

Faculté des sciences

Ecole de Physique

Automatic Delineation of Primary Tumor Volumes from MR Images in Head and Neck Cancer Patients with Deep Learning

Author : Eléonore LIEFFRIG

Supervisors : Edmond STERPIN, John A. LEE

Reader : Ana Maria BARRAGAN MONTERO

Academic year 2021-2022

Master [120] in Physics, Medical Physics

Abstract

Lately, Deep Learning techniques have shown promising results on performing various tasks related to the medical field. More specifically, Convolutional Neural Networks have been used to try to automate tasks linked to the radiation therapy workflow. The latter includes the precise delineation of the tumor and the organs at risk, the dose prescription, orientation of the beams, and so on. Radiotherapy is involved in about half of cancer cases and has to be accurate in order to destroy all the cancer cells while sparing the healthy tissues and organs. In order to achieve this goal, treatment is carefully planned for each patient.

During the course of treatment, changes might arise in the patient's anatomy. Thus, the treatment plan might not be adapted to this new anatomy anymore. Adaptive Radiotherapy addresses this problem by adapting the treatment plan to these changes on the fly. To change the plan so rapidly, the delineation task has to be automated and fast. Convolutional Neural Networks are one candidate for this task. The aim of this thesis is to train a state-of-the-art Convolutional Neural Network (nnU-Net) to automatically delineate the Primary Tumor on Magnetic Resonance Images of 20 Head and Neck Cancer patients to assess how close to a physician's contours the network can get. Provided the database size, the 3D version of nnU-Net showed acceptable results with mean Dice, Hausdorff Distance and Surface Dice of respectively 0.61 ± 0.21 , 16.52 ± 4.25 mm and 0.56 ± 0.23 and median of 0.65, 16.12 mm and 0.56. The network inference on 7 patients showed clinically acceptable results, especially for the lower and posterior part of the Primary Tumor.

In addition, an adaptation of nnU-Net is also proposed. This new network, called nnX-Net, allows taking advantage of different imaging modalities even if they are not registered while keeping the advantages offered by nnU-Net.

Acknowledgments

This master's thesis is the result of months of work and would not be what it is today without the support of people I would like to personally thank.

First, I would like to express my gratitude to my supervisors. I want to thank Professor Edmond Sterpin and Professor John Lee for their availability and good advice throughout the realization of this work. Thank you too for sharing your passion for Medical Imaging and Medical Physics during your lectures which are the ones that definitely got me interested in the field.

I would also like to warmly thank Ana and Margerie. Thank you Ana for your support, for your answers to my countless questions and for the time you spent advising me on how to improve the first version of this thesis. Thank you Margerie for your kindness, good advice and for all the time you dedicated to help me with this project, with a special and huge thank you for the nnX-Net development. I am really going to miss spending Mondays in the lab with you ! I would also like to thank David for helping me getting started with nnU-Net, to Loïc for the fascinating discussions we had about the clinical aspects of this work, and to the other members of the MIRO lab for welcoming me.

Even if he was not directly involved in this project, I also would like to thank Professor John Onofrey from Yale University. Everything I learnt during my internship in his lab last summer definitely helped me a lot during the realisation of this work.

Finally, as this work marks the end of my studies at UCLouvain and thus closes a big chapter of my life, I also would like to thank some family and friends for their support during this project and during my studies in general.

First of all, words cannot express my gratitude to my parents who always supported me. They offered to my brother and me the best conditions to study, whereas it was good homemade meals and pastries, endless encouragements and, more importantly, the value of hard work. I also want to thank my little brother for his support and good advice.

These years at UCLouvain also gave me the chance to meet amazing people. Thank you to Florian and Benjamin for being the best classmates and, more recently, roommates I could have hoped for. I would also like to thank Maxime for all those shared coffees and songs. To conclude, thank you Manuel for always encouraging me to become the best version of myself, I cannot wait to begin a new adventure by your side.

Abbreviations

Abbreviation	Definition
HNC	Head and Neck Cancer
TV	Target Volumes
GTV	Gross Tumor Volume
CTV	Clinical Target Volume
PTV	Planning Target Volume
OAR	Organ at Risk
RF	Radio Frequency
FID	Free Induction Decay
TR	Repetition Time
TE	Echo time
NMR	Nuclear Magnetic Resonance
SE	Spin Echo
GRE	Gradient Echo
FLAIR	Fluid-Attenuated Inversion Recovery
STIR	Short-TI Inversion Recovery
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep learning
ANN	Artificial Neural Networks
CNN	Convolutional Neural Networks
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

DSC	Dice Sørensen Coefficient
JAC	Jaccard Index
IoU	Intersection over Union
HD	Hausdorff Distance
AVSSD	Average symmetric surface distance
AS	Auto-segmentation
ABAS	Atlas-based auto-segmentation
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
GPU	Graphics Processing Unit
CPU	Central Processing Unit

Contents

1	Introduction	1
1.1	Motivations	1
1.2	Objectives	2
1.3	Methodology	2
1.4	Thesis structure	2
2	Clinical and Theoretical Contexts	3
2.1	Head and Neck Cancer	3
2.1.1	Pathology overview	3
2.1.2	Diagnosis and treatment	4
2.2	Radiotherapy Workflow	5
2.2.1	Imaging	5
2.2.2	Contouring and prescription	8
2.2.3	Dose Calculation & Optimization and Plan Verification	11
2.2.4	Treatment Delivery	12
2.3	Magnetic Resonance Imaging	12
2.3.1	Nuclear Magnetic Resonance	13
2.3.2	Physical Principles	13
2.3.3	MRI Sequences	17
2.3.4	Image reconstruction	18
2.3.5	Advantages and drawbacks	18
2.3.6	MR-LINAC	19
2.4	Convolutional Neural Networks	20
2.4.1	From Artificial Intelligence to Convolutional Neural Networks	20
2.4.2	Artificial Neural Networks	21
2.4.2.1	Neurons	22
2.4.2.2	Activation Function	22
2.4.2.3	Hyperparameters	23
2.4.2.4	Learning types	24
2.4.2.5	Model training	24
2.4.2.6	Optimizer	25
2.4.3	Convolutional Neural Networks	25
2.4.3.1	Convolution operations	25
2.4.3.2	Pooling operations	26

3	State of the art in automated segmentation in HNC	28
3.1	Segmentation evaluation metrics	28
3.1.1	Set overlap measures	28
3.1.2	Set boundary agreement measures	30
3.2	Literature Review on automated Segmentation in HNC	31
3.2.1	Atlas-Based and Deep Learning Methods	31
3.2.2	Literature Review for ABAS Methods	32
3.2.3	Literature Review for Deep Learning Methods	33
3.2.3.1	U-Net	33
3.2.3.2	U-Net based CNN architectures	34
3.2.3.3	Other CNN architectures	35
4	Methods	37
4.1	Networks Architectures	37
4.1.1	nnU-Net	37
4.1.1.1	Motivations	37
4.1.1.2	The nnU-Net method	38
4.1.2	nnX-Net	40
4.1.3	Models hyperparameters	41
4.2	Dataset properties and splits	43
4.3	Network Trainings	45
4.3.1	nnU-Net trainings	45
4.3.1.1	nnU-Net 2D	45
4.3.1.2	nnU-Net 3D	45
4.3.1.3	Data augmentation	46
4.3.1.4	Adding of the CT as a channel	46
4.3.2	nnX-Net training	46
4.4	Performance assessment	46
5	Results and Discussions	48
5.1	nnU-Net results	48
5.1.1	nnU-Net 2D	48
5.1.2	nnU-Net 3D	50
5.1.2.1	General results	50
5.1.2.2	Inference patients	55
5.2	nnX-Net results	56
5.3	Discussions	56
5.3.1	nnU-Net 2D vs 3D	56
5.3.2	nnU-Net 3D	57
5.4	Limitations in the approach	59
5.4.1	CNN limitations	59
5.4.2	Data quantity and user-dependant variability	60
5.5	Perspectives	61
5.5.1	Extra input information	61
5.5.2	nnX-Net	61
5.5.3	Other DL techniques	61
5.5.3.1	Attention Networks	61
5.5.3.2	Transformers	62

6	Conclusion	63
A	nnU-Net 3D results	65
B	nnU-Net 3D - Inference results	75

Chapter 1

Introduction

In 2018, 18.1 million people were diagnosed with cancer worldwide and 9.6 million people died from this disease the same year [1]. Cancer is expected to become the leading cause of death during the 21st century. Therefore, being able to treat it is crucial for public health.

Currently, cancer treatments mainly comprise surgery, chemotherapy, radiation therapy, hormone therapy and, more recently, immunotherapy. Radiation therapy is involved in about half of cancer treatments and aims to get a loco-regional control of the tumor thanks to the use of ionizing particles. For this treatment to be as efficient as possible and induce the least side effects, a very accurate representation of the patient's anatomy and tumor location is needed. This is why patients undergo X-ray Computed Tomography (CT) scanning, on which segmentation is performed, meaning that the different organs, tumors and metastases are precisely localized and their volumes delineated.

1.1 Motivations

The image segmentation task is performed by a radiation oncologist. It is a time-consuming and cumbersome process, that also depends on the experience of the physician. Regarding the tumor especially, errors in its segmentation can impact the patient's chances of recovery. For these reasons, it would be of great interest to develop a standardized tool which assists the physician to perform segmentation or even achieves it fully automatically.

Since a few years, interest has grown in applying Machine Learning techniques to perform medical tasks. Indeed, Neural Networks have been trained to automatically perform biomedical images segmentation [2], dose prediction in radiation therapy [3], medical object detection [4] and so on. Many of these algorithms showed promising results. Therefore, scientists are currently conducting research on the development of models that could be used in clinical routine.

Furthermore, this thesis is focused on performing segmentation on Magnetic Resonance Imaging (MRI). This imaging technique has several advantages: indeed, compared to X-ray scanning, it is non-ionizing, it enables a better soft-tissue visualisation (which is particularly interesting in Head and Neck Cancer) and it would be a great candidate as the imaging method in adaptive radiation therapy using MRI linear accelerator [5]. In this kind of radiation therapy, the treatment strategy is adapted to anatomical changes during the treatment. More details about that specific topic can be found in subsection 2.3.6.

As the planning CT scan and diagnostic MR images are not acquired at the same time, there can be anatomical modifications (e.g., movement of the patients, cavity filling) in the time

interval between both acquisitions. Usually, the segmentation is performed in one of the images and registered to the remaining one. However, developing a model that is able to perform segmentation on both images while still combining their information avoids the errors caused during the registration process. This is why the considered model was adapted (see the following subsection).

1.2 Objectives

This thesis aims to explore and analyse the accuracy of a state-of-the-art neural network (nnU-Net) for segmentation of Primary Tumor volumes in Head and Neck Cancer from MR images.

In addition, since radiation therapy treatment planning must be performed with Computed Tomography images (see Section 2.2.1), the adaptation of this state-of-the-art model to segment the tumor in both MR and CT images simultaneously is also investigated.

1.3 Methodology

The primary tumor was delineated by a radiation oncologist on 20 MRIs of Head and Neck Cancer patients from Saint-Luc University Hospital. Next, these data were used to train a state-of-the-art Convolutional Neural Network called nnU-Net. This model is a state-of-the-art segmentation method belonging to AutoML (see subsection 2.4.2.3). The performances of nnU-Net were assessed. Its architecture and code were studied in order to build a second network called nnX-Net. It is a variation of nnU-Net allowing to input two images from different imaging modalities, even if they are not registered.

1.4 Thesis structure

This thesis is structured as follows: the second chapter first introduces the clinical context, i.e., which cancer type is of interest in this work, how radiation therapy is conducted globally and which task of the radiotherapy workflow is performed in this thesis. The Chapter also contains an overview of the Magnetic Resonance Imaging technique. Finally, a section is dedicated to an introduction to Artificial and Convolutional Neural Networks. Their different components are described, as well as how they are trained.

In Chapter 3, some metrics used to evaluate the performances of Convolutional Neural Networks are presented. The second part of this chapter contains a literature review on how segmentation is currently performed in Head and Neck Cancer in both CT scans and MRIs.

The fourth Chapter focuses on the methods that were used in this thesis. The two main networks that were used are described, as well as the database and how the latter was split to run the networks.

In Chapter 5, the results obtained from the different experiments are reported. The second main part of the Chapter discusses them and put on view limitations of the experiments.

Finally, Chapter 6 draws conclusions and sketches a few perspectives on how the present work could be enhanced.

Chapter 2

Clinical and Theoretical Contexts

This chapter is firstly devoted to an introduction to the pathology of interest in this work, which is Head and Neck Cancer (HNC) and to the Radiotherapy Workflow (more specifically to the segmentation part). Then, an overview of the Magnetic Resonance Imaging technique is proposed. Finally, the fourth section introduces Artificial and Convolutional Neural Networks.

2.1 Head and Neck Cancer

2.1.1 Pathology overview

HNC is the 4th most frequent type of cancer in men and 11th most frequent in women in Belgium [6]. This cancer may develop in several organs and structures of the head and neck represented in Figure 2.1. They include the oral cavity (lips, hard palate, ...), the three parts of the pharynx (nasopharynx, oropharynx, hypopharynx), the larynx, the paranasal sinuses, the nasal cavity and the salivary glands. Figure 2.2 details the different primary sites localisation and incidence of HNC in Belgium in 2017.

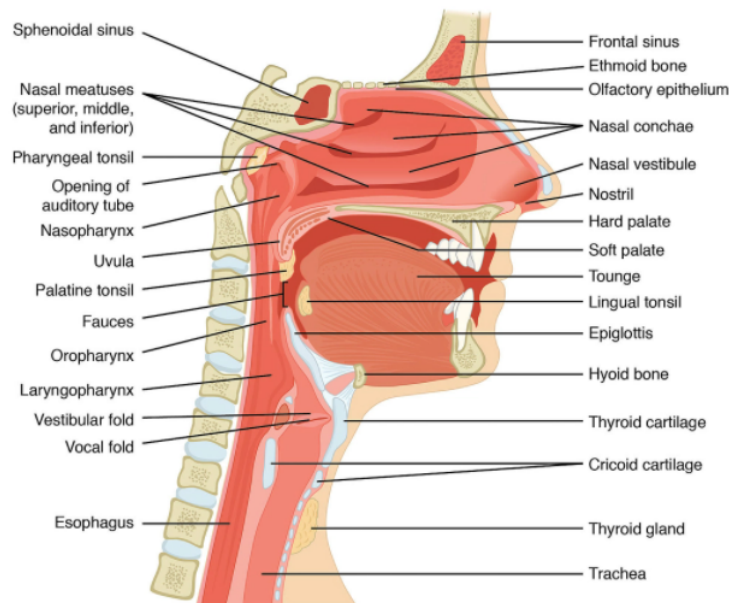


Figure 2.1: Head and Neck Anatomy - Sagittal Plane (from [7])

In 2019, 2766 HNCs (708 in women and 2058 in men) were diagnosed in Belgium, accounting for 4% of all cancers diagnosed that year, the survival rate at 5 years at the time being 51.9% for men and 59% for women [6]. Research has shown that many risk factors for HNC exist. They may be related to lifestyle (exposure to tobacco and alcohol), to a viral infection (Human Papillomavirus, Epstein-Barr virus) or to genetic disorders (Fanconi anemia) [8].

