cubic interpolation.

Translation	PSNR	MAE	MSE	SSIM
T1 → T2	29.45±1.72	0.0124±0.0027	0.0012±0.0004	0.932±0.023
T2 → T1	29.44±1.85	0.0149±0.0050	0.0012±0.0005	0.937±0.020
T1 → FLAIR	33.25±1.55	0.0086±0.0020	0.0005±0.0002	0.946±0.013
FLAIR → T1	30.73±1.81	0.0125±0.0038	0.0009±0.0004	0.946±0.017
T2 → FLAIR	33.01±1.66	0.0089±0.0027	0.0005±0.0002	0.944±0.015
FLAIR → T2	29.57±1.78	0.0120±0.0036	0.0012±0.0005	0.936±0.022

Table 4.6. Experimental results as presented in Osman and Tamam [2022] showing metric values for all synthesis configuration taken into account.

The U-Net model has been trained from scratch for different image translation configurations (T1 → T2, T2 → T1, T1 → FLAIR, FLAIR → T1, T2 → FLAIR and FLAIR → T2). Adam optimizer with a learning rate of 0.001 was selected to minimize the mean-squared error (MSE) loss until convergence. Training was planned for 120 epochs but an early-stopping technique was set to terminate it if no improvement was achieved during a patience time of 20 epochs.
Four pixel-wise metrics were chosen in order to evaluate the quality of the generated MR images: MAE, MSE, peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM). PSNR measures if the artificial output is a sparsely or evenly distributed prediction by taking into account both the largest intensity value and the MSE, and is defined as

$$\mathrm{PSNR} = 10 * \log_{10}\left(\frac{I_{\max}^2}{\mathrm{MSE}}\right). \tag{4.3}$$

SSIM tries instead to capture the quality of the images as perceived by humans by comparing them. Its formula is defined as

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{4.4}$$

where μ is the mean image intensity, σ^2 is the variance of the image, σ_{xy} is the covariance between ground truth and prediction, while c_1 and c_2 are constants added for division stability.
The authors compared their U-Net results with several other models adopting BraTS to validate their deep learning models for image modality translation. The quantitative comparison of SSIM, PSNR and MAE metrics (see Table 4.6) indicates that such an U-Net model behaves as good as the best-reported results in literature, achieving the best MAE in all synthesis configurations taken into account in the study.

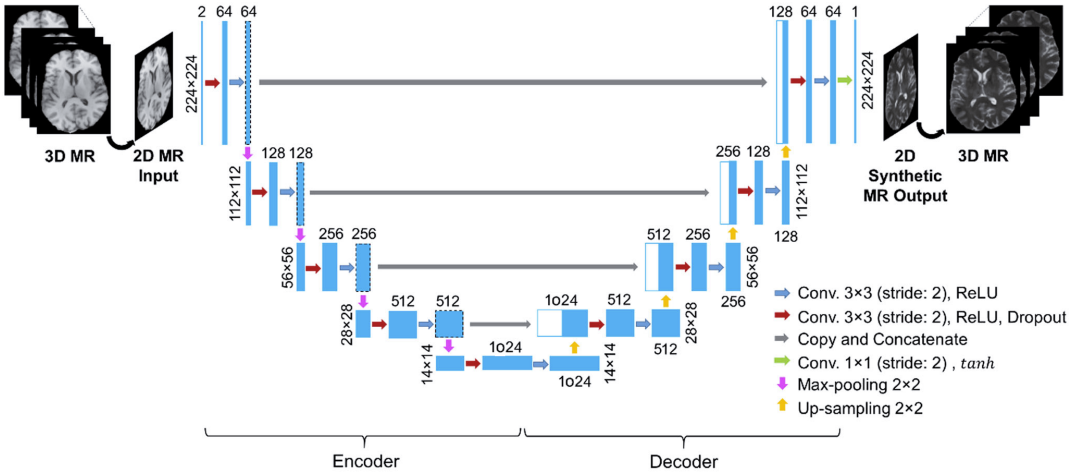


Figure 4.11. Modified version of the U-Net model for IMT originally introduced by Osman and Tamam [2022].

In order to evaluate the effect on post-operative segmentation of the additional information present in the missing two sequences (T1, T2), the model presented by Osman and Tamam [2022] is chosen, expanded and adopted to artificially synthesize the required MRI modalities. More in detail, two are the main architectural modifications implemented: multichannel input layer allowing for multiple modalities being stacked as input with the goal of leveraging more contextual information and following layers presenting four times the number of filters of the original architecture.
Hence, current implementation receives as input both available modalities – i.e. FLAIR and T1ce – and returns as output the desired sequence. Figure 4.11 illustrates the U-Net structure chosen for the IMT task, together with the architectural changes presented above.

Chapter 5

Performance evaluation

The most amazing combinations can result if you shuffle the pack enough.

- Mikhail Bulgakov

5.1 Pre-operative segmentation

As anticipated, the goal of this study is to perform automatic post-operative brain tumor MRI segmentation via transfer learning from pre-operative one. Hence, the first step consists in reconstructing both the architecture and the training schedule in order to achieve comparable results to the ones presented by Futrega et al. [2021] on the BraTS 2021 dataset. Therefore, following the official NVIDIA GitHub repository[1], the nnU-Net variant, the preprocessing and the training schedule as discussed by the authors (see Subsection 4.1.1) have been implemented. Training was performed for 150 epochs while all other relevant hyperparameters have been kept as presented in literature.
Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino (http://www.hpc.polito.it). Tables 5.1 and 5.2 present the quantitative results obtained during a 5-fold cross-validation on the BraTS 2021 dataset for Dice and Hausdorff95 distance metrics respectively. Results are in line with those presented by Futrega et al. [2021] (DSC: 91.63), thus confirming the foundation of such an nnU-Net implementation for pre-operative brain tumor segmentation in magnetic resonance imaging. Evaluation is performed for two different configurations of available scans: the "complete" one, i.e. the one comprising all four modalities (FLAIR, T1, T1ce and T2), and the "most-informative subset" one, i.e. the one comprising just FLAIR and T1ce modalities. Figures 5.1, 5.2 and 5.3 respectively illustrate instead the trend of train and validation losses, dice scores and Hausdorff95 distance scores on the "complete" configuration.

[1] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/Segmentation/nnUNet.

Model	Available modalities	Whole	Core	Enh.	Mean
Fold 0	FLAIR, T1ce	93.59	92.56	85.52	90.56
	FLAIR, T1, T1ce, T2	93.52	92.66	85.86	90.68
Fold 1	FLAIR, T1ce	93.20	92.21	85.58	90.33
	FLAIR, T1, T1ce, T2	93.98	92.41	86.01	90.80
Fold 2	FLAIR, T1ce	93.12	92.31	87.31	90.92
	FLAIR, T1, T1ce, T2	93.70	93.25	87.40	91.45
Fold 3	FLAIR, T1ce	93.71	95.23	87.04	91.99
	FLAIR, T1, T1ce, T2	94.19	94.60	87.44	92.08
Fold 4	FLAIR, T1ce	92.90	92.83	84.50	90.08
	FLAIR, T1, T1ce, T2	93.51	92.62	85.13	90.42
Mean ± Std	FLAIR, T1ce	93.30 ± 0.30	93.03 ± 1.12	85.99 ± 1.04	**90.77 ± 0.67**
	FLAIR, T1, T1ce, T2	93.78 ± 0.27	93.11 ± 0.80	86.37 ± 0.91	**91.09 ± 0.60**

Table 5.1. Experimental Dice scores obtained during 5-fold cross-validation on the BraTS 2021 dataset for the two available modalities configurations.

Model	Available modalities	Whole	Core	Enh.	Mean
Fold 0	FLAIR, T1ce	7.13	6.13	11.74	8.33
	FLAIR, T1, T1ce, T2	6.46	5.29	12.47	8.07
Fold 1	FLAIR, T1ce	5.40	8.32	16.54	10.04
	FLAIR, T1, T1ce, T2	5.10	8.44	15.16	9.57
Fold 2	FLAIR, T1ce	5.31	3.74	10.69	6.58
	FLAIR, T1, T1ce, T2	4.67	3.66	10.55	6.29
Fold 3	FLAIR, T1ce	5.25	4.94	16.11	8.77
	FLAIR, T1, T1ce, T2	5.60	5.52	16.23	9.12
Fold 4	FLAIR, T1ce	5.64	5.79	12.67	8.03
	FLAIR, T1, T1ce, T2	5.44	5.90	12.83	8.06
Mean ± Std	FLAIR, T1ce	5.75 ± 0.70	5.77 ± 1.51	13.54 ± 2.35	**8.35 ± 1.12**
	FLAIR, T1, T1ce, T2	5.45 ± 0.60	5.76 ± 1.54	13.45 ± 2.02	**8.22 ± 1.13**

Table 5.2. Experimental Hausdorff95 distance scores obtained during 5-fold cross-validation on the BraTS 2021 dataset for the two available modalities configurations.

Results show that, if it is true, as known, that the optimal performance is obtained on the "complete" configuration, it also true that solid outcomes are achieved by only adopting the two most informative MRI scans. This is valuable since, in clinical practice, the availability of all four modalities is not granted.

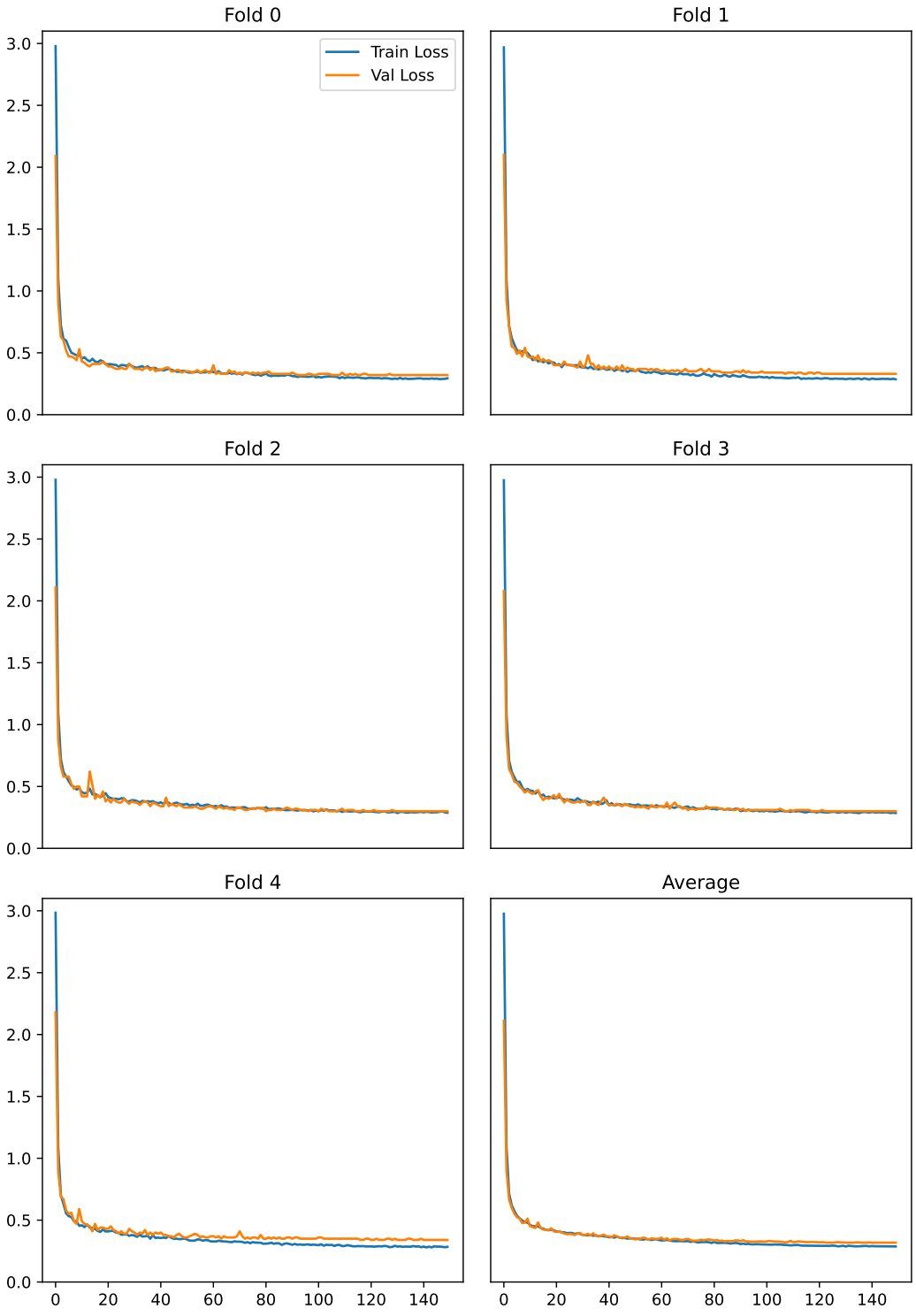


Figure 5.1. Train and validation losses for 150 epochs on the BraTS 2021 dataset.