Table 2 Head and neck cancer: Incidence by primary site and sex, Belgium 2017						
Belgium 2017	Total		Males		Females	
	N	%	N	%	N	%
Lip (C00)	32	1.2	23	1.2	9	1.3
Oral cavity (C02-C050,C06)	666	24.8	449	22.6	217	31.1
Tongue (C02)	271	10.1	188	9.5	83	11.9
Gum (C03)	101	3.8	54	2.7	47	6.7
Floor of mouth (C04)	197	7.3	143	7.2	54	7.7
Hard palate (C050)	23	0.9	13	0.7	10	1.4
Mouth, NOS (C06)	74	2.8	51	2.6	23	3.3
Pharynx (C01;C051-C059;C09-C13)	1,144	42.7	868	43.8	276	39.5
Oropharynx (C01;C051-C059;C09-C10)	779	29.0	574	28.9	205	29.4
Base of tongue (C01)	191	7.1	140	7.1	51	7.3
Soft palate, uvula and palate NOS (C051-C059)	100	3.7	74	3.7	26	3.7
Tonsil (C09)	304	11.3	224	11.3	80	11.5
Oropharynx, other and NOS (C10)	184	6.9	136	6.9	48	6.9
Nasopharynx (C11)	63	2.3	50	2.5	13	1.9
Hypopharynx (C12-C13)	302	11.3	244	12.3	58	8.3
Pyriformis sinus (C12)	202	7.5	165	8.3	37	5.3
Hypopharynx, other and NOS (C13)	100	3.7	79	4.0	21	3.0
Larynx (C32)	590	22.0	495	24.9	95	13.6
Glottis (C320)	350	13.0	307	15.5	43	6.2
Supraglottis (C321)	192	7.2	152	7.7	40	5.7
Larynx, other and NOS (C322-C329)	48	1.8	36	1.8	12	1.7
Nasal cavity and paranasal sinuses (C30-C31)	123	4.6	84	4.2	39	5.6
Nasal cavity and middle ear (C30)	57	2.1	31	1.6	26	3.7
Accessory sinuses (C31)	66	2.5	53	2.7	13	1.9
Salivary glands (C07-C08)	111	4.1	52	2.6	59	8.5
Parotid gland (C07)	89	3.3	46	2.3	43	6.2
Salivary glands, NOS (C08)	22	0.8	6	0.3	16	2.3
Lip, oral cavity and pharynx, NOS (C14)	16	0.6	13	0.7	3	0.4
Head and neck (C00-C14;C30-C32)	2,682	100.0	1,984	100.0	698	100.0

Figure 2.2: HNC incidence by primary site and gender in Belgium (2017) (from [6])

2.1.2 Diagnosis and treatment

Patients affected by HNC may suffer from different symptoms depending on the location and size of the tumor. For instance, they may experience pain when swallowing (throat tumor), trouble breathing or speaking and ear pain (larynx tumor), white or red patch on the gums, the tongue, or the lining of the mouth (oral cavity tumor).

Clinical examination and imaging procedures such as CT scan, MRI or 18F-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET or PET scan) allow to formulate an appropriate diagnosis of the cancer and to stage it (i.e., determine the type and malignancy of the tumor and identify eventual metastasis).

For early-stage HNC, the treatment will consist in surgery, radiotherapy or both. If it is advanced stage, the patient will undergo either a combination of surgery and radio(chemo)therapy or a radiotherapy. Thus, radiotherapy plays a very big part in treating HNC since it is involved in about 75 % of HNC treatment [9]. The following section is dedicated to describing the differ-

ent steps of the radiotherapy treatment and aims to point out the one that will be of particular interest in this work, which is segmentation.

2.2 Radiotherapy Workflow

As previously mentioned, radiotherapy ¹ is presently involved in about half of cancer treatments in general [10]. The use of ionizing radiations coming from charged particles (e.g., protons or carbon ions) or high energy photons (generally about 6MV) to kill cancer cells and thus cure the patient, has proven itself throughout the last decades. Planning and delivery techniques are in continuous evolution to improve the efficiency of treatments. The aim of radiotherapy is to ensure loco-regional control of the cancer cells by damaging their DNA so much that they are no longer able to repair it and thus die.

However, the damaging interactions between the particles and the tumor can also happen between the particles and the healthy cells. This leads to unwanted side effects which can strongly impact the patient's life quality or even cause radiation-induced cancer years later. In HNC especially, the most frequent side effects are xerostomia (dry mouth), dysphagia (difficulty swallowing), and dermatitis (skin inflammation). Therefore, a good trade-off has to be found between delivering enough radiation to kill all the malignant cells while sparing the surrounding healthy tissues.

Whatever the cancer type a patient is suffering from, the steps leading to a complete radiotherapy (called the Radiotherapy Workflow) are identical. They are illustrated in Figure 2.3 and detailed in the following subsections.

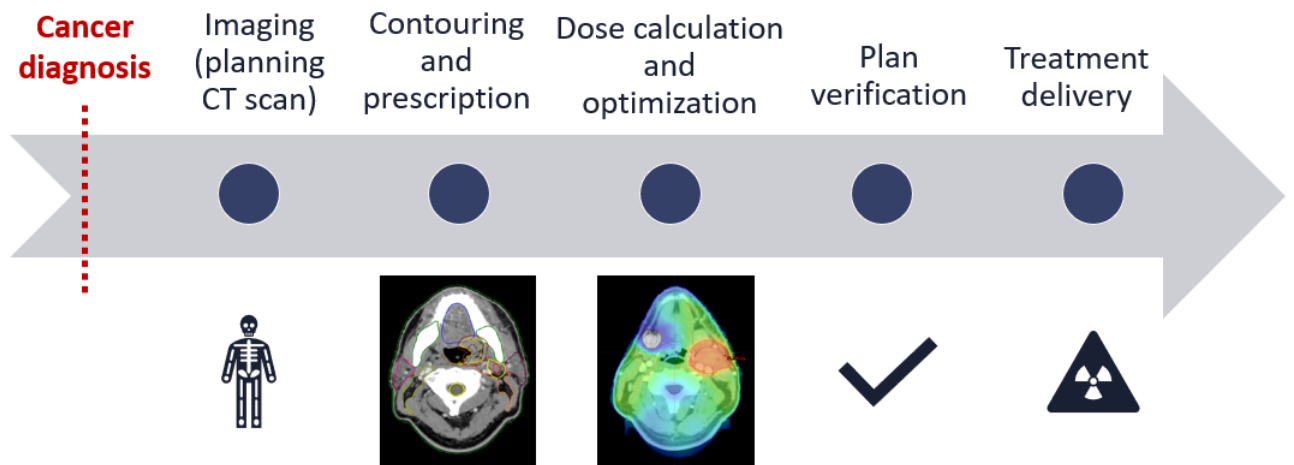


Figure 2.3: Schematic of the Radiotherapy Workflow (medical images from [9])

2.2.1 Imaging

For the treatment to be as accurate as possible, it is essential to have a thorough representation of the patient's anatomy and tumor location. To achieve this goal, the patient undergoes a CT scan. During the examination, in the case of HNC, the patient lies on the couch and is immobilized with a customized thermoplastic mask covering the head and shoulders, which is

¹In this work, "radiotherapy" will refer to external radiotherapy which is the most common type, but other ways to deliver a dose to a tumor and thus perform radiotherapy exist, such as brachytherapy or metabolic radiotherapy.

fixed to the table as shown in Figure 2.4. In other types of cancers, different immobilization tools exist and foam pads may also be added to ensure that the patient is not too uncomfortable during the imaging process (e.g., under the knees, arms, and so on). All the accessories and how they are used are carefully recorded in the patient's file because this exact position and set-up will be the one for each treatment session, since the positions and fluences of the photon beams to kill the tumor will be computed based on this reference position. This CT scan is of high spatial resolution and since it will be the reference to achieve treatment planning, it is called the planning CT scan.

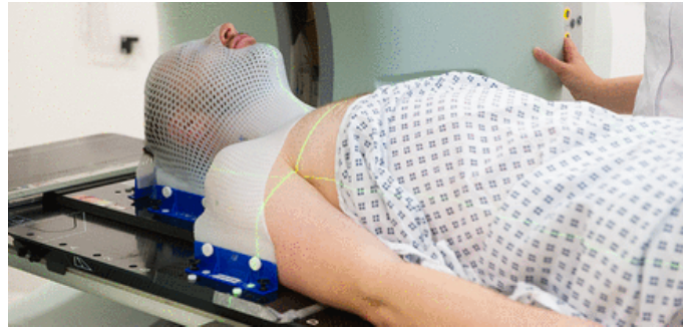


Figure 2.4: Patient lying on the table wearing a thermoplastic mask (from [9])

The CT scan, shown in Figure 2.5 uses an X-ray tube rotating around the patient while he/she is sliding inside the gantry. A range of detectors is facing the X-ray tube and allows measuring the X-ray attenuation by the traversed tissues. These attenuations can be related to tissue types thanks to the Hounsfield Unit (HU), given by the formula

$$HU = 1000 \cdot \frac{\mu_X - \mu_w}{\mu_w} , \quad (2.1)$$

where μ_X and μ_w are the linear attenuation coefficients of the traversed tissue and water, respectively. The Hounsfield scale is represented in Figure 2.6. This relationship is a big advantage of CT scans compared to MR machines, for example. There is indeed a quantitative mapping between the nature of the tissue and how it attenuates the photons, which is the radiation that will also be used for treatment.

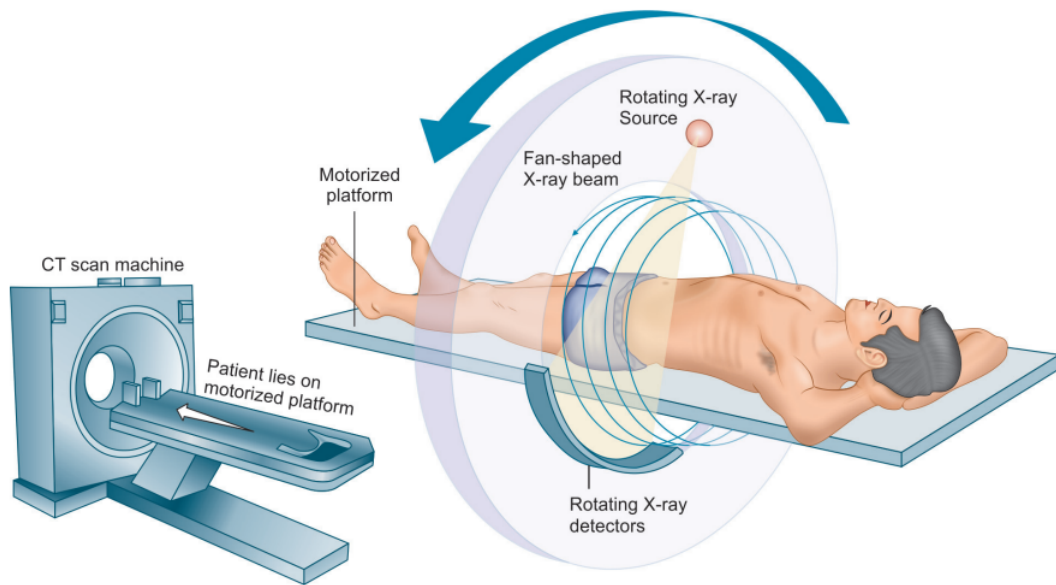


Figure 2.5: CT scan machine overall geometry and working principle (from [11])

In helical CT, the X-ray tube and detectors follow an helical path with respect to the patient. This leads to attenuation measures for different angles and heights in the craniocaudal direction of the patient. Thanks to reconstruction algorithms, these measures on an helical path are interpolated into circles corresponding to the different patient slices.

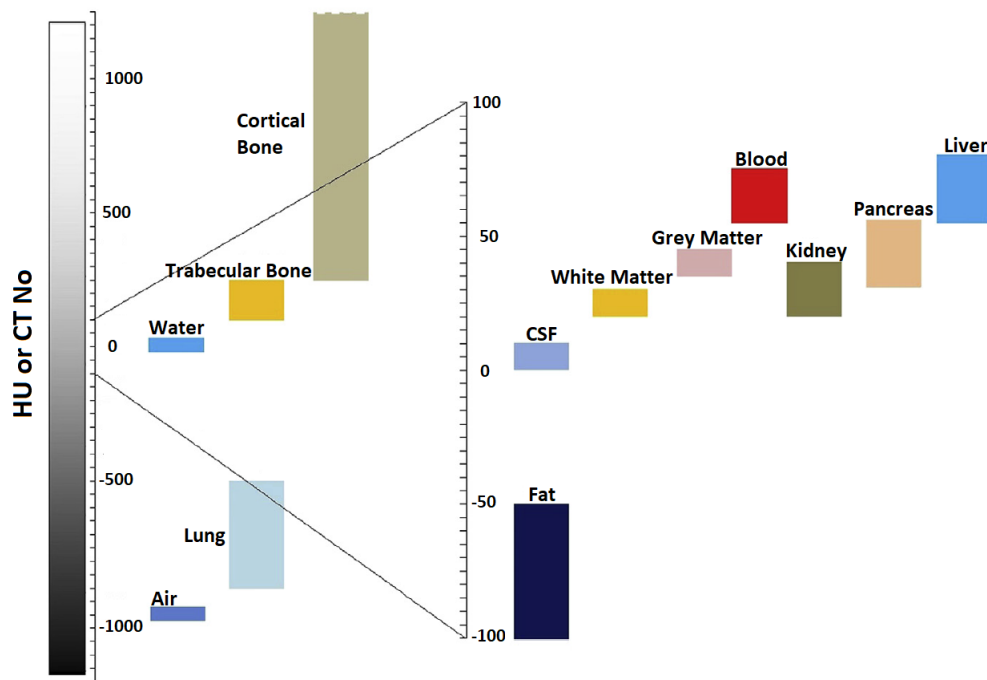


Figure 2.6: The Hounsfield scale allows to link tissue type with its X-ray attenuation (adapted from [12])

For HNC, in Saint-Luc University Hospital the protocol requires other imaging exams to be done: a Positron Emission Tomography (PET) scan and an MRI. These imaging techniques have different advantages. MRI enables a better soft tissue differentiation than CT. This property is especially interesting when imaging HNC patients (see subsection 2.3.5). A PET scan highlights higher metabolic activity which can be characteristic of cancer cells, allowing the tumor to be located and the spread of the cancer to be assessed in the whole body (lymph nodes, metastasis). As the clinical practice changes from one hospital to another, some of them might not perform PET scans or MRIs, although they are encouraged to do so, since these extra images enable better tumor visualization.

Thus, these imaging modalities can be combined with the CT scan to help localising the tumor, metastases, and anatomical structures as accurately as possible. Unfortunately, the different images are often acquired at different moments and in different positions. Therefore, they can not just be superimposed, they have to be registered to align the anatomical structures correctly. It is also possible to use images from an endoscopy. All these imaging modalities examples are represented in Figure 2.7.

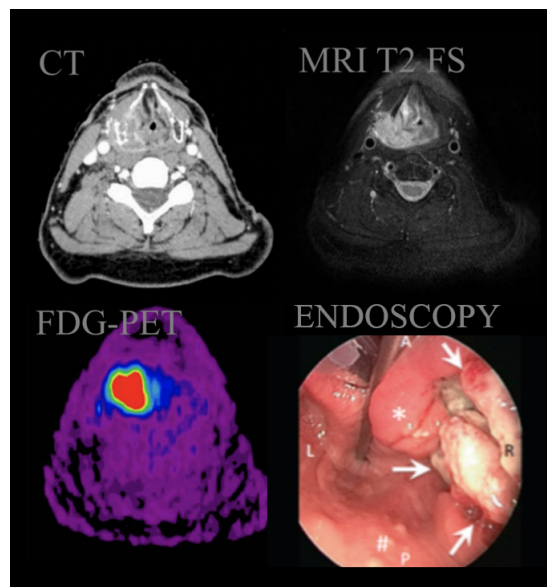


Figure 2.7: Different imaging modalities of a patient suffering from glottic cancer (images from [9])

2.2.2 Contouring and prescription

Once an accurate representation of the patient's anatomy is available, the treatment plan can start being designed. It includes deciding what dose the tumor should receive, in how many fractions², with how many beams, what is the maximum dose that the surrounding organs can receive, and so on. This task is performed by both the radiation oncologist and the medical physicist and/or dosimetrist working together.

²Treatment fractionation consists in dividing the total dose that the tumor should receive into smaller doses delivered at different moments corresponding to the different fractions. The purpose of fractionation is to take advantage of the better repairing capabilities that the healthy cells have over the cancer cells.

First of all, the radiation oncologist contours the Target Volumes (TVs) and the Organs at Risk (OARs) on the planning CT scan. The TVs are different volumes encapsulating the tumor that take into account geometrical and medical uncertainties. They are represented in Figure 2.8. Based on the ICRU (International Commission on Radiation Units) Report 62 [13], the TVs can be defined by:

- Gross Tumor Volume (GTV): it is the gross demonstrable location and extent of the malignant growth (definition from [13]). It is composed of the GTV related to the primary tumor (called GTV-T or GTV-P), of the potential metastatic lymphadenopathy (lymph nodes in which the cancer has spread, called GTV-N) and metastasis (GTV-M). This target volume is determined thanks to the combination of the clinical examination and several imaging modalities such as CT, ultrasound, MRI, PET, and so on. As the GTV generally indicates anatomical locations where the tumor cell density is very high, it has to receive an appropriate dose on all the volume in order to kill all tumor cells. In this work, the GTV-T will be the target volume of interest because it is mostly determined on medical images.

When referring to GTV delineation performed by a computer, one will rather speak of segmentation instead of contouring. To be precise, segmentation consists in assigning a class label to each pixel of the image.

- Clinical Target Volume (CTV): it contains the GTV and additional volume of normal tissues with suspected microscopic presence of the tumor that might not be visible in the images. The CTV may correspond to the GTV plus a margin, to anatomical volumes with suspected microscopic tumor presence, or eventually a volume including the zone where the tumor was removed for a post-surgery irradiation. For the first type of CTV (GTV plus a margin), the margin depends on the type and location of the primary tumor. To delineate the CTV, doctors refer to guidelines that were published after analyzing surgically-resected tumor spread with a microscope. The CTV margins are adapted to the anatomic boundaries (i.e., air cavities, muscles, and bones are excluded from the CTV). In HNC, the margin around the GTV is typically of 5 mm [14]. Guidelines exist regarding the delineation of the CTV in HNC and also for the OARs like in the paper of Grégoire et al. [15] in which a consensus was found on CTV-P delineation in laryngeal, hypopharyngeal, oropharyngeal, and oral cavity squamous cell carcinomas. An example of this type of OARs delineation atlas can be found in Figure 2.9.
- Planning Target Volume (PTV): it is a geometrical concept created in order to be sure that the CTV will receive the prescribed dose despite the occurrence of geometrical uncertainties caused by patient setup errors, target motion (e.g., motion due to breathing) or morphological changes. These errors are divided into 2 categories: the systematic errors (when the deviation has a non-zero average during the treatment or if it always happens in the same direction and with the same magnitude throughout the treatment, e.g., an error in the tumor delineation) and the random errors (when the deviation can change in direction and magnitude throughout the treatment, e.g., patient re-positioning error compared to the planning CT scan reference position).

Figure 2.8 displays a schematic of the different TVs, and a clinical example showing the GTV and CTVs:

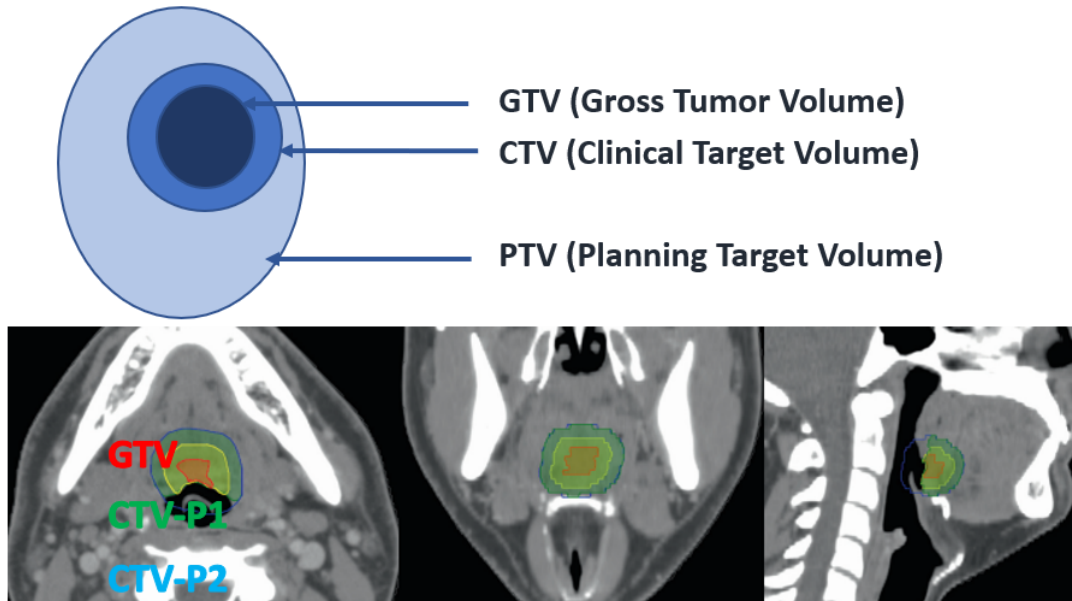


Figure 2.8: Qualitative schematic of the different Target Volumes and an example of GTV and CTV in an HNC from [15]

As soon as the TVs are delineated, the radiation oncologist carries on with the contouring of the OARs. As a matter of fact, the tumors can be surrounded by healthy tissues that are especially vulnerable to radiation. These particular tissues must therefore be avoided at all costs during treatment, in order to prevent their partial or total dysfunction. In some cases, some vulnerable tissues may even develop radiation-induced cancers years after the treatment if they are not adequately protected. Since the anatomy of every human is more or less the same, reference guidelines exist for the contouring of the OARs. For instance, Brouwer et al. [16] defined guidelines based on discussions with radiation oncologists from all over the world. This led to anatomical descriptions explaining how to delineate 25 OARs of the head and neck region.

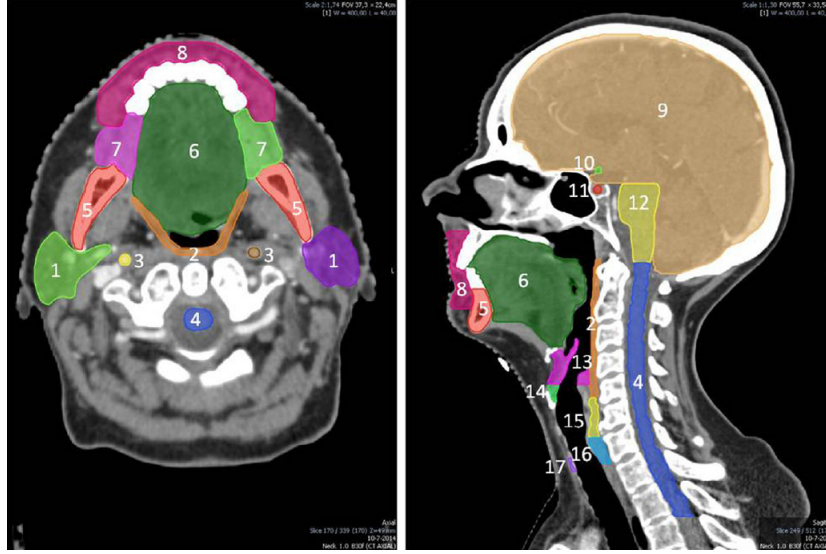


Figure 2.9: Axial and sagittal views of the delineations of the parotid glands (1), pharyngeal constrictor muscles (2), carotid arteries (3), spinal cord (4), mandible (5), extended oral cavity (6), buccal mucosa (7), lips (8), brain (9), chiasm (10), pituitary gland (11), brainstem (12), supraglottic larynx (13), glottic area (14), cricopharyngeal inlet (15), cervical esophagus (16) and thyroid (17). (from [16])

After contouring, the radiation oncologist moves on to the prescription step, i.e., he/she decides the dose that has to be delivered to each TV. The dose is expressed in Grays ($1 \text{ Gy} = 1 \text{ J/kg}$) and depends on the cancer type (typically, the prescription for the primary tumor will be several dozens of Gy). Thresholds for the maximum dose the OARs can receive are also defined individually.

2.2.3 Dose Calculation & Optimization and Plan Verification

The medical physicists and medical dosimetrists are responsible for the next part of the treatment planning, which consists in finding the best physical parameters, e.g., number of beams, angles, use of attenuating wedges and so on, to achieve the radiation oncologist's goals. This is an iterative optimisation process. At each step, a configuration is chosen and the simulated doses that the TV and OARs would receive in this configuration are computed and compared to the physician's prescriptions to determine if they are close enough to the clinical goals or not. Dose calculation can be achieved via analytical algorithms (e.g., pencil beam algorithm) or Monte Carlo simulations. The dose assessment is made thanks to the Dose Volume Histograms (DVH), representing which part of the tumor or OAR volume will receive which dose.

Figure 2.10 represents an example of dose distributions and DVH for the PTV, CTV, and a few OARs in an HNC.

The most widely used radiotherapy technique is Intensity Modulated Radiation Therapy (IMRT). It aims to shape the dose so that it fits the tumor and avoids the OARs. In order to do so, the linear accelerator (LINAC) that delivers the photon beam has a multileaf collimator, composed of leaves of lead moving independently. They move while the LINAC rotates around the patient, in order to shape the beam.

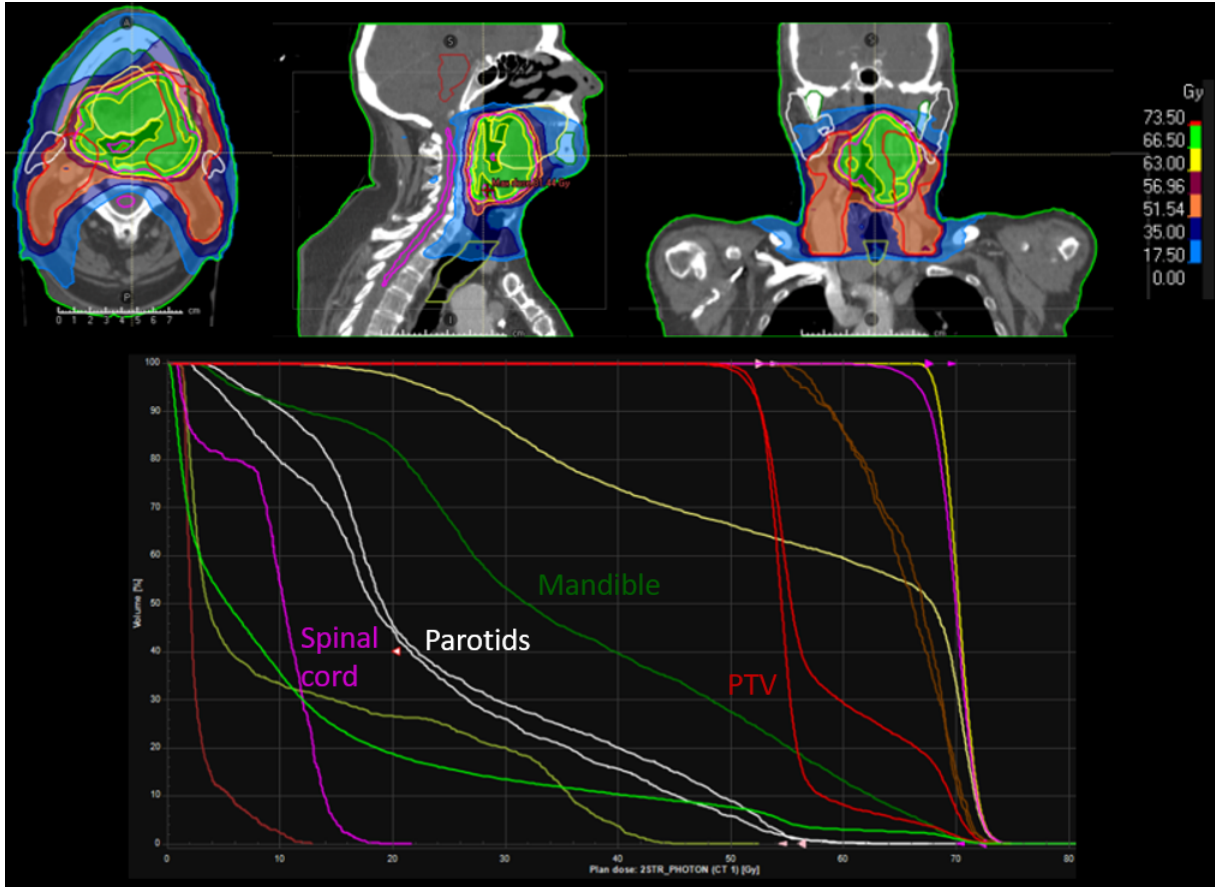


Figure 2.10: Dose distributions and Dose Volume Histograms for a HNC treatment plan. The DVH curve for the PTV is the red one, CTV is in yellow, mandible in green, parotids in white and spinal cord in pink.