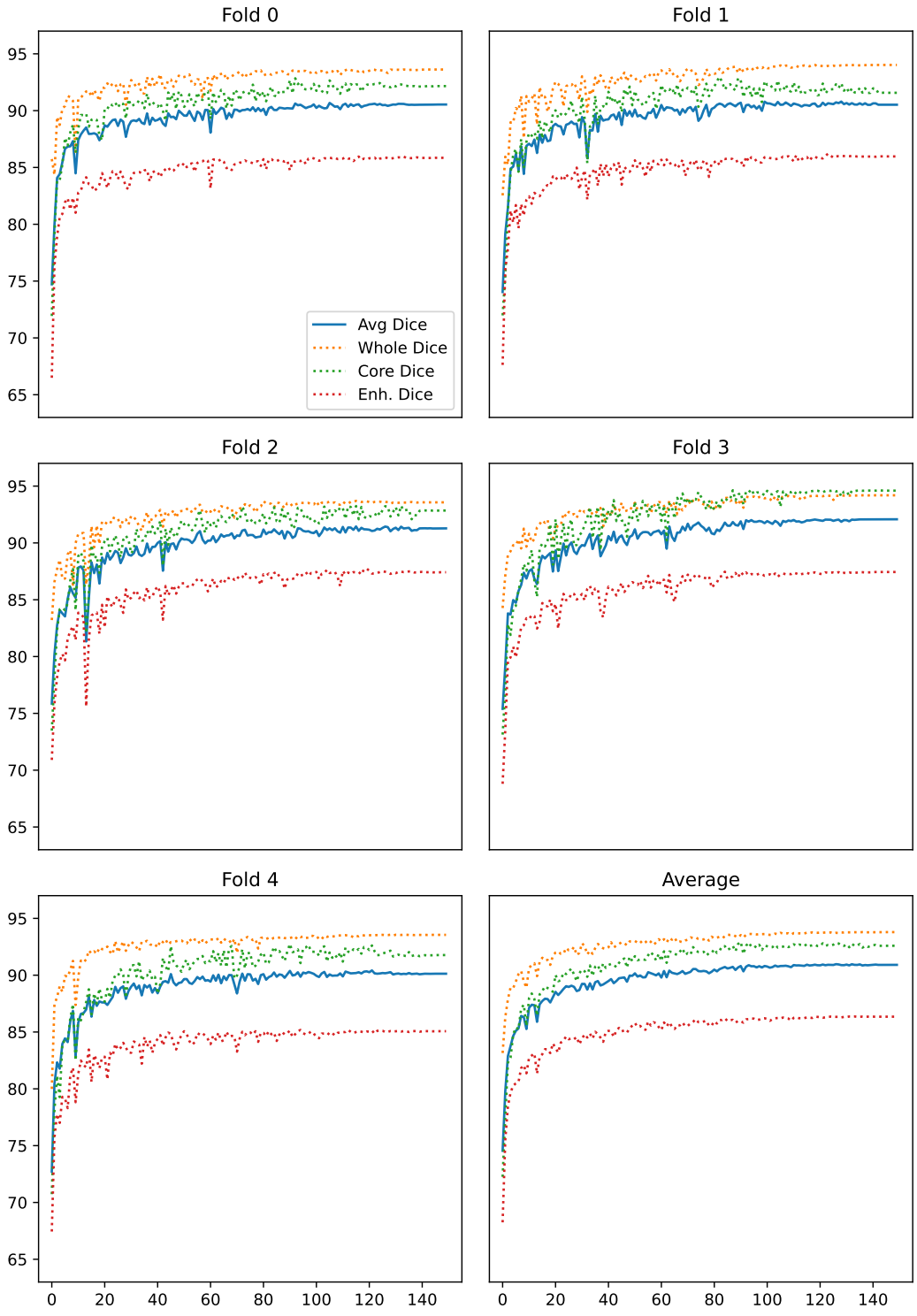


Figure 5.2. Dice scores for 150 epochs on the BraTS 2021 dataset.

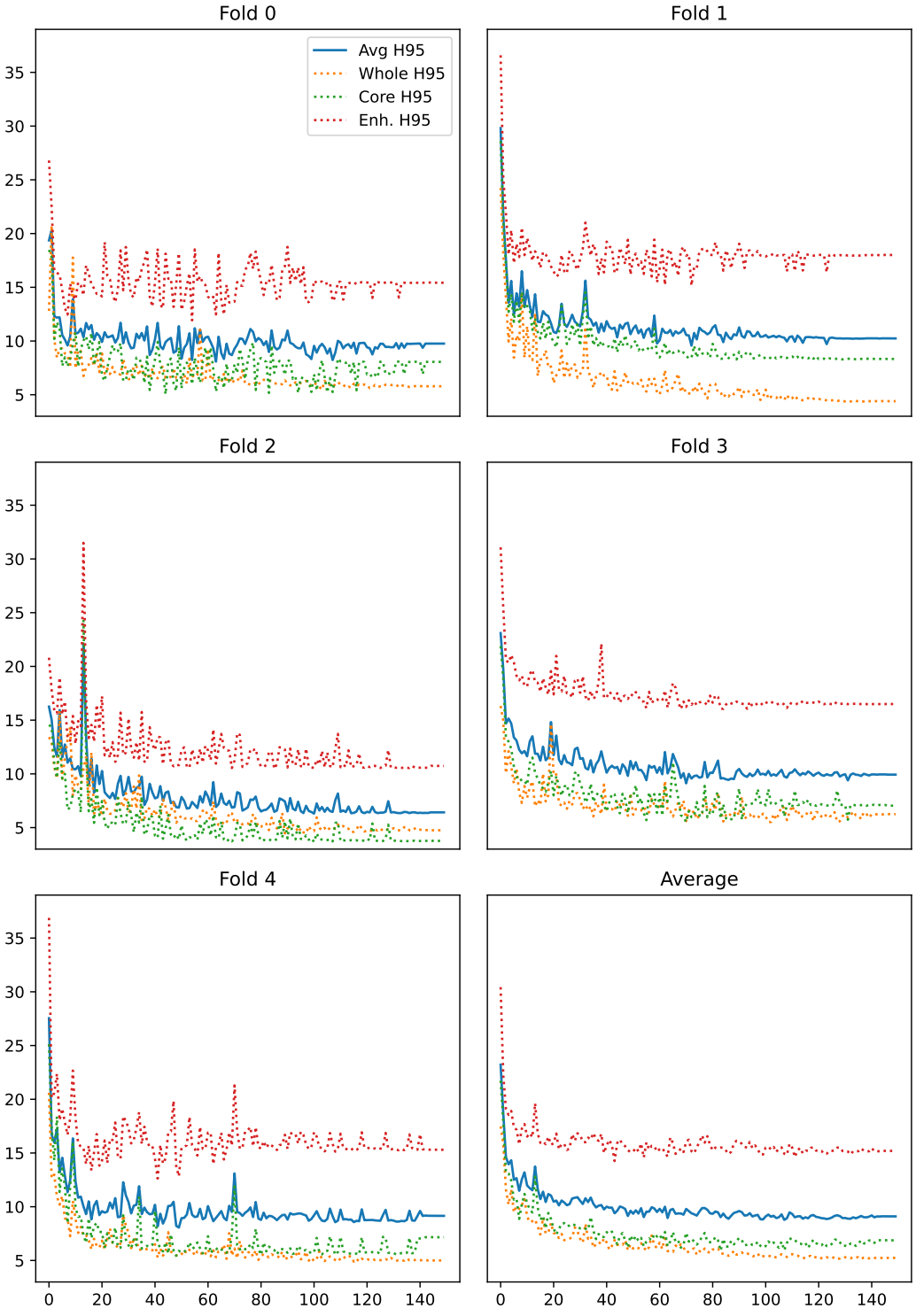


Figure 5.3. Hausdorff95 scores for 150 epochs on the BraTS 2021 dataset.

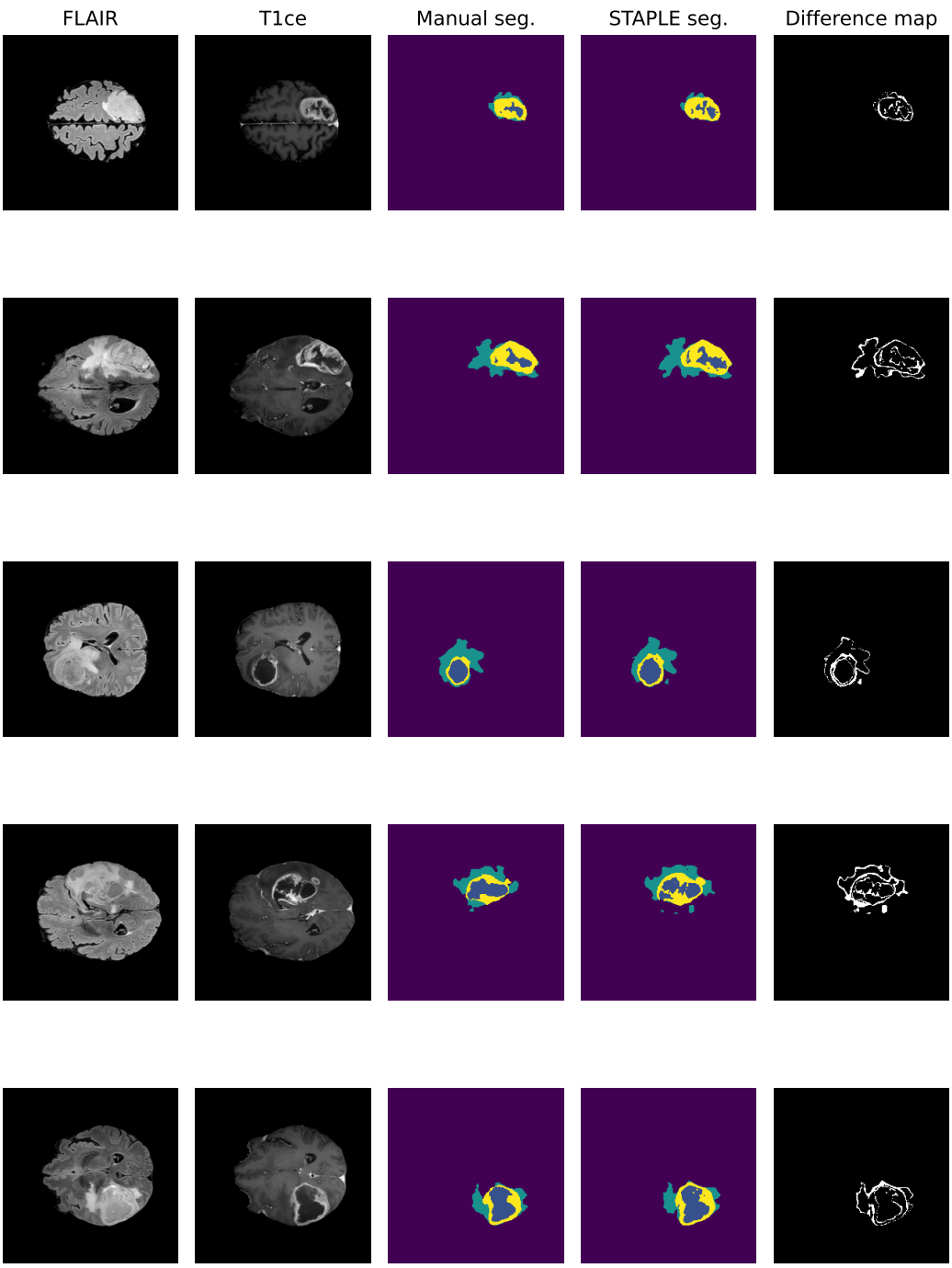


Figure 5.4. STAPLE pre-operative segmentation on patients from Molinette institute.

Figure 5.4 shows qualitatively the pre-operative GBM segmentation of patients from the Molinette dataset obtained fusing with the STAPLE algorithm (Warfield et al. [2004]) all outputs retrieved from the application of the five folds trained on the BraTS 2021 dataset. It is interesting to notice how STAPLE outperforms the noisy segmentation acquired as ground truth after the whole data harmonization pipeline from the radiologists' one, being able to segment each class with an higher precision and recognize smaller portions which are lost as reference because of interpolation.

5.2 Intermodality synthesis

Given the absence of both T1 and T2 MRI modalities, the modified U-Net structure proposed in its original form by Osman and Tamam [2022] and presented in Section 4.3 is implemented in order to artificially synthesize them and leverage such further information. The network is trained for 50 epochs with mixed precision on the BraTS 2021 dataset. The underlying U-Net requires 2D inputs, therefore each iteration a 3D scan is taken and a random selection of 64 2D slices (possibly flipped along the two dimensions) is extracted and fed as batch. The network is trained with the default Adam optimizer with a learning rate of 0.0003 which decays following a cosine schedule. Output images have shape 224×224 so post-processing is applied to resize, re-orient and pad them in their BraTS form. Gaussian sharpening is also applied since it increases slightly SSIM performances. Even if the model was extended to admit more than one modality as input, actual syntheses were carried out starting with only the T1ce MRI acquisition since the presence of non-volumetric FLAIR scans would lead to noisy and coarse-grained artificial outputs. Table 5.3 summarized the results obtained with the two synthesis configurations (T1ce → T1, T1ce → T2) in terms of mean-squared error, mean-absolute error, peak signal-to-noise ratio and structural similarity index. Results are slightly worse than the ones reported by the authors, which is due to both having the T1 contrast-enhanced modality as single input and the increasing in acquisition protocols heterogeneity from BraTS 2018 to BraTS 2021. Figure 5.5 and 5.7 present qualitatively the comparison between real and synthesized T1 (and T2) MRI modalities for patients from the BraTS 2021 validation dataset. The slight difference in visual perception is due to the [0, 1]-scaling applied as pre-processing step which modifies the voxel intensity distribution. This difference will be eventually mitigated by the segmentation pre-processing pipeline first, where all voxels are normalized, and by the network learning abilities then. Figure **??** illustrates instead the application of the IMT U-Net architecture to some patients from the Molinette Hospital dataset, being that the final goal of this process.

Synthesis	MSE (↓)	MAE (↓)	PSNR (↑)	SSIM (↑)
T1ce → T1	0.0045	0.0329	23.42	0.87
T1ce → T2	0.0025	0.0225	25.99	0.87

Table 5.3. Experimental metrics results obtained for the two synthesis configuration.

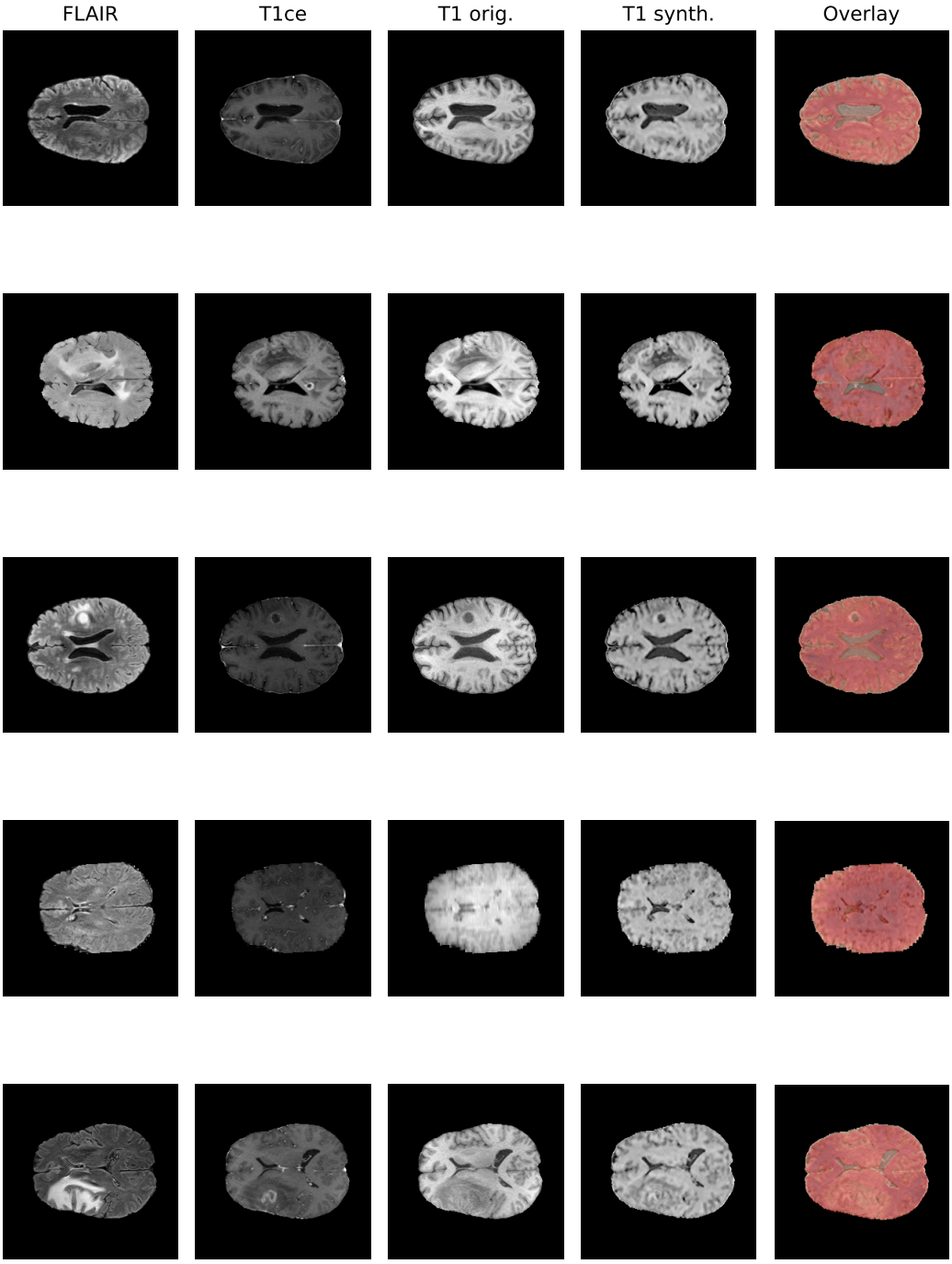


Figure 5.5. Qualitative comparison of synthesized T1 scans from BraTS 2021 patients.

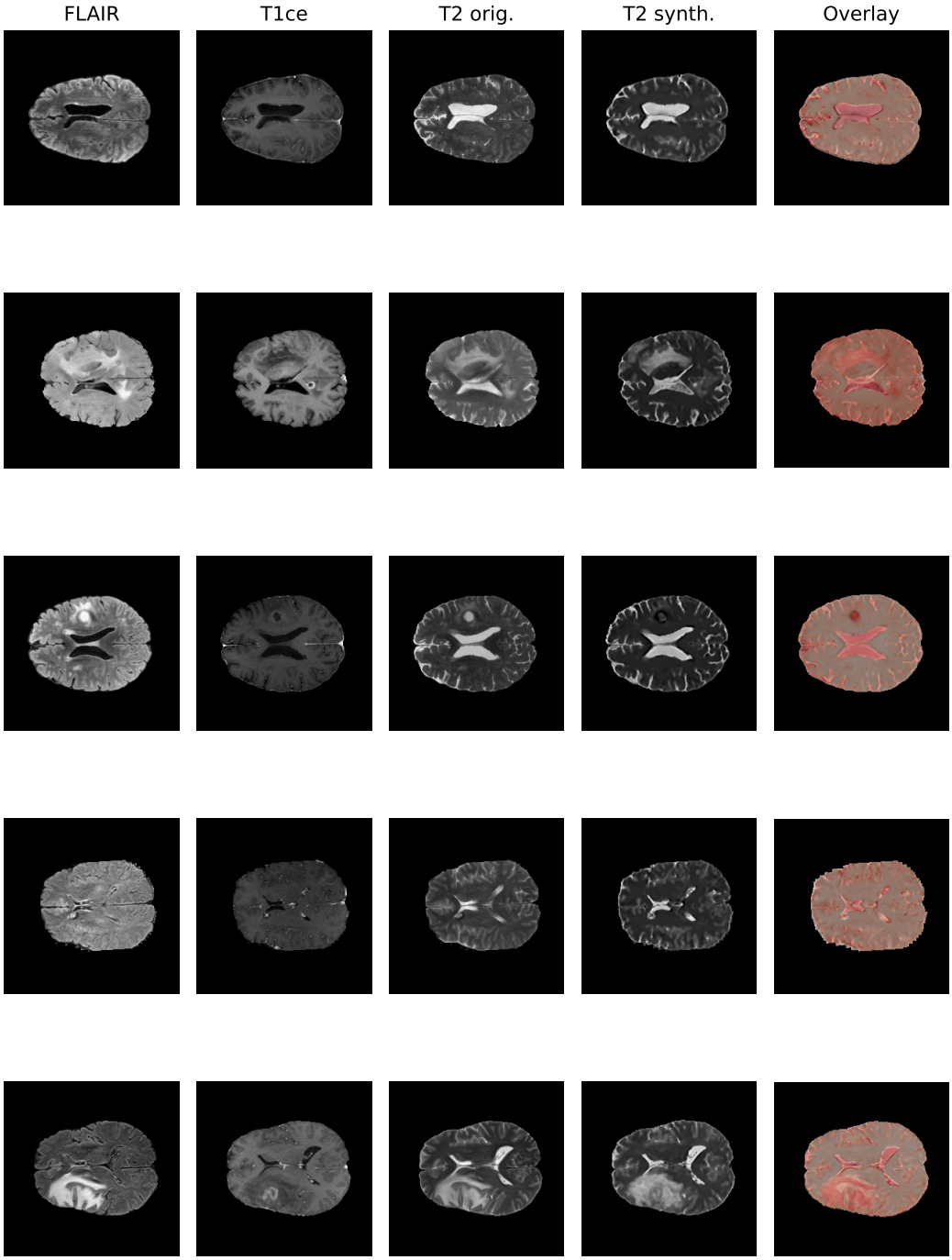


Figure 5.6. Qualitative comparison of synthesized T2 scans from BraTS 2021 patients.

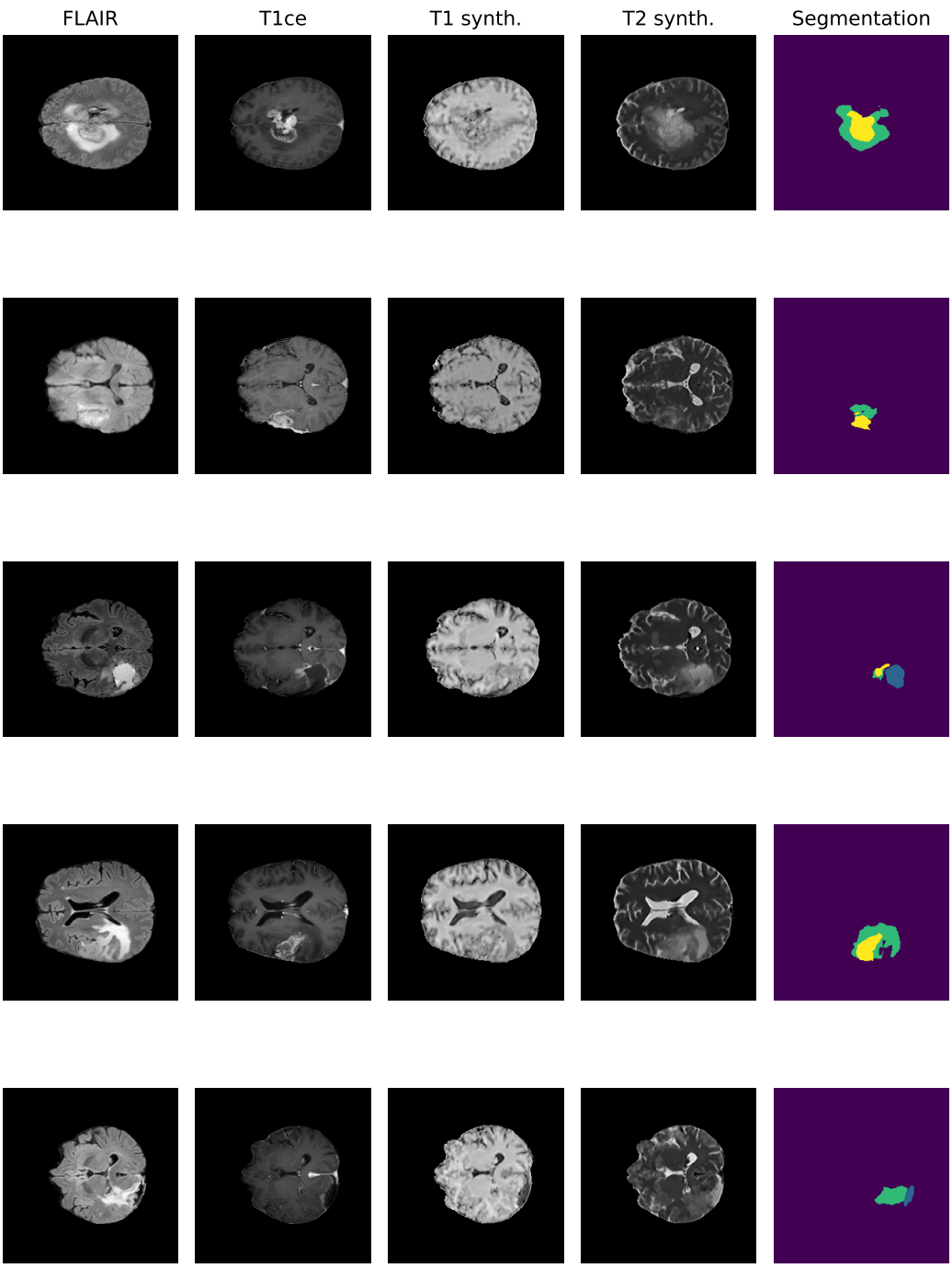


Figure 5.7. IMT application on post-operative scans from the Molinette Hospital dataset.