Finally, some Quality Assurance tests are realized to ensure patient's safety. For instance, a treatment fraction can be delivered in a water phantom containing an ionization chamber to verify that the delivered dose corresponds to the simulated one. Delivering a dose in a phantom can also make sure that the machine can realize physical constraints like the speed at which the leaves of the collimator have to move. [17]

2.2.4 Treatment Delivery

Once the treatment plan has been approved, the patient can start receiving it, either every day or at other frequencies depending on the treatment fractionation. At each fraction, the patient will be positioned exactly in the same position as at the planning CT acquisition (thanks to the mask, other immobilization material, and lasers) since it was the reference position that was used for treatment planning. A low-dose CT scan can be realised before each fraction to ensure the correct position on the couch.

2.3 Magnetic Resonance Imaging

This section is dedicated to an introduction on how MRI works. Since it is a very complex and wide topic, the next subsections will only focus on a reminder of the physical principles that

underlie this technique. More specifically, a general description of how a final MR image is built is done, and the advantages and drawbacks of the technique are presented and compared with other imaging modalities.

The main references that were used to write this section are the book "Guide des technologies de l'imagerie médicale et de la radiothérapie" [18] and the lectures on MRI by Nicolas Michoux at UCLouvain [19].

2.3.1 Nuclear Magnetic Resonance

In 1938, the American physicist Isidor Isaac Rabi discovers the Nuclear Magnetic Resonance (NMR) phenomenon. It happens when atomic nucleus with certain spin properties and undergoing a strong constant magnetic field is perturbed by a weak magnetic field oscillating at a special frequency. In this setting, the nucleus then reacts by producing an electromagnetic signal at a characteristic frequency.

This phenomenon has many applications in several fields. For instance, NMR spectroscopy is widely used to study the structure of chemicals or determine the purity of samples. Moreover, NMR is also used in magnetometers, in quantum computing and, of course, in medicine. Indeed, Magnetic Resonance Imaging is a tomographic, non invasive and non ionizing imaging modality taking advantage of the Nuclear Magnetic Resonance phenomenon. It can yield 2D or 3D representations of the body at a spatial resolution of $\sim 1\text{mm}$ and also picture physiological processes.

2.3.2 Physical Principles

In MRI, changes in magnetization of hydrogen atoms nuclei (containing only one proton) are triggered and studied. Indeed, hydrogen is of particular interest as it is the most abundant element in the body: it is part of water, but also of sugar, proteins, starch, and fats. Given that an hydrogen atom nucleus magnetization is affected by the fields of the surrounding atoms in a tissue, hydrogen protons in a liquid or in adipose tissue will not have the same magnetization properties. This is one of the key effects that permit tissue differentiation and getting a contrasted image when performing MRI.

A proton carries a positive charge and a spin of $\frac{1}{2}$. Due to its spin, it induces a magnetic field associated with a magnetic moment $\vec{\mu}$ aligned with the direction of the spin vector $\vec{S}$ representing the microscopic magnetization.

Though the spin is a notion intrinsic to quantum physics, and complex developments are required to properly describe the behaviour of one single proton, the functioning of several protons can be explained with a more classical point of view. This is what will be done in the following subsections in order to simplify the analysis. More specifically, a great number of hydrogen atoms will be represented as generating small magnetic moments which precess at the Larmor frequency.

If no magnetic field is applied to a sample (i.e., the typical everyday situation for human tissue), due to thermal motion its proton magnetic moment vectors will be oriented randomly. This results in a null macroscopic magnetization as represented in Figure 2.11: $\vec{M} = \vec{0}$.

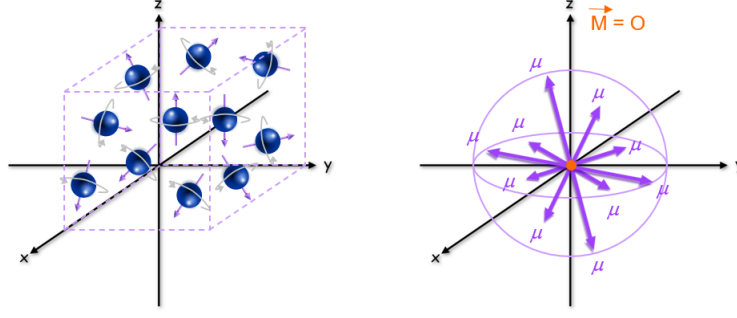


Figure 2.11: Example of macroscopic sample magnetization (from [18])

During an MRI exam, the patient's body is immersed in a strong magnetic field $\vec{B}_0 = B_0 \hat{z}$ (its intensity is typically of 1.5 T, but can vary in a range from 0.2 to 7 T). Consequently, the proton magnetic moment vectors will progressively orient themselves in the field direction. Once this alignment is completed, there is a nonzero macroscopic magnetization $\vec{M}$ in the direction of $\vec{B}_0$. Because of quantum mechanics effects, the magnetic moment vectors are slightly tilted as represented in Figure 2.12 and more of them are oriented in $\hat{z}$ than in $-\hat{z}$. The $\vec{\mu}$ are tilted because the protons precess. More precisely, the magnetic moment $\vec{\mu}$ is tilted from an angle θ (compared to $\hat{z}$) and precess at an angular speed ω_0 so that

$$\omega_0 = \gamma B_0 \quad , \quad (2.2)$$

where γ is the gyromagnetic ratio in MHz/T and B_0 the strength of the external magnetic field in teslas. This relation is called the Larmor equation.

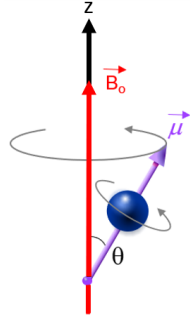


Figure 2.12: Representation of a proton precessing (adapted from [19])

Once dived in the magnetic field, protons will either orient themselves in upward (north pole facing the south pole of $\vec{B}_0$) or the downward (north pole facing the north pole of $\vec{B}_0$) way compared to the magnetic field. There are more protons aligned upward because the system in this state has a lower energy. This induces a macroscopic magnetization $\vec{M} \neq \vec{0}$ in the same direction than $\vec{B}_0$ as displayed in Figure 2.14. In most MRI machines, $\vec{B}_0$ is oriented towards the tunnel. The z-axis is called the longitudinal axis, and the xy-plane is the transverse plane (see Figure 2.13).

³ $\hat{z}$ refers to the unit vector in the z-direction.

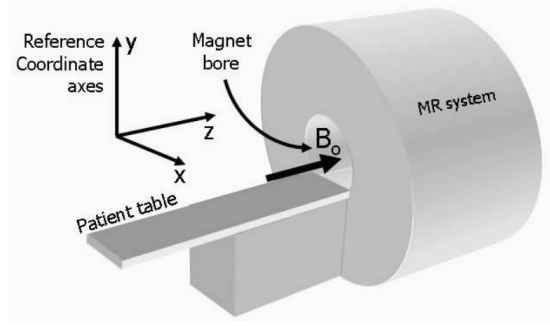


Figure 2.13: Coordinate system associated with an MRI machine (from [20])

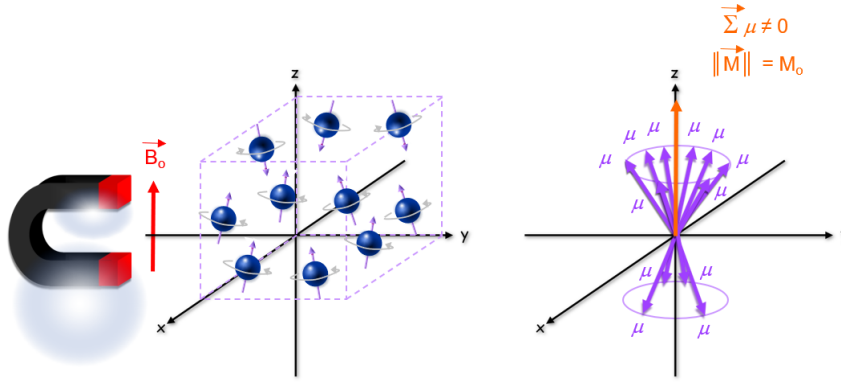


Figure 2.14: Sample once immersed in a magnetic field (from[18])

Unfortunately, since $|\vec{M}| \ll |\vec{B}_0|$ and $\vec{M}$ and $\vec{B}_0$ are in the same direction, it is not possible to measure $\vec{M}$ directly, this information is hidden in $\vec{B}_0$. Thus, it has to be tilted in the transverse plane like shown in Figure 2.15. This is done thanks to a rotating magnetic field (a Radio Frequency Wave (RF) corresponding to the resonance frequency of the atoms of interest) applied during a short period of time. By applying the latter, denoted by $\vec{B}_1(t)$, the system receives energy and is therefore able to have a higher energy. It is applied just during the time needed for half of the proton population to jump into the higher energy level, therefore removing the longitudinal component in the macroscopic magnetization vector. The radio frequency impulse also induces a microscopic magnetization vectors rephasing, allowing the growth of a transverse component M_{xy} . Since $\vec{M}$ was tilted of 90° , this special RF pulse is called a 90° pulse.

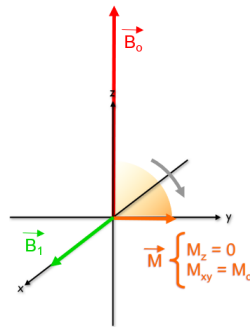


Figure 2.15: Macroscopic magnetization tilted in the transverse plane

When $\vec{B}_1(t)$ is not applied anymore, the relaxation phase, illustrated in Figure 2.16 begins. It is during this phase that the information about the different tissues will be collected. During the relaxation, the system that had jumped to a higher energy level goes back to the initial one, so the longitudinal component of $\vec{M}$ grows back and the transverse component of $\vec{M}$ vanishes. During this process, the system releases energy. These two processes occur at the same time, the $\vec{M}$ grows making a spiral movement to get back to its initial alignment with $\vec{B}_0$. This change in this magnetic field induced by the macroscopic magnetization produces a current that can be measured by a receiving coil in the transverse plane. This is the signal of interest, it is a sinusoid called the Free Induction Decay (FID)

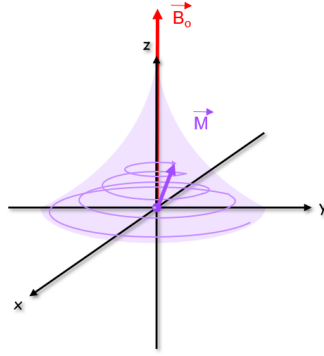


Figure 2.16: Relaxation phase

T1 is the time needed for the longitudinal component of $\vec{M}$ to grow back to 63 % of what it was before applying $\vec{B}_1(t)$. This longitudinal relaxation is also called the T1 relaxation, or spin-lattice relaxation.

T2 is the time corresponding to a 63 % decrease of the transverse magnetization. This transverse relaxation, the T2 relaxation or spin-spin relaxation is induced by the fact that protons interact with each other and therefore loose spin phases coherence. The T1 and T2 relaxations are represented in Figure 2.17:

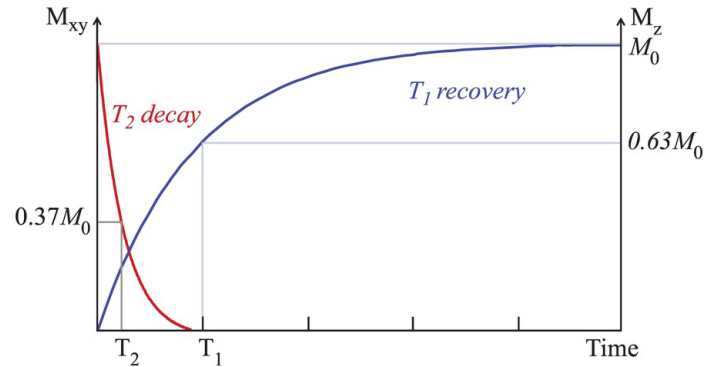


Figure 2.17: The spin-lattice and spin-spin relaxations

2.3.3 MRI Sequences

Here are some examples of values of T1 and T2 for biological tissues and fluids:

	T1 (ms)	T2 (ms)
Water	3000	3000
Muscles	870	47
Fat	260	84
Cerebro-spinal liquid	>4000	>2000
White matter	790	92
Grey matter	920	101
Kidney	650	58
Liver	490	43

Table 2.1: T1 and T2 for various tissues and fluids of the body.

T1 and T2 depend on the molecular organisation of the medium in which the hydrogen atom lies. Actually, in denser systems, the collision frequency with the surrounding atoms is higher and thus maybe close to Larmor frequency. As a consequence, the system will easily release energy to its environment and the spin-lattice relaxation will be fast (and vice-versa). This is why T1 and T2 are smaller in soft tissues than in fluids.

It is possible to enhance the contrast between tissues by choosing appropriate parameters. Among them are the repetition time (TR) which is the time between two RF pulses and the echo time (TE) which is the time after which the RMN signal is collected. There is also the RF pulse angle and the application of different magnetic field gradients. A particular setting of these parameters creates what is called an MRI sequence. Many different sequences exist, resulting in different image appearance. The sequences can be classified as different groups, among which Spin Echo (SE), Gradient Echo (GRE), Inversion recovery, and so on. The FLAIR (Fluid-Attenuated Inversion Recovery) and STIR (Short-TI Inversion Recovery) sequences belong to the GRE group. A STIR sequence leads to fat suppression and a FLAIR sequence to fluid signal suppression. Two common MRI sequences are the T1 and T2-weightings (SE group).