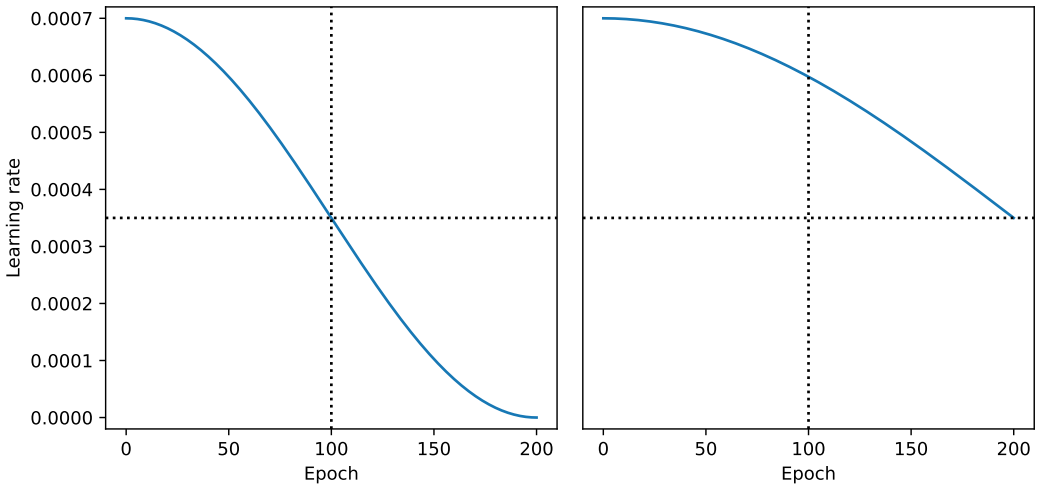


Figure 5.8. Standard learning rate cosine decay schedule (cycles: 0.5, on the left) against the implemented one (cycles: 0.25, on the right).

5.3 Post-operative segmentation

The architecture proposed by Futrega et al. [2021], pre-trained on pre-operative brain tumor segmentation, is therefore extended to allow for transfer learning applicability. Both TL philosophies – i.e. feature extractor and fine-tuning (see Section 2.2) – have been investigated. Specifically, for the former, all configurations with frozen layers in depth levels 1–7 have been studied but no significant improvement or relevant advancement was observed. For such reason, the latter strategy, i.e. tuning all weights from the pre-trained model, is chosen for the actual experiments.
Here again, both configurations (the "complete" one comprising all four MRI modalities and the "most-informative subset" one comprising only FLAIR and T1ce) have been explored, including as T1 and T2 the ones artificially synthesized through the 2D U-Net IMT method described above. Training is performed for 200 epochs with a learning rate of 0.0007 which decays following a cosine schedule having 0.25 as cosine cycles parameter (see Figure 5.8). The other relevant hyperparameters have been kept as in the pre-operative case. A more aggressive data augmentation strategy has been adopted by increasing the probability of applying a given transformation from 0.15 to 0.5 because it is shown to improve the model's generalizability if trained on a small dataset (Zhang et al. [2020a]).
Tables 5.4 and 5.5 present the quantitative results obtained during a 5-fold cross-validation on the Molinette Hospital dataset for Dice and Hausdorff95 distance metrics respectively. What emerges is the high volatility in metric results – distinctly regarding the resection cavity segmentation – due to the limited size of the dataset. Indeed, precisely identifying the cavity is often arduous even for the naked eye of an expert, which translates in either the network oversegmenting the region by including hypointense areas or missing some parts by undersegmenting it.

Model	Available modalities	All	GTV	Cav.	Mean
Fold 0	FLAIR, T1ce	74.45	74.52	69.77	72.91
	FLAIR, T1, T1ce, T2	73.07	72.70	71.23	72.33
Fold 1	FLAIR, T1ce	77.80	74.89	54.78	69.16
	FLAIR, T1, T1ce, T2	76.62	82.39	50.05	69.69
Fold 2	FLAIR, T1ce	76.10	71.41	55.93	67.81
	FLAIR, T1, T1ce, T2	75.07	75.48	55.85	68.80
Fold 3	FLAIR, T1ce	77.16	81.23	71.82	76.74
	FLAIR, T1, T1ce, T2	76.86	81.38	71.96	76.73
Fold 4	FLAIR, T1ce	77.02	80.43	69.31	75.59
	FLAIR, T1, T1ce, T2	73.93	79.58	68.50	74.00
Mean ± Std	FLAIR, T1ce	76.51 ± 1.16	76.50 ± 3.75	64.32 ± 7.38	**72.44 ± 3.49**
	FLAIR, T1, T1ce, T2	75.11 ± 1.48	78.31 ± 3.67	63.52 ± 8.90	**72.31 ± 2.88**

Table 5.4. Experimental Dice scores obtained during 5-fold cross-validation on the Molinette Hospital dataset for the two available modalities configurations.

Model	Available modalities	All	GTV	Cav.	Mean
Fold 0	FLAIR, T1ce	15.20	9.96	8.78	11.31
	FLAIR, T1, T1ce, T2	14.77	10.70	12.07	12.51
Fold 1	FLAIR, T1ce	8.09	16.36	59.18	27.88
	FLAIR, T1, T1ce, T2	7.99	15.87	57.70	27.19
Fold 2	FLAIR, T1ce	11.40	11.60	75.23	32.74
	FLAIR, T1, T1ce, T2	13.71	12.55	75.98	34.08
Fold 3	FLAIR, T1ce	7.83	7.03	47.00	20.62
	FLAIR, T1, T1ce, T2	7.83	6.59	46.30	20.24
Fold 4	FLAIR, T1ce	9.34	6.56	57.94	24.61
	FLAIR, T1, T1ce, T2	10.00	6.96	58.70	25.22
Mean ± Std	FLAIR, T1ce	10.37 ± 2.72	10.30 ± 3.56	49.63 ± 22.32	**23.43 ± 7.24**
	FLAIR, T1, T1ce, T2	10.86 ± 2.88	10.53 ± 3.49	50.15 ± 21.27	**23.85 ± 7.20**

Table 5.5. Experimental Hausdorff95 distance scores obtained during 5-fold cross-validation on the Molinette Hospital dataset for the two available modalities configurations.

The triplets including Figures 5.9, 5.10, 5.11 and Figures 5.12, 5.13, 5.14 illustrate instead the trend of train and validation losses, dice scores and Hausdorff95 distance scores on the "most informative" and "complete" configurations, respectively.

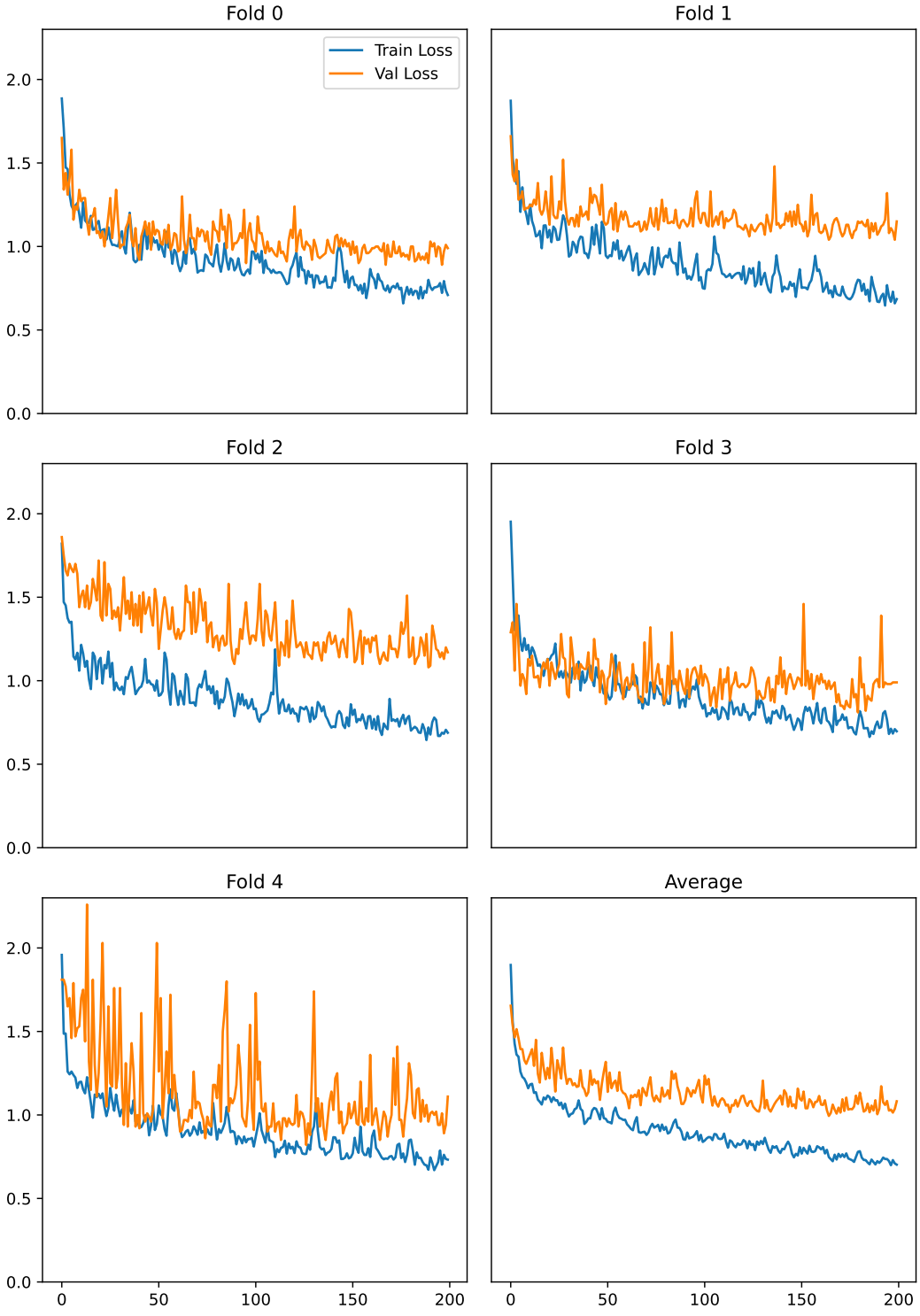


Figure 5.9. Train and validation losses for the reduced configuration on the Molinette Hospital dataset.

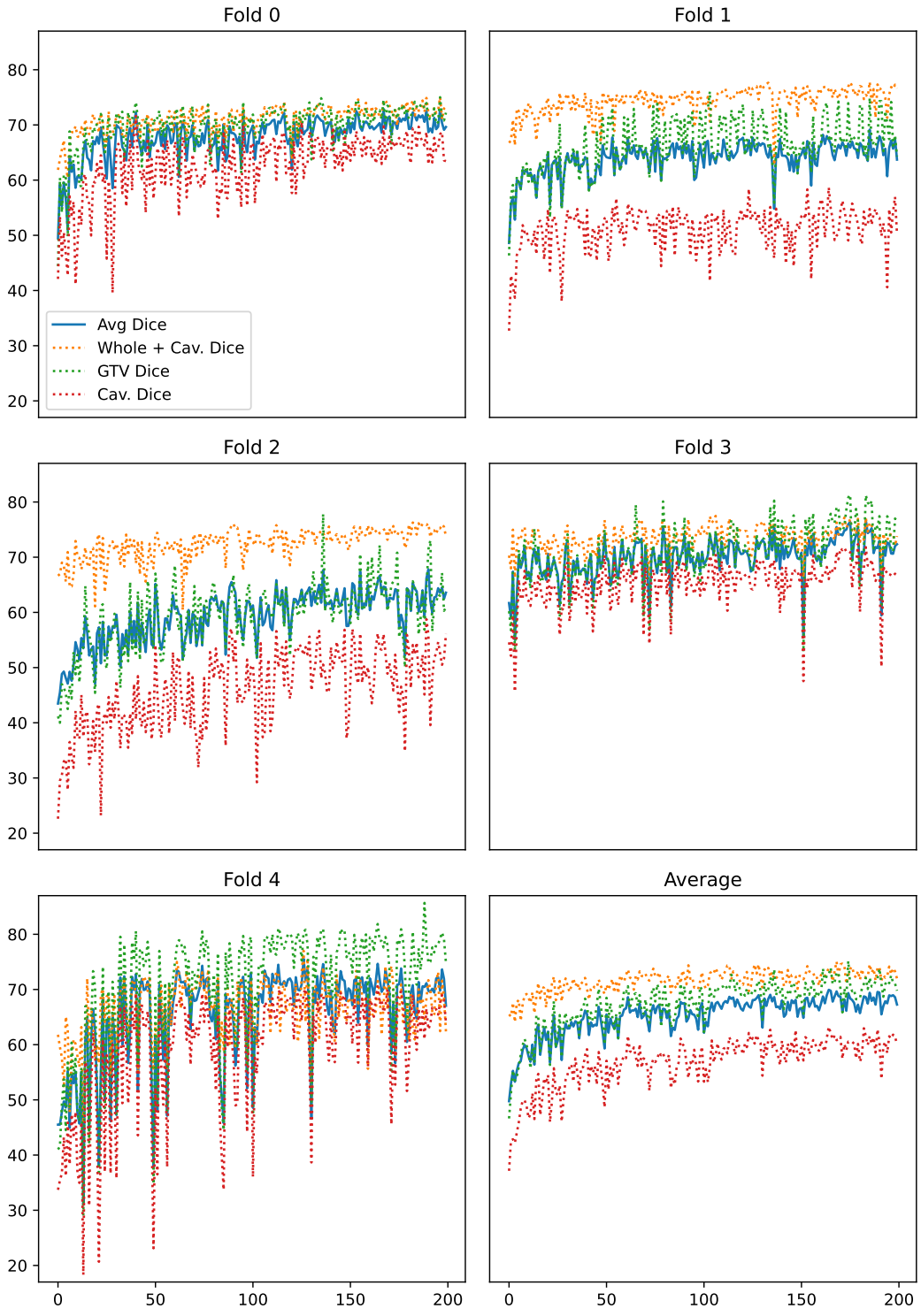


Figure 5.10. Dice scores for the reduced configuration on the Molinette Hospital dataset.

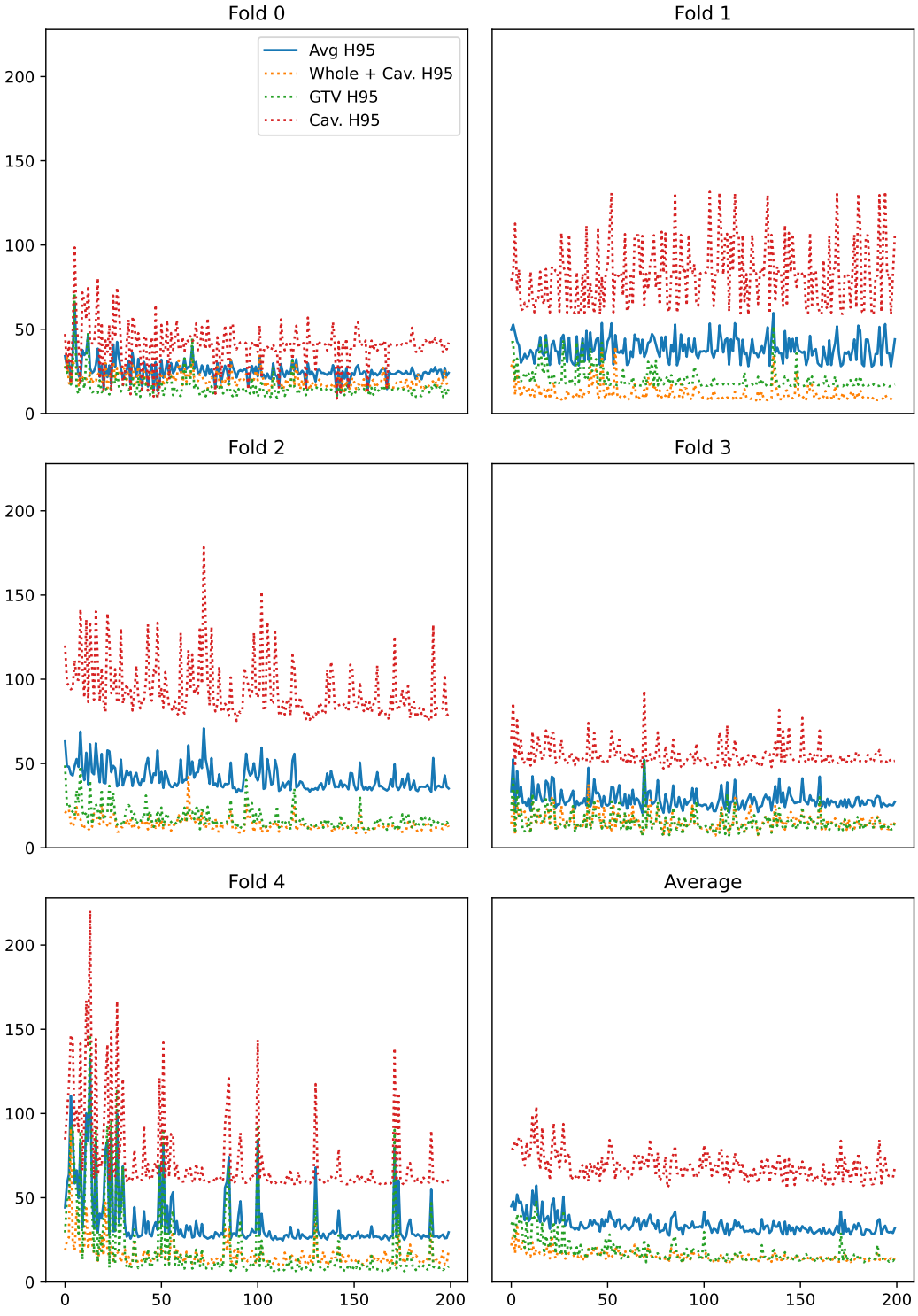


Figure 5.11. Hausdorff95 scores for the reduced configuration on the Molinette Hospital dataset.

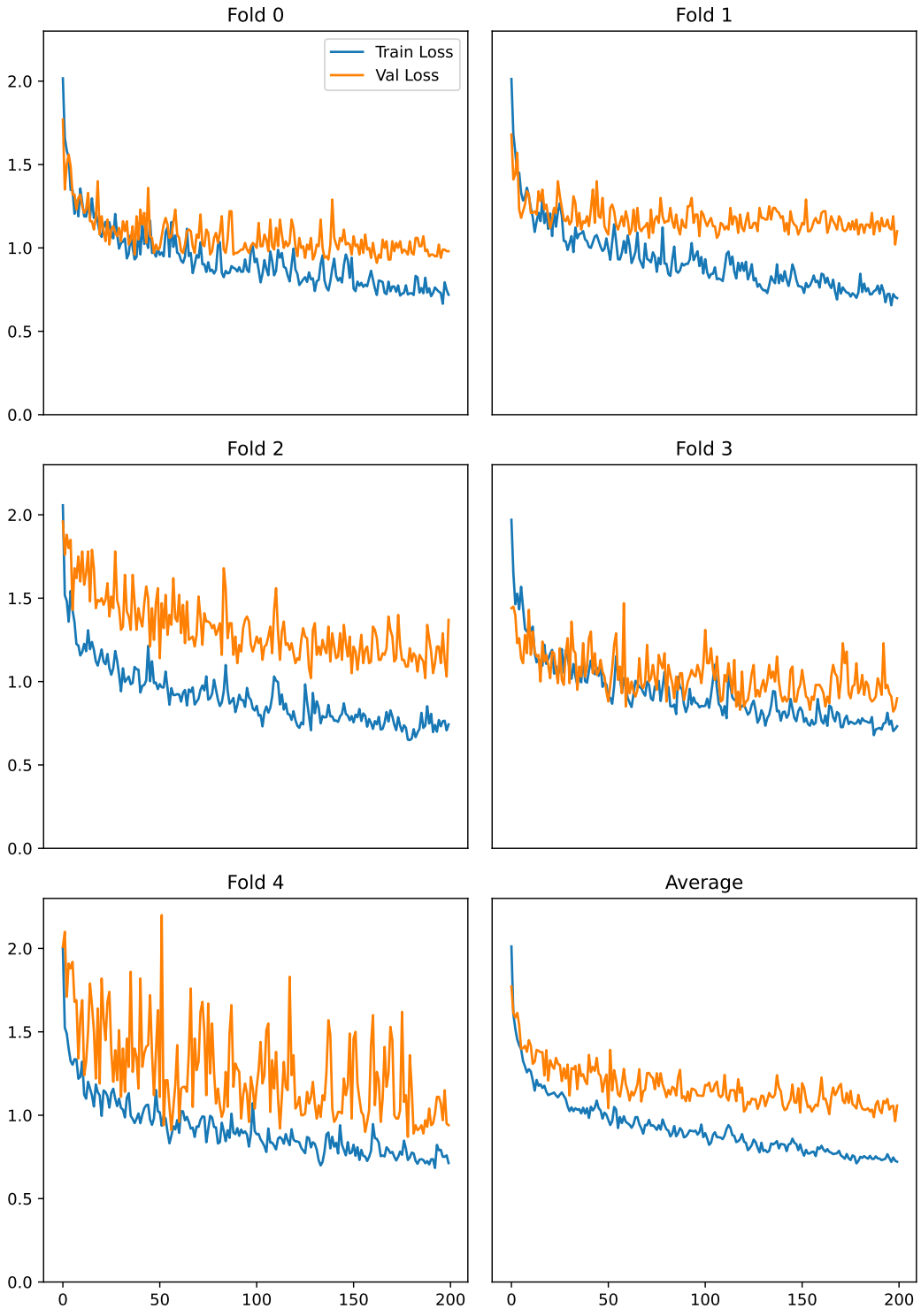


Figure 5.12. Train and validation losses for the complete configuration on the Molinette Hospital dataset.

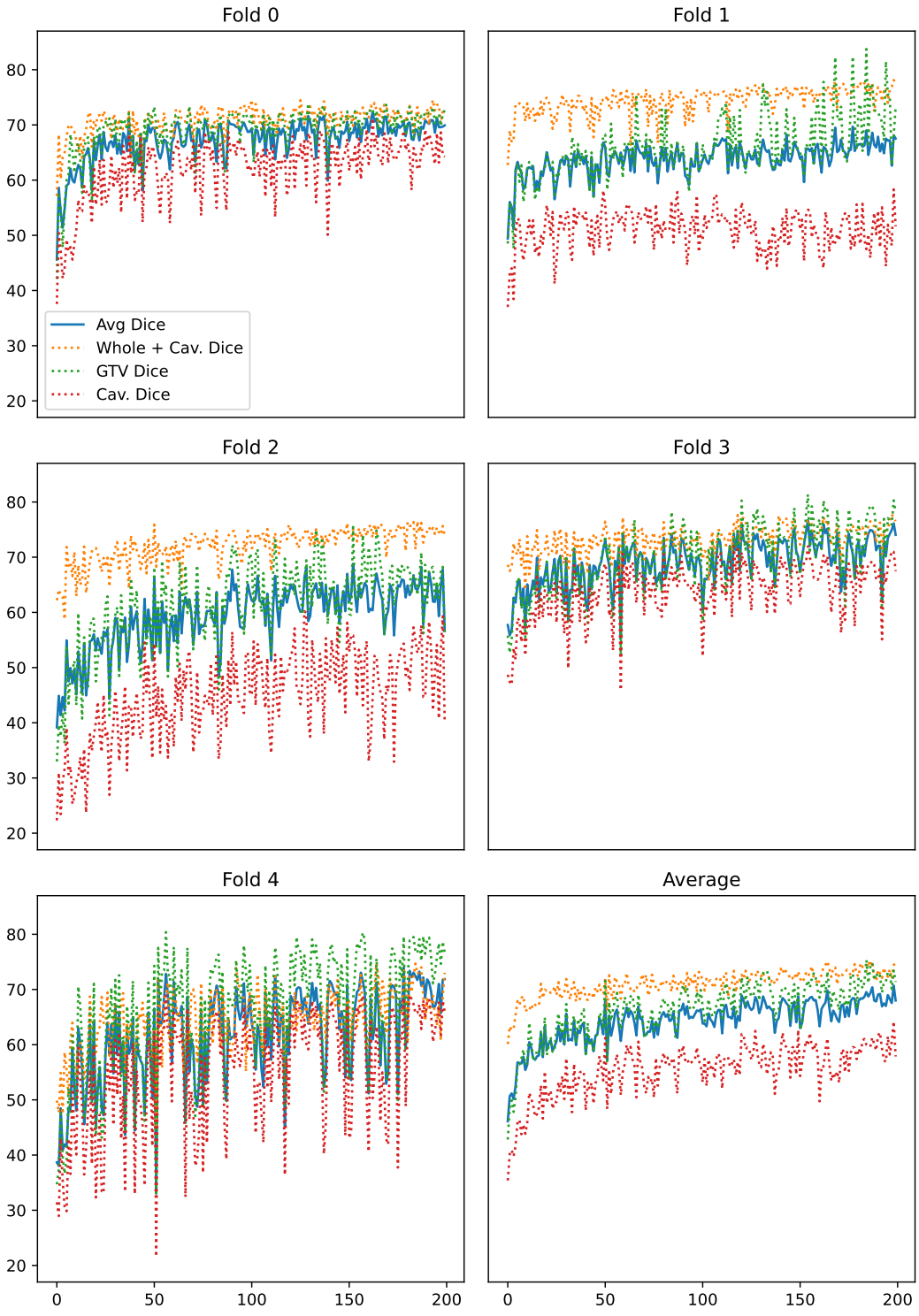


Figure 5.13. Dice scores for the complete configuration on the Molinette Hospital dataset.