A T1-weighted image is obtained by choosing short TR and TE, in order to get more information on M_{xy} . In contrast, long T1 and T2 give a T2-weighted image. Contrast agent such as Gadolinium (Gd) might also be used to increase contrast.

Several other MRI sequences exist (see Figure 2.18 for some examples) but detailing them is beyond the scope of this work. The next paragraph is anyway dedicated to the T2 Fat-Sat ponderation, since it is the MRI sequence of the database images used in this work.

Sequence	T1-weighted	T2-weighted	FLAIR	T2*-weighted
Duration of scan	2–5 min	2–5 min	2–5 min	2–5 min
Sequence type	Spin echo (FSE or TSE)	Spin echo (FSE or TSE); maximum echo train length 14	Spin echo	Gradient echo preferred to ensure comparability between sites
TR	~600 ms	~3200–4000 ms or per institutional protocol	At least 9000 ms	As per institutional protocol
TE	As per institutional protocol	~85 ms, or as per institutional protocol	Recommend 100–150 ms	Recommend at least 12 ms at 3 T and 20 ms at 1.5 T

Figure 2.18: Sequences parameters examples (adapted from [21])

The Fat-Sat technique takes advantage of the chemical shift existing between hydrogen protons of fat and water. At the end of each TR, a low pass band RF pulse centered on the fat resonance peak frequency is applied to get rid of the longitudinal magnetization of the fat. Therefore, only the water hydrogen protons magnetization is tilted in the transverse plane and contributes to the image. This Fat-Sat technique can be applied to any sequence (in the case of this work, T2-weighted MRIs).

2.3.4 Image reconstruction

The image construction from the information collected at the different echos happens in two main parts. First, the Fourier plane is filled with the signal information, and then an inverse Fourier transform is applied to go from the frequency domain to the spatial domain and get the final MR image.

The electric signal collected by the antenna is carrying information from the whole patient volume. To be able to identify which part of the signal corresponds to which part of the body, magnetic field gradients (additional magnetic fields of small intensities) are added during the sequence to achieve spatial coding. The image reconstruction process being very complex, it will not be more detailed.

2.3.5 Advantages and drawbacks

MRI is a very interesting technique because it is non-ionizing, meaning that it has the advantage of not using radiation and thus avoid the side effects related to it. It is also non-invasive. Compared to CT scan, it allows adjacent soft tissues to be better differentiated, which is very useful in HNC, since the head and neck both contain several different organs and tissues that are close on the Hounsfield scale and thus hard to differentiate on the CT scan. Figure 2.19 shows the difference between these two modalities, the GTV being identified more easily.

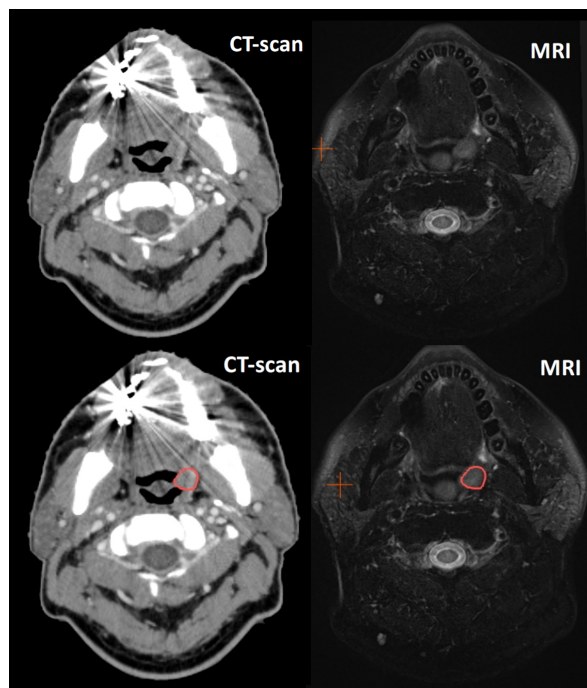


Figure 2.19: Tonsil tumor on MR and CT. The GTVs are delineated in red (images from [9]).

MRI is also indifferent to artifacts that might be created by dental fillings, such as in Figure 2.19. In addition, sagittal, coronal, and oblique images can be obtained directly, which is impossible with radiography and CT, in which only axial slices can be acquired and other planes are computed indirectly. [22]

The MRI technique still has its drawbacks, among the main ones are the high costs related to buying, operating, and maintaining the equipment, and the duration of the exam which can be pretty long (most CT scans are done within 10-15 minutes whereas it takes 30 minutes to an hour and sometimes even more for an MRI to complete). The interaction between strong magnetic fields and cardiac pacemakers is also a disadvantage of the technique. Finally, the duration and noise due to current switching in the coils that are creating the magnetic fields and make it vibrate during the exam can be uncomfortable for some patients who might feel claustrophobic.

2.3.6 MR-LINAC

During the treatment, changes might arise in the patient's anatomy (due to weight loss for instance) and in the tumor size and position (e.g., tumor shrinking). The goal of adaptive radiotherapy (ART) is to take into account these changes by creating a new plan adapted to them during the course of treatment. To monitor these anatomical changes, imaging has to be performed regularly. This is already sometimes the case for some patients who get a Cone Beam CT scan (CBCT) right before each treatment fraction. Generally, the CBCT volumetric image is used to check their position matches the one used for treatment planning. However, it can also be used for online/offline treatment adaptation.

MRI seems to be a more promising imaging modality in the context of HNC. As explained before, it enables improved soft tissue differentiation and it is a non-ionizing imaging technique. It would be especially interesting in the context of adaptive radiotherapy to associate an MRI machine with the linear accelerator that delivers the treatment. It already exists and it is called

a Magnetic Resonance Imaging Guided Linear Accelerator (MR-LINAC). By using the MRI as imaging technique, the radiation oncologist is able to get high-definition images, to have a great contrast between soft tissues, and to see the tumor and organs movements in real-time.

This is why this work is dedicated to trying to automate the GTV segmentation on MRIs, since being able to perform it automatically, accurately, and fast would be of great interest in the adaptive radiotherapy workflow using MR-LINAC. Being able to generate contours on MRIs could also be useful even without a MR-LINAC, since the anatomical information captured by the MRI can be used to delineate on the CT scan. This will be explained with more details in Chapter 4.

2.4 Convolutional Neural Networks

2.4.1 From Artificial Intelligence to Convolutional Neural Networks

To introduce Convolutional Neural Networks, it is useful to first take a look at the big picture in order to understand where these models are located in the field of Artificial Intelligence (AI). Figure 2.20 represents the hierarchy between the different algorithms families, which are briefly defined hereafter:

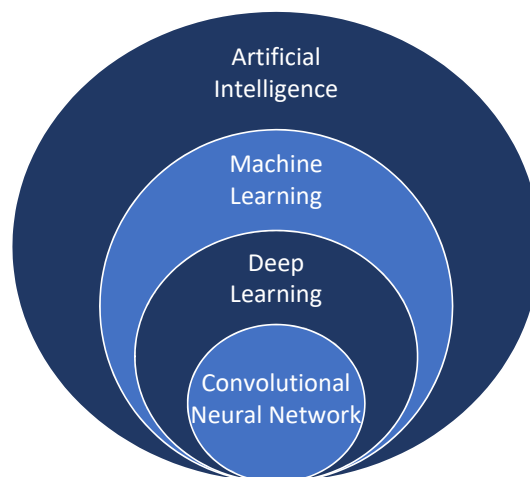


Figure 2.20: Overview of the hierarchy from Artificial Intelligence to Convolutional Neural Networks

- Artificial Intelligence (AI), or Computational Intelligence has many definitions. David Poole, Alan Mackworth and Randy Goebel who wrote one of the first and leading book on AI [23], define it as the study of the design of intelligent agents. They define an agent as something acting in its environment (like a human, an airplane, ...) and acting intelligently by "being doing what is appropriate for its circumstances and goals. It is flexible to changing environments and changing goals, it learns from experience and makes appropriate choices given perceptual limitations and finite computation." Subfields of AI include Machine Learning, Robotics and Natural Language Processing, for instance.
- Machine Learning (ML) can be defined as "the study of computer algorithms that can improve automatically through experience and by the use of data." [24] ML algorithms use data called training data to tune their parameters and become more efficient. They learn

from this data, so they are not explicitly programmed to perform the task they are asked to complete. After that, the method efficiency can be assessed by using it to perform the same task on unseen examples of the so-called test dataset.

- Deep learning (DL) is a subset of Machine Learning based on Artificial Neural Networks (ANN) with features learning. The "Deep" in Deep Learning finds its meaning in the fact that the networks of this type of methods are composed of many stacked layers. More details about ANN and deep networks can be found in subsection 2.4.2.
- A Convolutional Neural Network (aka CNN, or ConvNet) is a subtype of ANN generally used to perform image analysis. Many different architectures exist such as LeNet, VGGNet, U-Net⁴, and so on. They are studied in more details in subsection 2.4.3.

It is interesting to highlight the fact that a lot of the mathematical developments underlying these models and the intuition behind them have existed for decades (in the late 1940s, Warren McCulloch and Walter Pitts developed the first computational model of a neuron). However, their development skyrocketed only recently thanks to the new availability of better computational resources, big databases and storage facilities.

2.4.2 Artificial Neural Networks

Artificial Neural Networks (ANN), or simply Neural Networks (NN) are computational models inspired by the human brain structure, more specifically by the relations existing between its neurons. In fact, an ANN is a collection of units called artificial neurons that are connected with each other just like the neurons are connected thanks to the synapses in the brain. In addition to the neurons and synapses, the other basic components of every ANN are weights, biases, and activation functions. A schematic representation of all these elements is available in Figure [25].

The very first ANN is the Perceptron, it is a single layer NN giving only one output which was used for binary classification. It is represented in Figure 2.21, side by side with a biological neuron to highlight the structural similarity between the two. The Perceptron, just like the brain neuron, takes multiple inputs and can produce an output transmitted to other neurons thanks to a weighted sum and an activation function. This process is detailed in subsection 2.4.2.1.

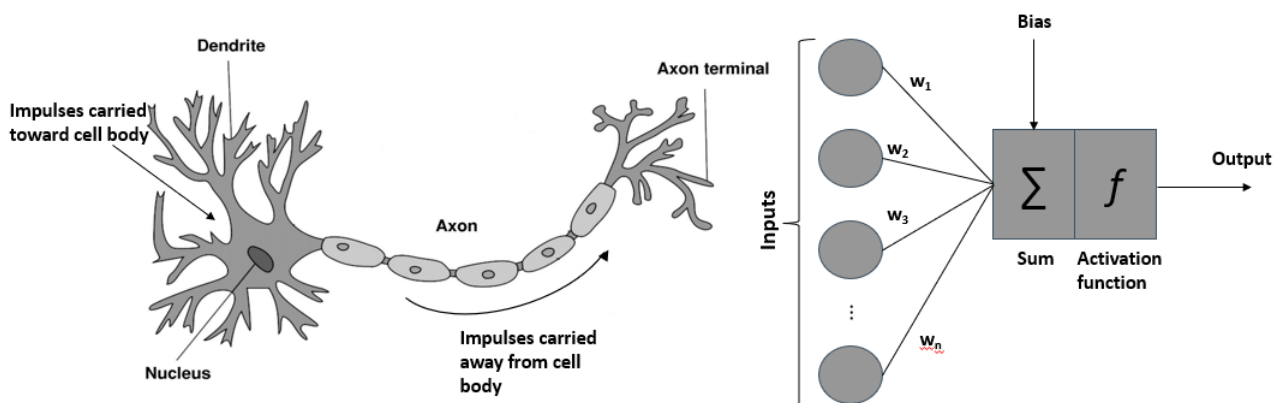


Figure 2.21: Schematic comparison between a biological neuron and the Perceptron (left side adapted from [26])

⁴Which will be studied in details in Chapter 3

The ANNs, which are the evolution of the perceptron, are composed of multiple layers of several neurons. Figure 2.23 shows an example of a classic ANN structure. Overall, they are composed of an input layer (here of m neurons), many hidden layers and an output layer (here of n neurons). The lines show the relationships between the different neurons, while the weights are stored in the W matrices. The back propagation arrow corresponds to an important step of the learning process that will be detailed later. The different components of the ANN are detailed hereafter.

A Deep Neural Network (DNN) is an ANN containing a lot of hidden layers between the input layer and the output layer and a lot of neurons in these hidden layers.

2.4.2.1 Neurons

As briefly explained in the beginning of subsection 2.4.2, an ANN is composed of layers of neurons. To each neuron-neuron connection is assigned a weight reflecting the importance of the connection. A bias is also associated with each neuron.

Each neuron of a layer can trigger and transmit information to other neurons to which it is connected. This information takes the form of a real number. A neuron can get its input directly from data (input layer neurons) or from the outputs of other neurons (hidden layers neurons or output neurons). The value a neuron outputs is the sum of all the inputs it receives weighted by the connections between the different neurons and the neuron of interest, plus the bias term. Finally this weighted sum goes through an activation function to give the final output that will either be passed to further neurons as input or be the final prediction. A fully-connected NN is a network where all the neurons of a layer are connected to all the neurons of the following layer.

2.4.2.2 Activation Function

Choosing an activation function (or transfer function) is critical for the model performance. It defines how the weighted sum of a neuron inputs is modified to produce the neuron output. All activation functions are non-linear. Their application thus aims to add non-linearity to the network so that the network is able to learn complex non-linear relationships between the inputs and outputs. Typically, the same activation function will be used in all the hidden layers, and the output layer will use a different one, depending on the prediction type needed (probability, label, binary output, and so on). Activation functions have to be differentiable in order to be able to perform backpropagation. Among the most popular ones are Rectified Linear Unit (ReLU), Logistic (Sigmoid), and Hyperbolic Tangent.

The ReLU is probably the most commonly used activation function for hidden layers. For any real number z , it is defined as

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

LeakyReLU is a modification of ReLU, in which was added a small slope for negative input values instead of a flat one.

These 2 functions, along with the sigmoid and hyperbolic tangent activation functions, are represented in Figure 2.22

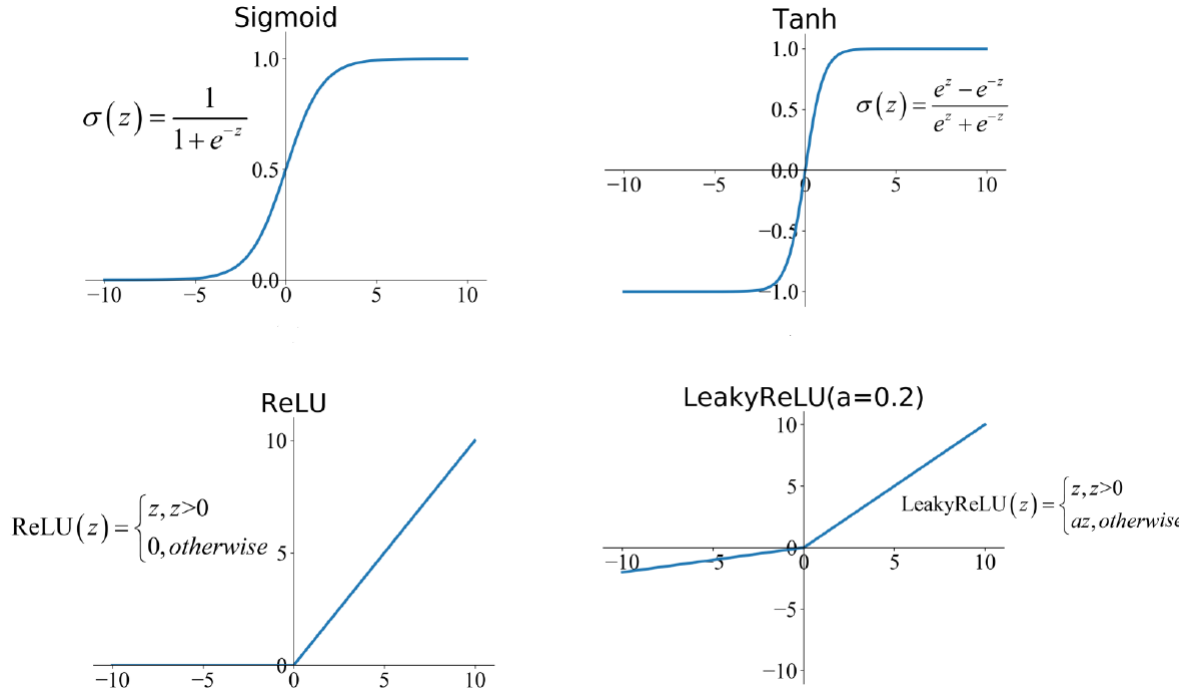


Figure 2.22: Sigmoid, hyperbolic tangent, ReLU, and LeakyReLU activation functions mathematical expressions and graphs (from [27]).

2.4.2.3 Hyperparameters

Hyperparameters are design choices related to the network architecture and the learning process. They include for example the number of hidden layers, the learning rate (i.e., the step size of network parameters change at each iteration).

These choices are extremely important for the performance of the network and require trial and error in classical architecture. An alternative to hyperparameter tuning would be the Automated Machine Learning (AutoML). Indeed, AutoML provides methods for the automation of the labour-intensive and iterative tasks of ML model development (algorithm selection, hyperparameter tuning of models, model evaluation) [28] [29].

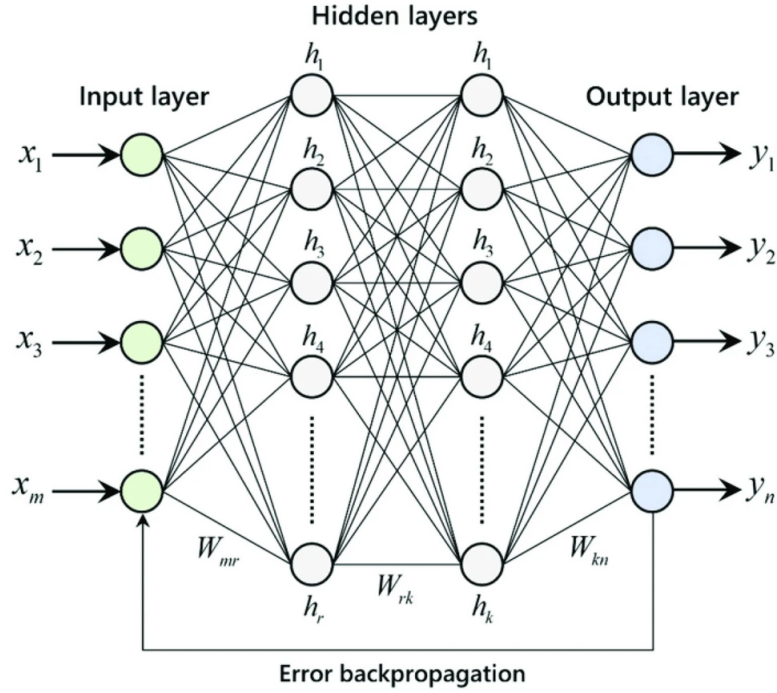


Figure 2.23: Example of Artificial Neural Network (from [25]).

2.4.2.4 Learning types

Several approaches to perform the learning task exist and are subsets of ML. Among them are Supervised Learning, Unsupervised Learning, and Semi-Supervised Learning. In Supervised Learning, which is the learning type considered in this work, a labeled dataset is used. Concretely, the available data is split in 3 sets: the training, validation, and test sets. The NN learns from the training set (it has access to the answers of the problem it is asked to learn to solve). Then the NN hyperparameters are adjusted thanks to the validation set. Finally, the model performance is assessed on the test set, which remains unseen until the testing operation. In Unsupervised Learning, however, the network learns on unlabeled datasets. These models find the hidden patterns in the dataset without the need for human intervention. In this type of learning, it is assumed that the dataset is distributed according to an unknown statistical distribution. The algorithm tries to find this distribution. In Semi-Supervised Learning, a small amount of labeled data is combined to a great amount of unlabeled data.

2.4.2.5 Model training

In this subsection, the network training step is developed in the context of Supervised Learning. During the training process, the NN tries to formulate the relationship between the examples from the training data and the outputs. The output can be of many different types, depending on the task the NN is asked to perform. The two main tasks are classification and regression. For a classification task, the output can be the class label with associated probability and for a regression task it is a numerical output. After the network has predicted outputs for each training object, the difference between the ground-truth and the prediction is computed thanks to what is called a loss function. The result represents the error made by the model (see subsection 4.1.3 for examples of loss functions). The network then adjusts its parameters accordingly to minimize the error thanks to an optimizer (see next subsection). More concretely, the network weights

and biases change during training. After that, the network will go through all the training data again to predict new outputs, recompute the error, readjust and repeat the process until it meets a certain criterion that allows it to stop. This is called the backpropagation.

A k -fold cross-validation is a technique used during the training to fully exploit the available data and to maximize the statistical power of the performance indicator. In a 5-fold cross-validation for instance, which is a very common choice, the dataset is first split into 5 subsets. At the first iteration, 4 subsets are used for network training and the last one for testing. The training is repeated 4 times, the testing set being different at each iteration so that each one is used once for testing and 4 times in training. The performances reported after k -fold cross-validation is the average of the performances obtained at each loop.

Overfitting happens when the model learns too well on the data, leading to difficulties to perform the same task on new data. On the contrary, underfitting happens when the model fails to catch the data properties that allow it to learn how to perform the task.

2.4.2.6 Optimizer

During the optimization process in supervised learning, the optimizer tries to minimise a loss function $L(y, y^*)$ that compares the prediction y^* with the real value y . By minimizing this function, the optimal weights for the neurons in the network are found. A popular optimizer is Stochastic Gradient Descent, which is developed further in section 4.1.3. An example of loss function for segmentation is described in section 4.1.3 too.

2.4.3 Convolutional Neural Networks

Convolutional Neural Networks are a specific type of Deep Neural Network that use convolution operations to be able to perform various tasks on images, including medical images. They may also include pooling operations such as max-pooling and stride, which are detailed below.

2.4.3.1 Convolution operations

An image being just a matrix of numbers (one matrix for each color channel in the case of an RGB image for instance), it is possible to apply convolution operations between this image and different filters that aim to capture image properties. The result of a convolution operation is a matrix where each cell correspond to the sum of the products of the corresponding cells of the image and the filter. Figure 2.24 shows an example of convolution. To fill in the feature map, the kernel will slide of one pixel to the right, repeat the operation until the end of the line, then jump to the next one and repeat (raster scanning). In a CNN, the values in the filter will be adapted during learning.

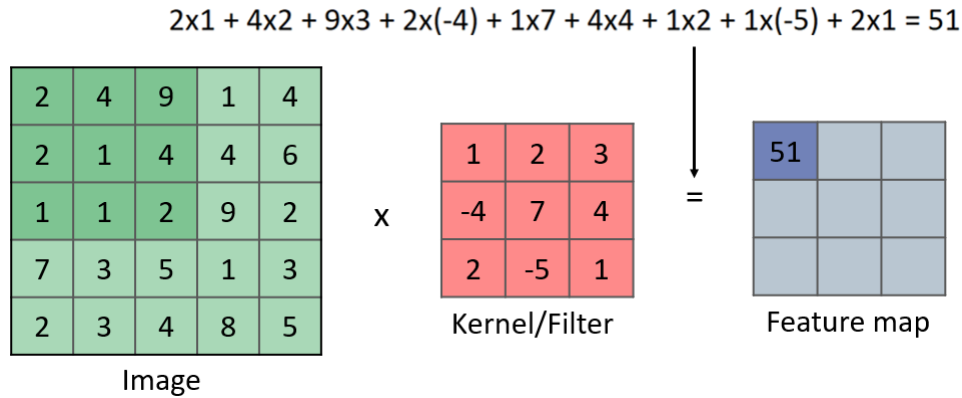


Figure 2.24: First step of a convolution operation between an image and a filter.

2.4.3.2 Pooling operations

Pooling in a NN allows reducing the dimensionality of the data given as input. A popular method is the **max pooling**. It slides a window with a set size on the image and extracts the pixel with highest value. The step size as the window slides across the image is called a **stride**.

The application of the convolutional filters allows patterns to be detected in the image and features to be extracted. Pooling allows the U-Net to learn features in multiple resolutions. In image analysis, the CNN extracts high-level features, like objects, which would totally make sense in the learning process of face recognition for example, to low-level features that do not make sense anymore to the human eye but help the network to perform its task. Figure 2.25 shows an example of different level features in the framework of face recognition.

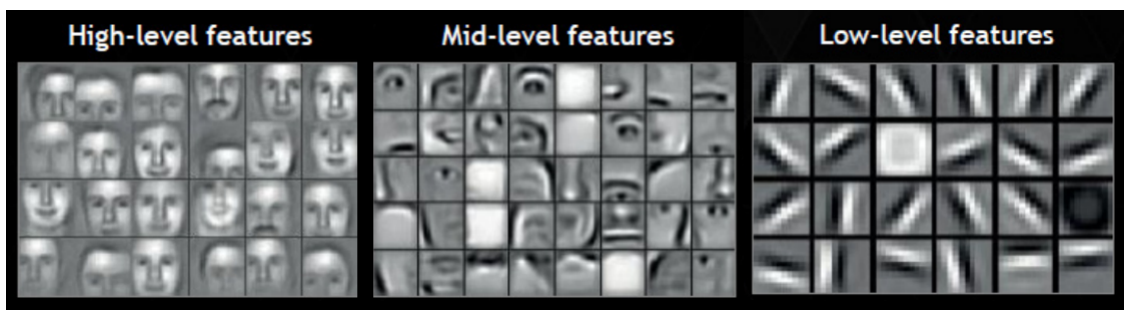


Figure 2.25: Different features level in the framework of face recognition (adapted from [30])

Figure 2.26 shows an example of typical CNN architecture. At each step, convolutions and max-pooling operations are applied to the image. The convolutions leads to features maps and the max pooling allows the dimensions of these feature maps to be reduced, while the number of channels increases because of the multiple filters. The output of the last convolution layer is flattened and fed into a fully connected mini-ANN. The output is generally a probability associated with each pixel.

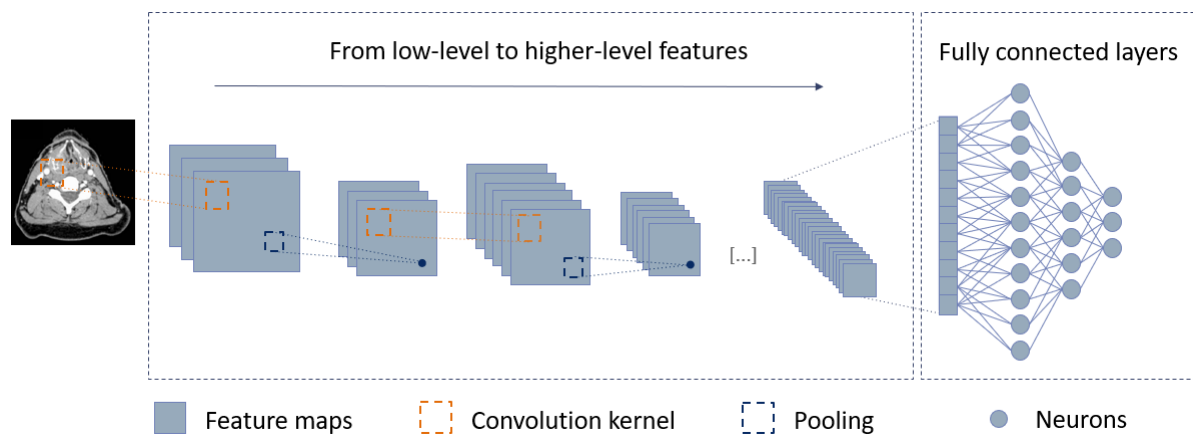


Figure 2.26: Basic example of a Convolutional Neural Network (adapted from [31])

Chapter 3

State of the art in automated segmentation in HNC

This chapter starts with an introductory section on how to assess CNNs performances in the framework of medical objects segmentation. The limitations of the metrics are further discussed in Chapter 5. The second section is dedicated to a literature review on TV auto-segmentation in HNC using Atlas-Based and Deep Learning methods.

3.1 Segmentation evaluation metrics

Before diving in the description and analysis of models currently used to perform auto-segmentation in HNC, it seems useful to describe the metrics that are used to assess their performances and briefly mention what are their advantages, drawbacks, and how they complete each other.

To evaluate a segmentation model means to compare the ground-truth segmentation to the segmentation predicted by the model. In this work context, it is the similarity between the GTV delineated by the radiation oncologist and the GTV predicted by the network that has to be quantitatively assessed. In order to do that, one can look at the level of overlap between the GTVs, but also check the quality of the boundaries and the distance between the ground-truth and predicted GTVs.