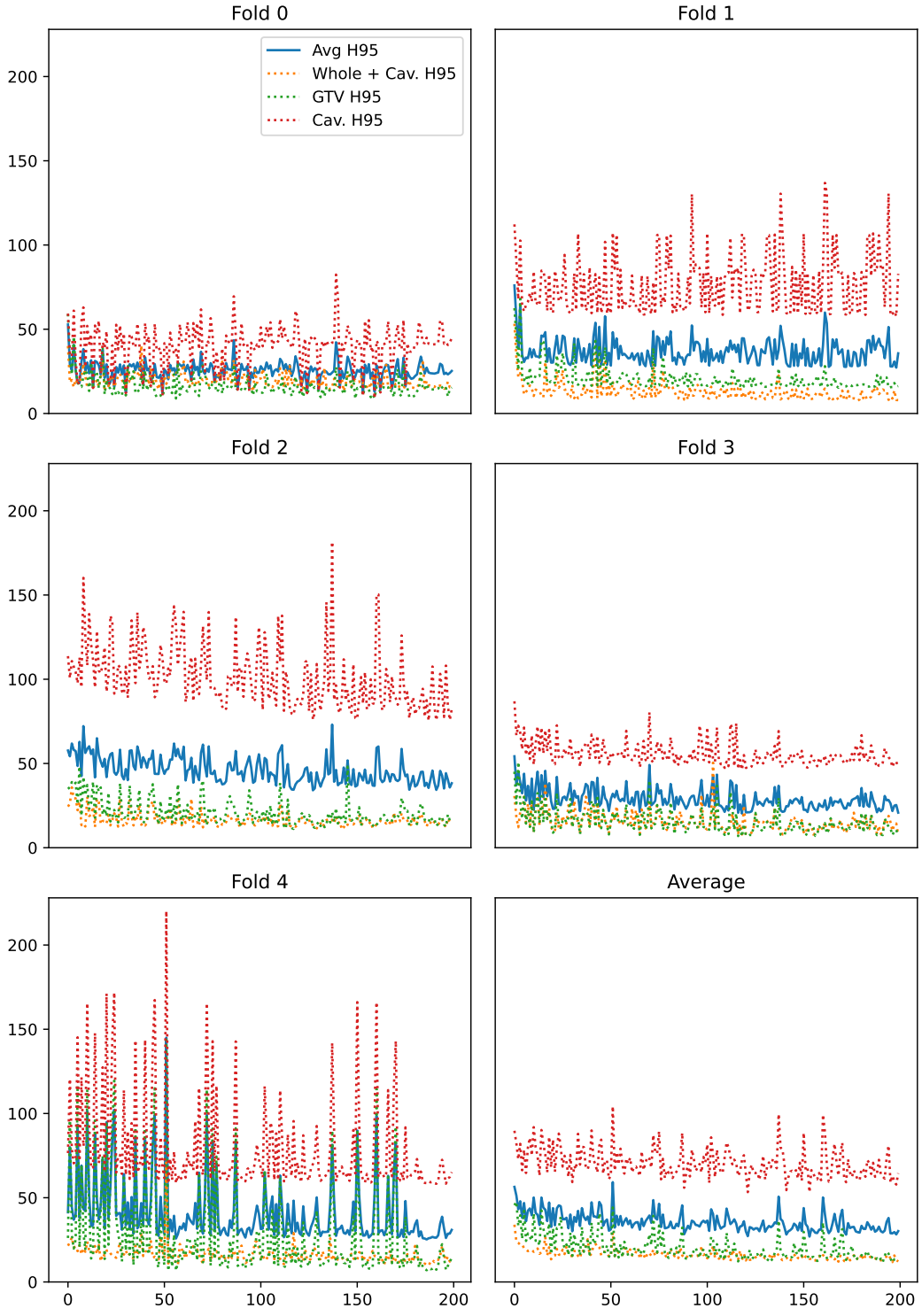


Figure 5.14. Hausdorff95 scores for the complete configuration on the Molinette Hospital dataset.

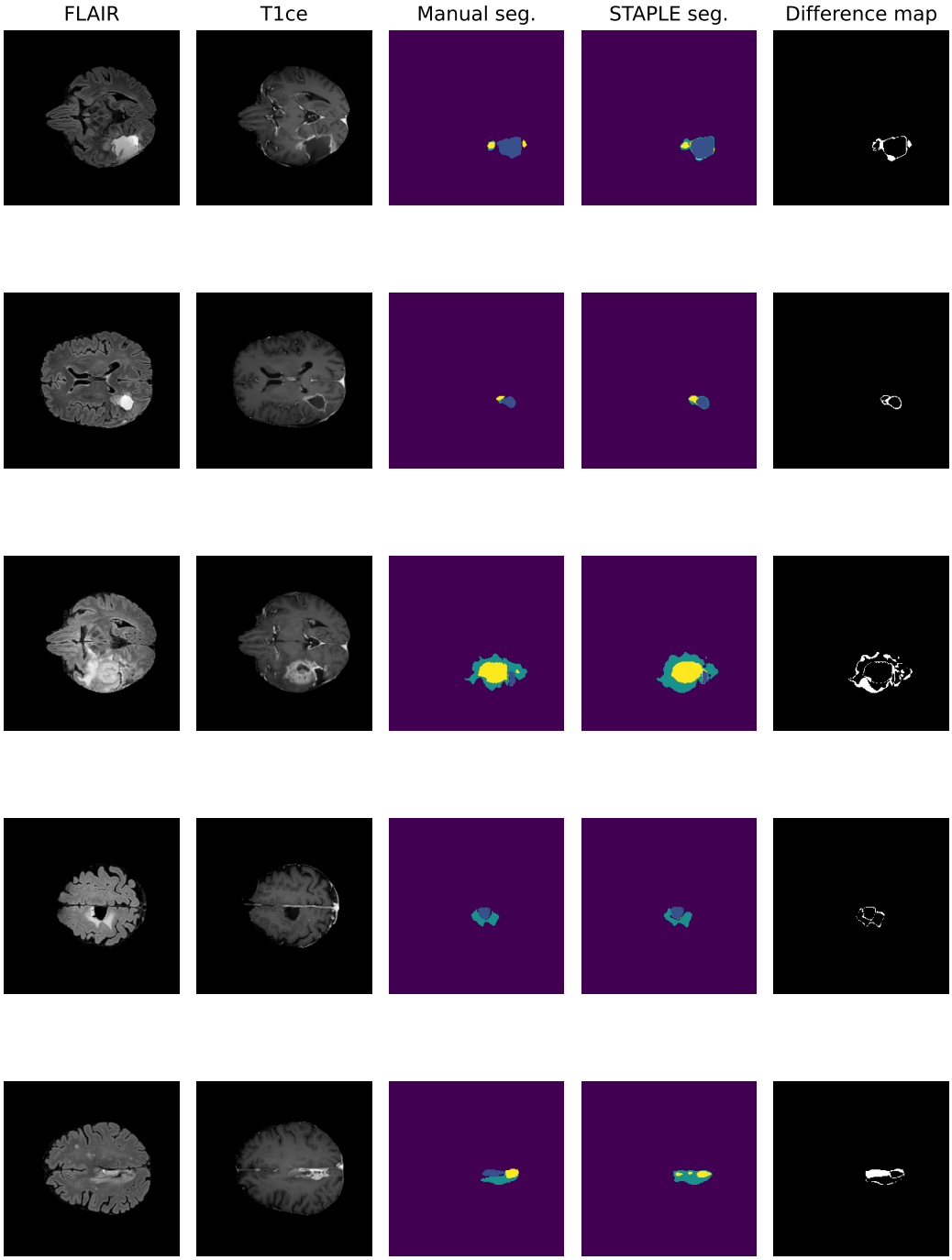


Figure 5.15. STAPLE post-operative segmentation on patients from Molinette institute.

Figure 5.15 shows some example of post-operative segmentation on those 4 patients randomly extracted from the dataset for visual inspection. Similarly to the pre-operative case, in order to bring back the three overlapping regions to the default ones of edema, enhancing tumor and resection cavity, some post-processing is applied. Specifically, if the probability for a voxel of being classified as either GTV or edema is ≤ 0.45, it is classified as background. Instead, if the probability of being classified as GTV is ≤ 0.4, then the voxel is classified as edema. If, on the contrary, the probability of being a GTV voxel is > 0.4 and the same holds for the probability of being classified as cavity, then the voxel is identified as such (or simply enhancing tumor if the latter is not true). These values are determined via a gridsearch on the 5 folds, starting from the parameters proposed by the authors for the pre-operative case. Moreover, similarly to what has been proposed by the authors, in order to eliminate irrelevant segmentations, any connected component identified as resection cavity smaller than 16 voxels with an overall probability smaller than 0.9 is ignored and classified as enhancing tumor instead.

Chapter 6

Discussion and conclusions

A conclusion is simply the place where you got tired of thinking.

- Dan Chaon

6.1 Final remarks

If it was expected that the vast heterogenity that characterizes post-operative segmentation (see Figure 6.1) would affect negatively the performances of the model, a slightly further downgrading is observable for resection cavity recognition due to the extent of MRI scans admitted in the dataset, for reasons ranging from including non-volumetric magnetic resonances to considering a temporal range of acquisitions that goes beyond 12 months. Results are nontheless promising, when considering that this study is the first omni-comprehensive analysis on post-operative GBM segmentation without any kind of limitation regarding MRI scans inclusion. Indeed, current literature regarding post-operative brain tumor segmentation often imposes strong criteria for inclusion/exclusion of MRI scans in the final dataset such as, among others, newly-diagnosed GBM only, availability of all imaging modalities or resection cavity clearly present on visual inspection. These constraints undoubtedly ease the learning process but are not generalizable to the clinical practice where, as it has been shown, an incommensurable diverseness and variety is present. Oversegmentations of resection cavities are however mitigated by the STAPLE fusion, often able to converge towards a reasonable and less noisy segmentation. Figure 6.2, as an example, illustrates how the STAPLE convergence is able to forget the oversegmentation of an hypointense region as resection cavity. However, little to no help comes from it in the case of classes missegmentation. The data scarcity problem makes it impossible for the network to learn thoroughly how to segment each class and "outlier patients", i.e. those patients that present an out-of-the-ordinary behaviour with respects to the ones present in the dataset, are wrongly segmented by a network still unable to recognize such new patterns.