Achieving segmentation of a TV or OAR, i.e., labelling each pixel of the image can be considered as a classification problem: either the pixel is correctly labelled or not. This lead to the possibility of using the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) indicators. They respectively refer to the number of voxels correctly classified as belonging to the tumor (or OAR), correctly classified as belonging to the background, wrongly classified as belonging to the tumor (or OAR), and wrongly classified as belonging to the background. A lot of metrics are defined from TP, TN, FP, and FN. Metrics describing the overlap, distance and boundaries quality are described in the next subsections.

3.1.1 Set overlap measures

It is assumed for this subsection and the following ones that G is the point set containing all ground-truth GTV voxels, and P the point set containing the voxels falling into the predicted GTV.

Among the segmentation overlap evaluation metrics, the Dice Sørensen Coefficient (DICE or DSC) is probably the most widely used. It is defined as

$$\text{DICE} = \frac{2|G \cap P|}{|G| + |P|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (3.1)$$

The Jaccard index (JAC) or Intersection over Union (IoU) is defined as

$$\text{JAC} = \frac{|G \cap P|}{|G \cup P|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (3.2)$$

These two indicators are related to each other by

$$\text{DICE} = \frac{2\text{JAC}}{2 + \text{JAC}} \quad (3.3)$$

Therefore using both of them to assess performance is redundant. They are illustrated in Figure 3.1.

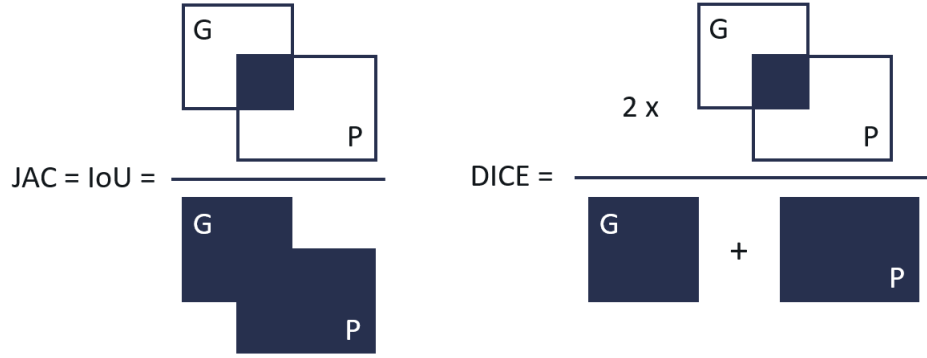


Figure 3.1: Jaccard index and DICE coefficient schematic definitions

The Dice coefficient is often used because it allows direct comparison of the ground truth and predicted segmentation, but one should be careful when performing segmentation of small objects. In this case, the Dice coefficient is not really appropriate because a few pixels difference between the ground truth and prediction can lead to a big drop in the Dice coefficient and thus a bad performance score whereas the model performance is good.

While measuring the overlap between the ground truth and the prediction volumes is a good indicator of the model performance, it is not enough. As shown in Figure 3.2, a notion of distance has to be considered in the performance measurement because two predictions can not overlap, but one might still be much closer to the ground truth than the other.

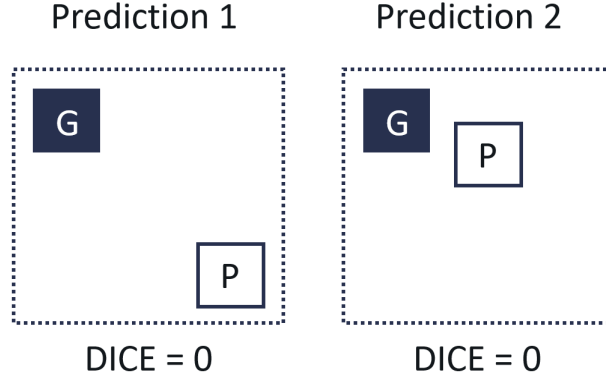


Figure 3.2: Predictions example showing one limitation of using only the Dice coefficient to assess model performance

3.1.2 Set boundary agreement measures

The Hausdorff Distance (HD) allows introducing a notion of distance. The HD between two point sets G and P is given by ([32])

$$HD(G, P) = \max(h(G, P), h(P, G)) , \quad (3.4)$$

where $h(G, P)$ is the directed Hausdorff distance, defined by

$$h(G, P) = \max_{g \in G} \min_{p \in P} \|g - p\| , \quad (3.5)$$

with $\|g - p\|$ being some norm, like the Euclidean distance, for instance.

The directed HD corresponds to the maximum distance between a set of point to the nearest point in the other set. Variants of the HD exist. Among them, the 95% Hausdorff Distance is often used to assess segmentation performance. It corresponds to the 95th percentile of the distances between the points of the 2 sets. Taking the 95th percentile allows to get rid of the impact of a few outliers.

To evaluate the boundaries and thus the shape of the prediction, 3 metrics are commonly used: an adapted Hausdorff Distance, the Average symmetric surface distance (ASSD), and the Surface Dice.

The HD can be adapted to evaluate the boundaries of the segmented volumes. This variant is sometimes referred to as the maximum symmetric surface distance and is defined as the maximum of the Hausdorff directed distances between the boundaries of the ground-truth and auto-segmented regions [33].

The ASSD is the average of all the distances from points on the boundary of machine segmented region to the boundary of the ground truth, and vice versa (definition from [33]).

The Surface Dice is a modification of the Dice. This indicator assesses the surfaces overlap. Indeed, it has the same formula than the Dice but the overlap is defined differently: a predicted surface voxel is considered as overlapping with the ground-truth surface if the closest distance to this reference surface is below a certain threshold. $|G|$ and $|P|$ are the length of the ground-truth and predicted contours in mm in 2D (or surface elements area in mm^2 in 3D).

Many other metrics exist, some of them penalizing over or under-segmentation for example, but this work will be focused on assessing the overlap and boundaries quality.

3.2 Literature Review on automated Segmentation in HNC

While the problem of automated OAR segmentation in HNC has already been extensively studied and that the models show clinically acceptable results [34] [35] [36], fewer studies have been conducted on the segmentation of the target volumes, especially with MRI. This section presents a literature review of the current methods used for CTV and GTV auto-segmentation in HNC on CT scan, MRI, and PET scan, with a special focus on Deep Learning methods, since they are the techniques of interest in this work.

3.2.1 Atlas-Based and Deep Learning Methods

As shown in Figure 3.3 auto-segmentation (AS) can be performed using Deep Learning techniques and Atlas-based auto-segmentation (ABAS) methods.

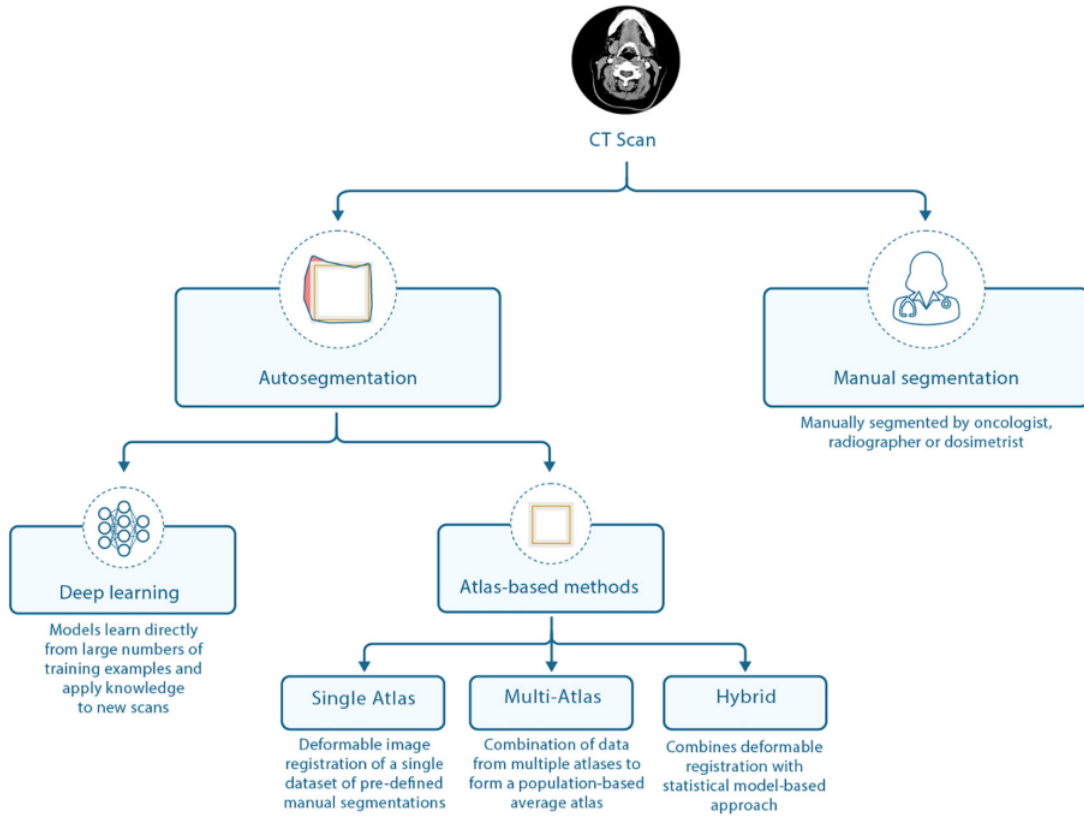


Figure 3.3: Segmentation techniques overview (from [37])

Atlas-based auto-segmentation

ABAS techniques are based on the idea of using already labelled data (atlases) which are typically annotated images to perform segmentation on new data. In fact, an atlas includes information about the locations, shapes, and spatial relationships between anatomical structures. It can be generated by manual segmentation of an image or by combining information from several manually segmented images. Then, a new image can be segmented by mapping the coordinate space of the atlas to that of the image, meaning that registration is performed between the atlas and the image of interest [38]. Once the deformable image registration has been established, the reference contours can be transformed according to the latter.

As shown in Figure 3.3, different ABAS methods exist, the difference between them being how the registration is performed. In Single Atlas techniques, one single dataset of contours is used, whereas in Multi-Atlas methods data from multiple datasets are combined to ensure that the atlas is representative of the anatomical variations that exist among a population. Finally, in Hybrid auto-segmentation (HAS), the registration idea is combined with the knowledge extracted from a statistical analysis of the reference segmentations. By doing so, shape constraints can be imposed and help the segmentation process for example.

Deep Learning models

Deep Learning techniques have already been introduced through the section 2.4, more specifically with the overview of ANN and CNN in subsections 2.4.2 and 2.4.3.

3.2.2 Literature Review for ABAS Methods

In 2007, motivated by the perspective of adaptive radiotherapy, Zhang et al. [39] tried to achieve automatic delineation of Regions of Interest (ROIs) on head and neck CT using an ABAS model. These ROIs comprised the GTV, mandible, brainstem, parotids, and lymph nodes. Thirty two daily CT scans from 7 HNC patients were segmented using the method and compared with manually drawn contours on these patients for model performance assessment. In this study, it is the planning CT of the patient scan and the corresponding contours that were used as the atlas. The CT was then non-rigidly registered to the daily treatment fraction CT and the ROIs were matched accordingly between the two images. The registration was performed thanks to a fast variational-based deformable image registration algorithm.

For most ROIs, the Dice coefficient was about 0.8, and according to the distance evaluation metric, the distance between the ground-truth ROIs and the automatically segmented ROIs was for most of structures below 3mm, with a mean of 1mm and a standard deviation below 2mm.

In 2014, Yang et al. [40] performed segmentation using single and multi-atlas methods and compared them. They built an atlas containing the CTV delineations on the planning CT scans of 6 patients suffering from unilateral tonsil cancers. Ten other patients were used as test set, and 8 oncologists manually delineated those 16 patients. The simultaneous truth and performance level estimation (STAPLE) algorithm was used to combine the 6 atlas segmentations, to perform the registration, and finally to generate contours on the 10 test patients. The STAPLE algorithm relies on the maximum-likelihood estimates of the TP and FN rates of individual segmentations. Moreover, a single atlas that gave the deformed contours best matching the doctor's contours was identified for each patient from the testing set.

Yang et al. observed that when considering a single atlas, no good results were obtained for any of the 10 test patients. In contrast, using multi-atlas AS showed a good agreement between the physician's contours and the AS contours. Actually, they got a median slice-wise Hausdorff distance of 7.4 ± 1.0 mm, Dice of $80.2 \% \pm 5.9\%$ and total volume overlap of $77.8 \% \pm 3.3 \%$ over the 10 test patients.

Many other papers on ABAS techniques exist [41] [42] [43], but they almost all only tackle the problem of OARs, not the GTV nor CTV (no paper was found about the GTV segmentation on MRI). In addition, according to Vrtovc et al. who performed a review of the OARs auto-segmentation methods in HNC from 2008 to 2020 [44], Deep Learning methods have now outperformed ABAS techniques so the remainder of this chapter will be dedicated to an overview of the DL methods.

3.2.3 Literature Review for Deep Learning Methods

Since few articles were found on GTV segmentation in HNC with MR images, this subsection will also review papers reporting how to perform this task on other imaging modalities (CT scan, PET scan) or combinations of several ones.

3.2.3.1 U-Net

A CNN of particular interest because of its popularity in the field of biomedical images segmentation is the U-Net, developed by Ronneberger et al. [45] in 2015. Its architecture is displayed in Figure 3.4 and easily justifies why it is called "U"-Net.

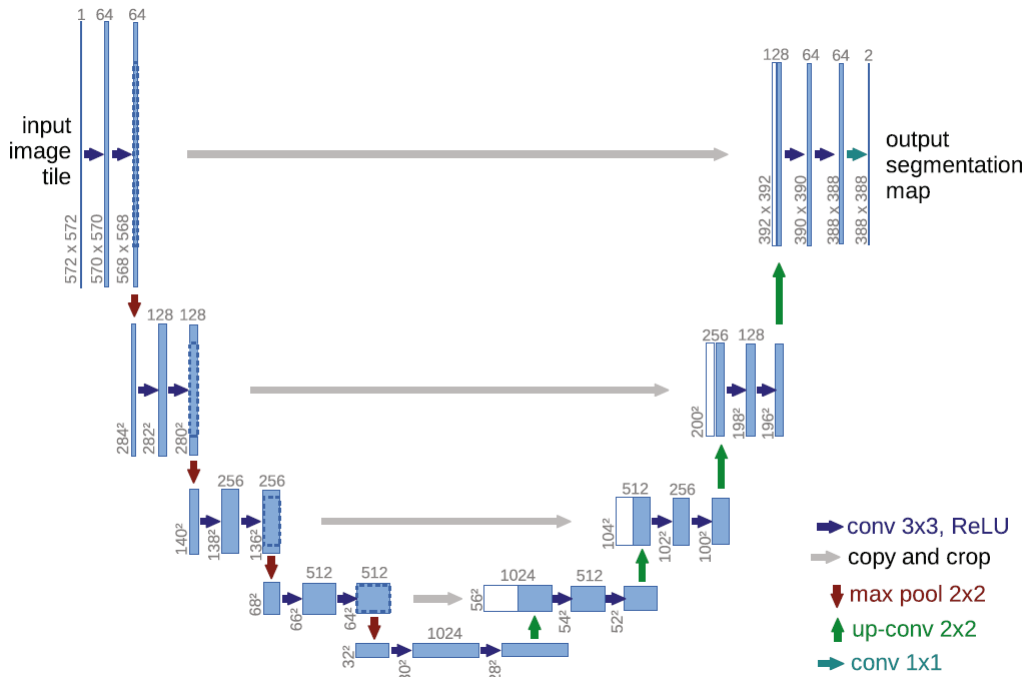


Figure 3.4: The U-Net architecture (from [45]). The blue boxes correspond to multi-channel feature maps, the number of channels being denoted on top of it. The x-y sizes are at the lower left edge of each box. The white boxes represent the copied feature maps.