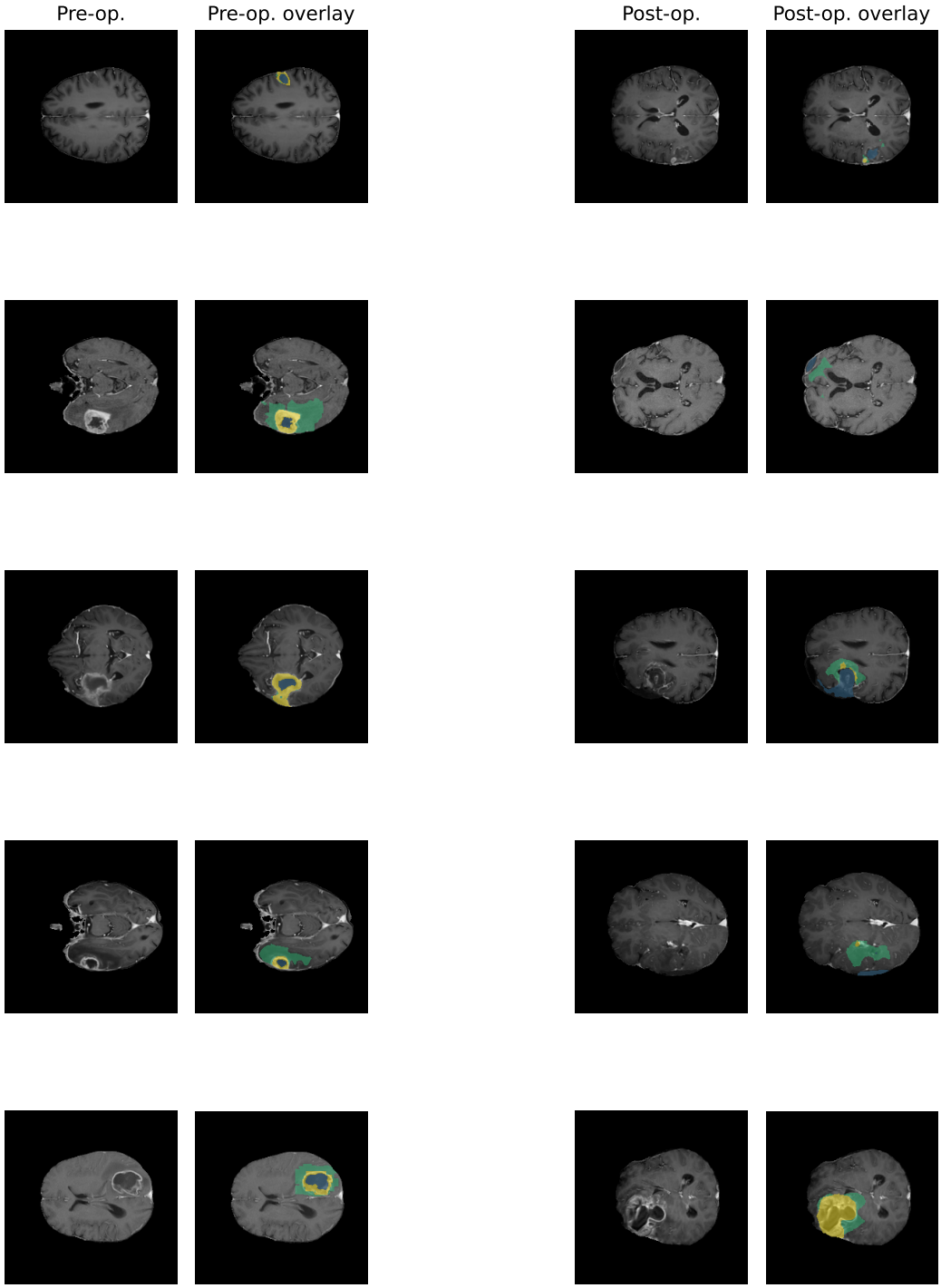


Figure 6.1. Visual summary of pre- and post-operative segmentation.

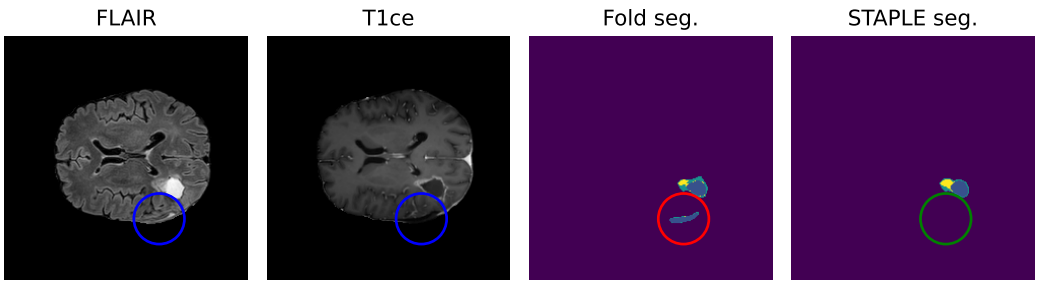


Figure 6.2. Positive effect of STAPLE fusion for resection cavity segmentation.

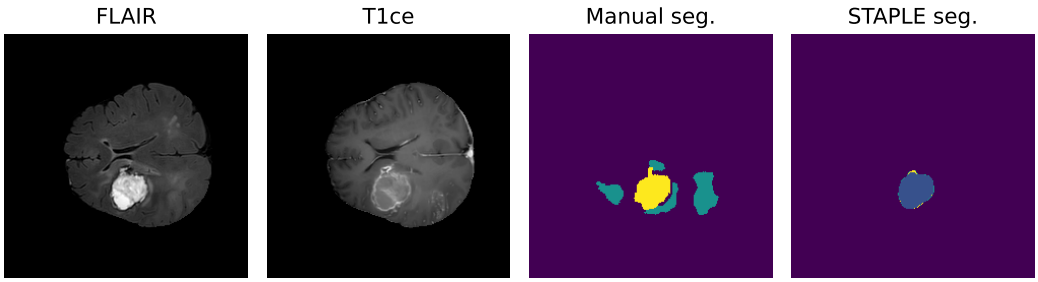


Figure 6.3. Worst-case scenario of post-operative classes missegmentation.

Figure 6.3 illustrates a worst-case scenario, with a patient presenting the enhancing tumor as an extremely hyperintense region (likely due to hemoglobine degradation) which is mistakenly classified by the network as resection cavity, with peripheral edema not being recognized at all.
Still, when considering its strongly diversified nature and the reduced dimension of the dataset, results are in line with the literature. While looking forward to a publicly available multi-site dataset like BraTS, this work presents itself as an interesting proposal and starting point in order to reduce the gap between pre- and post-operative GBM segmentation.

6.2 Limitations

Due to the absence of a publicly available post-operative brain tumor segmentation dataset for performance benchmarking, the current study (similarly to other works in such area) relied on a privately collected dataset from a single institution. Literature has shown that models developed and tested with data from one collection hardly achieve similar results when applied to data from a different site (Liu et al. [2019]). The ideal situation would

be therefore to collect multi-site data, assuring the data collection process to be diverse in terms of imaging/vendor system, acquisition protocol and subject population. Studies presenting results from data collected at a single institution are fundamental for advancing the current state of things but those limitations have to be acknowledged (Petrick et al. [2013]). Furthermore, image quality has a strong impact on the reported performances so it should be good practice to follow a quality assessment program. In the current study, performances are evaluated in the SRI-24 anatomical template space and not in the original raw one, so that the accurate segmentation is co-registered to the template and extracted ex post, eventually leading to slightly inaccurate anatomical segmentation. Several studies have also underlined that reference standards based on radiologists' opinions are subjective and model performance varies when trained on different ground truth (Revesz et al. [1983]). Ground truth for this work was usually determined by two neurosurgeons, but a consensus-based model from 3+ radiologist is preferred (Petrick et al. [2013]). A further limitation in the proposed work is the final post-processing pipeline proposed to bring back labels to edema, enhancing tumor and resection cavity. If it is true that the parameters are obtained averaging gridsearch outputs, it also worth noticing that the limited amount of data makes it impossible to extrapolate solid and trustworthy values. Indeed, since the network is way less confident in its predictions than in the pre-operative case, it is plausible that such prediction confidence might not be above the chosen threshold for some MRI scans, leading therefore to imprecise segmentations.

6.3 Future work

The ambitious aim of this work was to investigate a possible omni-comprehensive DL tool to investigate post-operative brain tumor segmentation, in all its forms. Given the shortage of available data, transfer learning was adopted as turnaround strategy to leverage the accessibility of pre-operative MRI scans. This led to interesting results but it is nevertheless undeniable that this procedure was hardly sufficient for a deep and thorough learning process. Increasing the dataset size, possibly with multi-site collections, would be therefore the preferred way to go. Since in the medical field this is often hardly viable, reducing the scope of the task might lead to more promising results, especially regarding the resection cavity, undoubtedly the Achilles' heel of post-operative segmentation.
Furthermore, no significant improvement was observed by moving from having only two modalities as input to dealing with the whole configuration, likely because the synthesis procedure led to artificial scans not precise or informative enough. It goes without saying that the optimal situation would be the one where all acquisitions are available but it is expected that, in the absence of non-volumetric MRI, implementing the synthesis procedure feeding to the network more than the single contrast-enhanced modality would lead to better and more meaningful reconstructions.

Bibliography

Nagwa M. Aboelenein, Piao Songhao, Anis Koubaa, et al. Httu-net: Hybrid two track u-net for automatic brain tumor segmentation. *IEEE Access*, 8:101406–101415, 2020. doi: 10.1109/ACCESS.2020.2998601.

Mahnoor Ali, Syed O. Gilani, Asim Waris, et al. Brain tumour image segmentation using deep networks. *IEEE Access*, 8:153589–153598, 2020. doi: 10.1109/ACCESS.2020.3018160.

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, et al. The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *CoRR*, abs/2107.02314, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 2017.

Mostefa Ben Naceur, Mohamed Akil, Rachida Saouli, and Rostom Kachouri. Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. *Medical Image Analysis*, 63, 03 2020. doi: 10.1016/j.media.2020.101692.

Bhavneet Bhinder, Coryandar Gilvary, Neel S. Madhukar, and Olivier Elemento. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discovery*, 11(4):900–915, 04 2021. ISSN 2159-8274. doi: 10.1158/2159-8290.CD-21-0090.

Erena S. Biratu, Friedhelm Schwenker, Yehualashet Megersa, and Taye Girma Debelee. A survey of brain tumor segmentation and classification algorithms. *Journal of Imaging*, 7, 2021. doi: 10.3390/jimaging7090179.

Aymen Bougacha, Jihene Boughariou, Mohamed Slima, et al. Comparative study of supervised and unsupervised classification methods: Application to automatic mri glioma brain tumors segmentation. pages 1–5, 2018. doi: 10.1109/ATSIP.2018.8364463.

Ronald A. Castellino. Computer aided detection (cad): an overview. *Cancer Imaging*, 5: 17–19, 2005. doi: 10.1102/1470-7330.2005.0018.

Liang Chen, Paul Bentley, and Daniel Rueckert. Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. *NeuroImage: Clinical*, 15: 633–643, 2017a. ISSN 2213-1582. doi: 10.1016/j.nicl.2017.06.016.

Wei Chen, Xu Qiao, Boqiang Liu, et al. Automatic brain tumor segmentation based on features of separated local square. pages 6489–6493, 10 2017b. doi: 10.1109/CAC.2017.8243946.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, et al. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2016*, pages 424–432. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46723-8_49.

Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Marco D. Cirillo, David Abramian, and Anders Eklund. Vox2vox: 3d-gan for brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 274–284, 2020.

Nicolas Coudray, Andre L. Moreira, Theodore Sakellaropoulos, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *bioRxiv*, 2017. doi: 10.1101/197574.

Robert W. Cox, John Ashburner, Hester Breman, et al. A (sort of) new image data format standard: Nifti-1. *10th Annual Meeting of the Organization for Human Brain Mapping*, 22, 2004.

Christos Davatzikos, Saima Rathore, Spyridon Bakas, et al. Cancer imaging phenomics toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of Medical Imaging*, 5:1, 01 2018. doi: 10.1117/1.JMI.5.1.011018.

Vincenzo Di Nunno, Mario Fordellone, Giuseppe Minniti, et al. Machine learning in neuro-oncology: toward novel development fields. *Journal of Neuro-Oncology*, pages 1–14, 06 2022. doi: 10.1007/s11060-022-04068-7.

Yi Ding, Fujuan Chen, Yang Zhao, et al. A stacked multi-connection simple reducing net for brain tumor segmentation. *IEEE Access*, 7:104011–104024, 2019a. doi: 10.1109/ACCESS.2019.2926448.

Yi Ding, Chang Li, Qiqi Yang, et al. How to improve the deep residual network to segment multi-modal brain tumor images. *IEEE Access*, 7:152821–152831, 2019b. doi: 10.1109/ACCESS.2019.2948120.

Jose Dolz, Ismail B. Ayed, and Christian Desrosiers. Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities. *CoRR*, abs/1810.07003, 2018.

Roelant Eijgelaar, Martin Visser, Domenique Müller, et al. Robust deep learning-based segmentation of glioblastoma on routine clinical mri scans using sparsified training. *Radiology. Artificial intelligence*, 2:e190103, 09 2020. doi: 10.1148/ryai.2020190103.

Ekin Ermis, Alain Jungo, Robert Poel, et al. Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning. *Radiation Oncology*, 15, 05 2020. doi: 10.1186/s13014-020-01553-z.

Andre Esteva, Katherine Chou, Serena Yeung, et al. Deep learning-enabled medical computer vision. *NPJ Digital Medicine*, 4, 2021.

Timea Fulop, Győrfi Ágnes, Szabolcs Csaholczi, et al. Brain tumor segmentation from multi-spectral mri data using cascaded ensemble learning *. pages 531–536, 06 2020. doi: 10.1109/SoSE50414.2020.9130550.

Michał Futrega, Alexandre Milesi, Michal Marcinkiewicz, and Pablo Ribalta. Optimized U-Net for Brain Tumor Segmentation. *arXiv e-prints*, 2021. doi: 10.48550/arXiv.2110.03352.

Sarah E. Gerard and Joseph M. Reinhardt. Pulmonary lobe segmentation using a sequence of convolutional neural networks for marginal learning. pages 1207–1211, 2019. doi: 10.1109/ISBI.2019.8759212.

Mina Ghaffari, Gihan Samarasinghe, Michael Jameson, et al. Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images. *Magnetic Resonance Imaging*, 86:28–36, 2022. doi: 10.1016/j.mri.2021.10.012.

McKinsey L. Goodenberger and Robert B. Jenkins. Genetics of adult glioma. *Cancer genetics*, 205(12):613–621, 2012. ISSN 2210-7762. doi: 10.1016/j.cancergen.2012.10.009.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial networks, 2014.

Toktam Hatami, Mohammad Hamghalam, Omid Reyhani-Galangashi, and Sattar Mirzakuchaki. A machine learning approach to brain tumors segmentation using adaptive random forest algorithm. 2019. doi: 10.1109/KBEI.2019.8735072.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

Ragnhild H. Helland, David Bouget, Alexandros Ferles, et al. Segmentation of post-operative glioblastoma. In *Medical Imaging with Deep Learning*. MIDL 2022 submission, 2022.

Théophraste Henry, Alexandre Carré, Marvin Lerousseau, et al. Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: A brats 2020 challenge solution. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 327–339. Springer International Publishing, 2021.

Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, et al. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119474.

Ahmed Hosny, Chintan Parmar, John Quackenbush, et al. Artificial intelligence in radiology. *Nat Rev Cancer*, 18:500–510, 2018. doi: 10.1038/s41568-018-0016-5.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning: Methods, Systems, Challenges*. 2019. doi: 10.1007/978-3-030-05318-5.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*, 18: 203–211, 2021a. doi: 10.1038/s41592-020-01008-z.

Fabian Isensee, Paul F. Jäger, Peter Full, et al. nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 118–132. Springer International Publishing, 2021b. doi: 10.1007/978-3-030-72087-2_11.

Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7:29, 07 2016. doi: 10.4103/2153-3539.186902.

Haozhe Jia, Weidong Cai, Heng Huang, and Yong Xia. H2nf-net for brain tumor segmentation using multimodal mr imaging: 2nd place solution to brats challenge 2020 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries-6th International Workshop, Springer*, pages 58–68, 2021.

Zeyu Jiang, Changxing Ding, Minfeng Liu, and Dacheng Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 231–241, Cham, 2020. Springer International Publishing.

Alain Jungo, Raphael Meier, Ekin Ermis, et al. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. *CoRR*, abs/1806.03106, 2018. doi: 10.48550/arXiv.1806.03106.

Konstantinos Kamnitsas, Liang Chen, Christian Ledig, et al. Multiscale 3d convolutional neural networks for lesion segmentation in brain mri. *Proc. MICCAI Ischemic Stroke Lesion Segmentation Challenge*, 01 2015.

Konstantinos Kamnitsas, Christian Ledig, Virginia Newcombe, et al. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36, 03 2016. doi: 10.1016/j.media.2016.10.004.

Hee Kim, Alejandro Cosa-Linan, Nandhini Santhanam, et al. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22, 04 2022. doi: 10.1186/s12880-022-00793-7.

Chin-Chi Kuo, Chun-Min Chang, Kuan-Ting Liu, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. *npj Digital Medicine*, 2, 12 2019. doi: 10.1038/s41746-019-0104-2.

Haichun Li, Ao Li, and Minghui Wang. A novel end-to-end brain tumor segmentation method using improved fully convolutional networks. *Computers in Biology and Medicine*, 108, 03 2019. doi: 10.1016/j.compbiomed.2019.03.014.

Heyi Li, Dongdong Chen, William H. Nailon, et al. Improved breast mass segmentation in mammograms with conditional residual u-net. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 81–89, 2018. doi: 10.1007/978-3-030-00946-5_9.

Qiang Li and Kunio Doi. Comparison of typical evaluation methods for computer-aided diagnostic schemes: Monte carlo simulation study. *Medical Physics*, 34(3):871–876, 2007. doi: 10.1118/1.2437130.

Geert Litjens, Thijs Kooi, Babak E. Bejnordi, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 1361-8415. doi: 10.1016/j.media.2017.07.005.

Xiaoxuan Liu, Livia Faes, Aditya Kale, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1, 2019. doi: 10.1016/S2589-7500(19)30123-2.

E. Lotan, B. Zhang, S. Dogra, et al. Development and practical implementation of a deep learning–based pipeline for automated pre- and postoperative glioma segmentation. *American Journal of Neuroradiology*, 43(1):24–32, 2022. doi: 10.3174/ajnr.A7363.

David N. Louis, Arie Perry, Pieter Wesseling, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8):1231–1251, 2021. ISSN 1522-8517. doi: 10.1093/neuonc/noab106.

Claudio Luchini, Antonio Pea, and Aldo Scarpa. Artificial intelligence in oncology: current applications and future perspectives. *Br J Cancer*, 126(1):4–9, 2022. doi: 10.1038/s41416-021-01633-1.

Huan Minh Luu and Sung-Hong Park. Extending nn-unet for brain tumor segmentation, 2021.

Chao Ma, Gongning Luo, and Kuanquan Wang. Concatenated and connected random forests with multiscale patch driven active contour model for automated brain tumor segmentation of mr images. *IEEE Transactions on Medical Imaging*, PP:1–1, 02 2018. doi: 10.1109/TMI.2018.2805821.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Kimberly Miller, Quinn Ostrom, Carol Kruchko, et al. Brain and other central nervous system tumor statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71, 08 2021. doi: 10.3322/caac.21693.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 311–320. Springer International Publishing, 2019.

Maximilian Niyazi, Michael Brada, Anthony J. Chalmers, et al. Estro-acrop guideline "target delineation of glioblastomas". *Radiotherapy and Oncology*, 118(1):35–42, 2016. ISSN 0167-8140. doi: 10.1016/j.radonc.2015.12.003.

Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, et al. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

Alexander F. I. Osman and Nissren M. Tamam. Deep learning-based convolutional neural network for intramodality brain mri synthesis. *Journal of Applied Clinical Medical Physics*, 23(4):e13530, 2022. doi: 10.1002/acm2.13530.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Anand K. Parekh, Richard A. Goodman, Catherine Gordon, et al. Managing multiple chronic conditions: A strategic framework for improving health outcomes and quality of life. *Public Health Reports*, 126(4):460–471, 2011. doi: 10.1177/003335491112600403.

Sarthak Pati, Saima Rathore, Aimilia Gastounioti, et al. *The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview*, volume 11993, pages 380–394. 05 2020. doi: 10.1007/978-3-030-46643-5_38.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, 2016. doi: 10.1109/TMI.2016.2538465.

Nicholas Petrick, Berkman Sahiner, Samuel G. Armato III, et al. Evaluation of computer-aided detection and diagnosis systems). volume 40, page 087001, 2013. doi: 10.1118/1.4816310.

Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *CoRR*, abs/1902.07208, 2019.

Muhammad I. Razzak, Muhammad Imran, and Guandong Xu. Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(5):1911–1919, 2019. doi: 10.1109/JBHI.2018.2874033.

Jacob C. Reinhold, Blake E. Dewey, Aaron Carass, and Jerry L. Prince. Evaluating the impact of intensity normalization on MR image synthesis. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109493H. International Society for Optics and Photonics, 2019. doi: 10.1117/12.2513089.

George Revesz, Harold L. Kundel, and Michael Bonitatibus. The effect of verification on the assessment of imaging techniques. *Investigative Radiology*, 25:461–472, 1983. doi: 10.1097/00004424-198303000-00018.

Thomas Roberts, Harpreet Hyare, Giulia Agliardi, et al. Noninvasive diffusion magnetic resonance imaging of brain tumour cell size for the early detection of therapeutic response. *Scientific Reports*, 10:9223, 2020. doi: 10.1038/s41598-020-65956-4.

Torsten Rohlfing, Natalie Zahr, Edith Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31: 798–819, 2009. doi: 10.1002/hbm.20906.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24574-4_28.

Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021. doi: 10.1109/access.2021.3086020.

Rahman Siddiquee and Andriy Myronenko. Redundancy reduction in semantic segmentation of 3d brain tumor mris. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 163–172. Springer International Publishing, 2022.

Jindong Sun, Yanjun Peng, Yanfei Guo, and Dapeng Li. Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3d fcn. *Neurocomputing*, 423:34–45, 01 2021. doi: 10.1016/j.neucom.2020.10.031.

Hui Tang, Huangxiang Lu, Weiping Liu, and Xiaodong Tao. Tumor segmentation from single contrast mr images of human brain. 2015:46–49, 07 2015. doi: 10.1109/ISBI.2015.7163813.

Guotai Wang, Wenqi Li, Maria Zuluaga, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37:1562 – 1573, 01 2018. doi: 10.1109/TMI.2018.2791721.

Tonghe Wang, Yang Lei, Yabo Fu, et al. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of Applied Clinical Medical Physics*, 22, 12 2020. doi: 10.1002/acm2.13121.

Simon Warfield, Kelly Zou, and William Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23:903–21, 08 2004. doi: 10.1109/TMI.2004.828354.

Wenting Wu, Kathleen Lamborn, Jan Buckner, et al. Joint ncctg and nabtc prognostic factors analysis for high-grade recurrent glioma. *Neuro-oncology*, 12:164–72, 02 2010. doi: 10.1093/neuonc/nop019.

Fan Xu, Haoyu Ma, Junxiao Sun, et al. Lstm multi-modal unet for brain tumor segmentation. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 236–240, 2019. doi: 10.1109/ICIVC47709.2019.8981027.

Yuan Xue, Tao Xu, Han Zhang, et al. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018. doi: 10.1007/s12021-018-9377-x.

Qianye Yang, Nannan Li, Zixu Zhao, et al. Mri cross-modality image-to-image translation. *Sci Rep*, 10, 2020. doi: 10.1038/s41598-020-60520-6.

Fangyan Ye, Yingbin Zheng, Hao Ye, et al. Parallel pathway dense neural network with weighted fusion structure for brain tumor segmentation. *Neurocomputing*, 425:1–11, 2021. doi: 10.1016/j.neucom.2020.11.005.

Yading Yuan. Automatic brain tumor segmentation with scale attention network. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 285–294. Springer International Publishing, 2021.

Ke Zeng, Spyridon Bakas, Aristeidis Sotiras, et al. Segmentation of gliomas in pre-operative and post-operative multimodal magnetic resonance imaging volumes based on a hybrid generative-discriminative framework. volume 10154, pages 184–194, 04 2017. ISBN 978-3-319-55523-2. doi: 10.1007/978-3-319-55524-9_18.

Ling Zhang, Xiaosong Wang, Dong Yang, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39(7):2531–2540, 2020a. doi: 10.1109/TMI.2020.2973595.

Ziang Zhang, Chengdong Wu, Sonya Coleman, and Dermot Kerr. Dense-inception u-net for medical image segmentation. *Computer Methods and Programs in Biomedicine*, 192: 105395, 2020b. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105395.

S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, et al. A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. *CoRR*, abs/2008.09104, 2020a. doi: 10.48550/arXiv.2008.09104.

Tongxue Zhou, Stéphane Canu, and Su Ruan. Fusion based on attention mechanism and context constrain for multi-modal brain tumor segmentation. *Computerized Medical Imaging and Graphics*, 86:101811, 12 2020b. doi: 10.1016/j.compmedimag.2020.101811.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, et al. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, 2018. doi: 10.1007/978-3-030-00889-5_1.

Qikui Zhu, Bo Du, Baris Turkbey, et al. Deeply-supervised cnn for prostate segmentation. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 178–184, 2017. doi: 10.1109/IJCNN.2017.7965852.