The U-Net is composed of three main parts: the encoder, the bottleneck and the decoder. As the input image goes deeper in the NN, the dimensions of the feature maps decrease, and the number of channels increases: this is the encoder part, which is the first vertical branch of the "U". It allows features extraction and captures information via the feature maps. It has a classical CNN architecture, composed of repeated convolutions followed by a ReLU activation and a max pooling with stride 2 for downsampling. The bottleneck is the horizontal part, it contains a compressed representation of the input data. The idea is that this compressed view should only contain abstract and high-level information to be able to construct the output. In order to perform image segmentation, the CNN has to output a matrix of class labels for each pixel/voxel and thus get back to original image size and fewer channels. To be able to do this, transpose convolution are applied to the image all along the decoder part of the network (the right branch of the "U"), and the feature maps are concatenated with the corresponding cropped feature maps from the encoder through the skip connections ("copy and crop" in Figure 3.4). This allows reintroducing the spatial information that was lost due to the pooling operations.

Finally, two convolutions each followed by a ReLU are applied. These steps are repeated along the decoder.

3.2.3.2 U-Net based CNN architectures

Many researchers used the original or modified U-Net architecture to perform segmentation. For instance, Cardenas et al. [46] used a 3D U-Net architecture to train 3 models for automatic segmentation of the CTV on oropharyngeal cancers. This architecture is an adaptation of the original 2D U-Net architecture proposed by Cicek et al. [47] to 3D images which enables, among other things, to take advantage of the contextual information from adjacent slices. Figure 3.5 displays the detailed network architecture of this study.

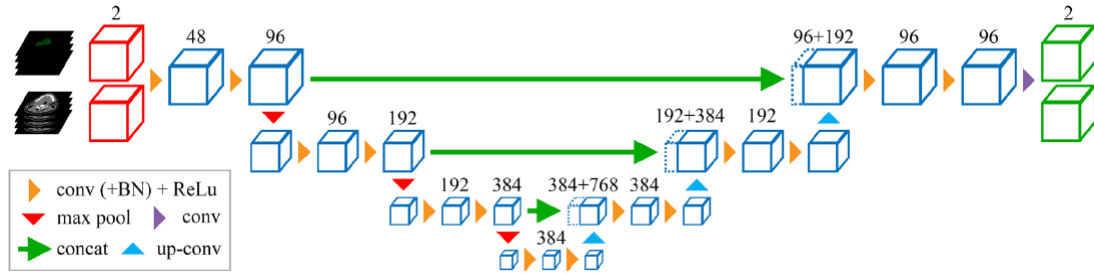


Figure 3.5: 3D U-Net architecture used in Cardenas et al. [46]. BN stands for batch normalization. Operations are 3x3x3 convolutions, 2x2x2 max pooling layers, 2x2x2 up-convolutions.

The networks were fed with two channels: the first one included the planning CT scans for anatomical context and the second one the contours of the GTV. As much as 285 patients were included in the database: 210 for training and 75 for testing. They trained 3 U-Nets: one for tight, one for moderate and one for wide CTVs. These different CTVs correspond to different margins around the GTV. The training was performed using 3-fold cross-validation. The hyperparameters selection was done using a score that takes into account the Dice and False Negative Dice. The model (and its hyperparameters set) with the lowest score is selected for the final predictions.

Feeding the test patients into the networks resulted in a Dice > 0.75 on 96% of cases and 97% showed a mean surface distance ≤ 5 mm to the ground-truth for the tight and moderate models.

Carlos Cardenas published along with other researchers another paper on HNC segmentation using 3D U-Net, but targeting the lymph nodes CTVs this time [48]. They collected the delineated CT scans for 51 training, 10 validation and 10 testing patients. Data augmentation (see subsection 4.3.1.3) was performed among the training patients (translation and rotation). The network contained residual connections and performed batch normalization. Again, the Dice and False Negative Dice were used to compute the loss function. They consider different node levels (IA-V, Ib-V, II-IV and retropharyngeal nodes (RP)), each of them was trained independently. The Dice coefficients for node levels Ia-V, Ib-V, II-IV and RP were respectively 0.9, 0.9, 0.89 and 0.81.

Comparing CNNs performance to the accuracy of contours delineated manually by radiation oncologists is also very interesting to study and is crucial to assess the viability of DL methods implementation in clinical routine. In most studies, DL generated contours are often compared to segmentation performed by only one radiation oncologist, thus ignoring the fact there can be a range of clinically acceptable contours. Wong et al. [49] addressed this problem by comparing

the accuracy of Deep Learning generated contours on OARs and CTV of the central nervous system, head and neck, and prostate (on planning CT scans) to contours of several physicians. Thus, they take into account the inter-observer variability existing among radiation oncologists. They trained a Deep Convolutional Neural Network based on the U-Net architecture with data delineated by one radiation oncologist only. Figure 3.6 shows the Dice (DSC) and 95% HD contours comparison between the physicians (Expert Contours - EC) and the DL based contours (DC). The DC-EC metrics are the metrics evaluated between the DC and different EC for each patient, and EC-EC between the expert contours allowing to quantify the variability between the physicians.

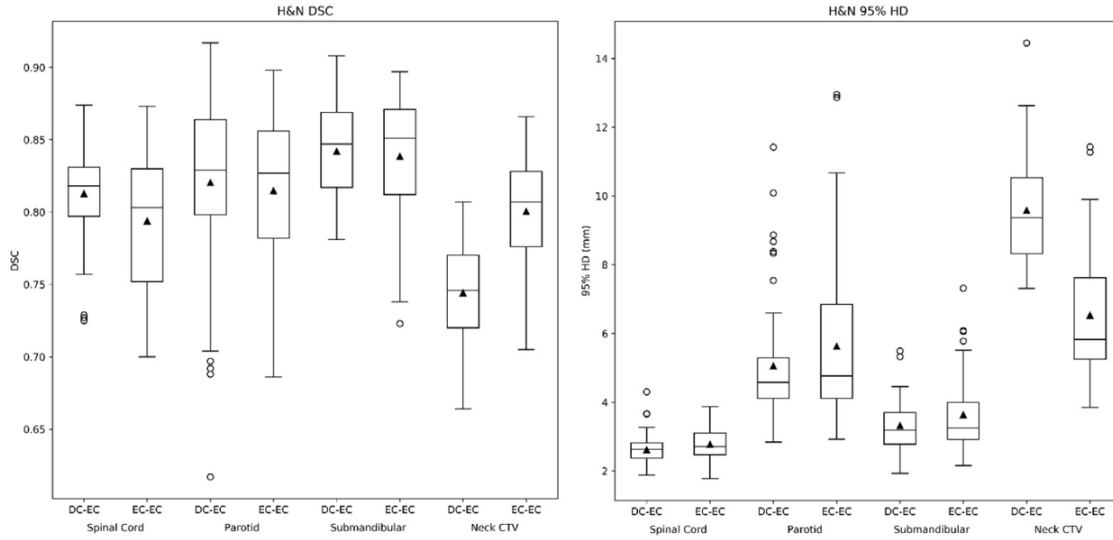


Figure 3.6: DSC and 95% HD evaluated between DL based contours and expert contours (DC-EC) and between experts contours (EC-EC) for HNC patients (from [49])

The results showed minimal differences in Dice and 95 % Hausdorff Distance regarding the OARs, but a bigger difference when comparing CTVs. The authors concluded anyway that the accuracy of automatically generated contours is comparable to expert inter-observer variability.

In June 2021, Ren et al. [50] published an article comparing the GTV segmentation results given by different imaging modalities (CT scan, PET scan, MRI T1 and T2-weighted) and different imaging modalities combinations (CT-PET-MRI, CT-MRI, CT-PET, and PET-MRI). As much as 153 patients were split into training (92), validation (31) and testing (30) sets. The Neural Network used was Residual 3D U-Net, an enhanced version of the original U-Net taking advantage of the Deep Residual Learning to decrease the number of parameters.

All combinations of images including the PET scan reached a Dice of 0.72-0.74, a HD of 8.8-9.5 mm and MSD (Mean Surface Distance) of 2.6-2.8 mm. The CT-MRI fusion showed lower scores (Dice = 0.58, HD=12.9 mm, MSD=3.7 mm). Therefore, the authors concluded that the inclusion of the PET scan is crucial to get good segmentation results.

3.2.3.3 Other CNN architectures

In 2020, Bielak et al. [51] used a CNN to perform GTV-T and GTV-N segmentation on 7-channel multi-parametric MRIs of HNC patients. What is very interesting in their study is that they used the CNN on two different MRI versions to assess the importance of each MRI weighting in the process of segmentation. Indeed, they first fed a reference CNN with the 7-channel multi-

parametric MRI including T2-weighted, T2*, perfusion and diffusion measurements and then they compared the results to those given by the same network, but fed with the MRIs containing one less channel (repeated multiple times, to exclude the channels one by one) in order to identify which MRI contrast contributes the most to the results. The CNN architecture was based on the DeepMedic architecture [52]

The reference CNN reached a Dice of 0.65. The CNN trained without T2* generated the worst results (the Dice dropped of 5.7% for the GTV-T and of 5.8% for the GTV-N) and removing the Apparent Diffusion Coefficient (ADC) input channel had the smallest impact on the DICE coefficient, which decreased of 2.4% for the GTV-T and 2.2% for the GTV-N. Thus, the authors concluded that this sequence could be removed without losing any interesting information for the segmentation task, allowing for reduced scan times in MRI, which are already long.

Guo et al. [53] also combined images, but from different modalities. Indeed, they used combined PET/CT and planning CT scans in order to take advantage of functional and anatomical information, respectively, offered by these two imaging modalities. The database contained 250 HNC patients (140 in training, 35 in validation, 75 in testing). The CNN used (Dense-Net) is based on 3D convolutions with dense connections (every neuron of a layer is connected to every neuron of the adjacent layer by a weight). The results given by the proposed network were compared to those given by 3D U-Net, and the authors also assessed the difference between the segmentation results using only one imaging modality (either the PET/CT or planning CT) or a multi-modality input (both of them).

The proposed Dense-Net performed better than 3D U-Net on multi-modality images (median Dice of 0.73 versus 0.71), and it has less trainable parameters, thus reducing the prediction variability. Multi-modality Dense-Net achieved a Dice of 0.73, whereas it was 0.67 for PET images only (probably due to the heterogeneities in the tumor metabolism on the PET images) and 0.32 for CT scan only.

Chapter 4

Methods

This chapter is dedicated to the description of the two main CNNs that were trained and used to perform GTV segmentation on HNC patients. In the first section, the two architectures are introduced, with a special emphasis on nnU-Net in the first subsection. The second subsection contains a description of nnX-Net, a modification of nnU-Net that was developed during this thesis. Models hyperparameters and design choices that are used in nnU-Net (and nnX-Net) such as the number of epochs, batch size and optimizer (Stochastic Gradient Descent) are explained in the third subsection. In the second section, the database characteristics are described, as well as how patients were split between the training and validation sets. Finally, in the third section, the different network trainings that were run are detailed, along with the slight modifications that were applied to them to try to improve the results.

4.1 Networks Architectures

4.1.1 nnU-Net

nnU-Net is a state-of-the art biomedical images segmentation¹ tool developed by researchers from the German Cancer Research Center. In their paper published in February 2021, Isensee et al. [55] proposed a new method combining the U-Net architecture to an automatic tool for adjusting the model hyperparameters. This is thus a method of automatic machine learning or AutoML. nnU-Net achieved impressive performances, setting a new state of the art in 33 of 53 tasks at international segmentation contests.

In this project, nnU-Net is used to perform segmentation of the GTV only, but the method was initially designed to perform segmentation of several classes, corresponding to different OARs for example.

4.1.1.1 Motivations

The nnU-Net developers wanted to tackle the problem of manual configuration that every Deep Learning application user has to address, in the specific case of biomedical images segmentation. In fact, when configuring a method of deep learning, the user has to make many design choices. They are related to the data pre-processing (image intensity normalization, resampling, ...), the network architecture, the training (choice of the loss function, the patch size, ...), and the post-processing. This leads to a long trial-and-error approach to find the best parameters

¹The method has recently been adapted to medical object detection tasks [54]

combination for which no standardized and clear rules existed. In addition to be cumbersome, this process also makes the results very dependent on each researcher's experience.

These statements are supported by Litjens et al. [56], who realised a global survey on the use of deep learning in medical imaging analysis in 2017. They confirmed that choosing the best set of hyper-parameters is a highly empirical task and that no clear recipe exists to achieve this goal, leaving most scientists with an intuition-based search. They also observed that different research groups who used exactly the same network architectures on the same data got very different results, highlighting the huge importance of good hyper-parameters choices and inter-user variability. Furthermore, the method configuration process has to be repeated for every new dataset to adapt to its specificities. Thus, a network performing well on the data of a specific hospital could hardly be used to perform the same task on the data from another hospital for example. For all these reasons, Isensee et al. wanted to develop a method that automatically makes the best design choices for a given task (here, segmentation) and data.

4.1.1.2 The nnU-Net method

The nnU-Net method relies on two main ideas: extracting useful data properties to generate a "dataset fingerprint", and then list all the design choices that have to be made and split them into three categories.

When asked to perform a new task, nnU-Net first extracts properties from the database available for network training and testing to create the dataset fingerprint. These properties include the image size (number of voxel along each direction) before and after cropping (see Figure 4.1). All images are actually cropped to their non-zero region, because it allows computational efficiency to be improved in some cases.

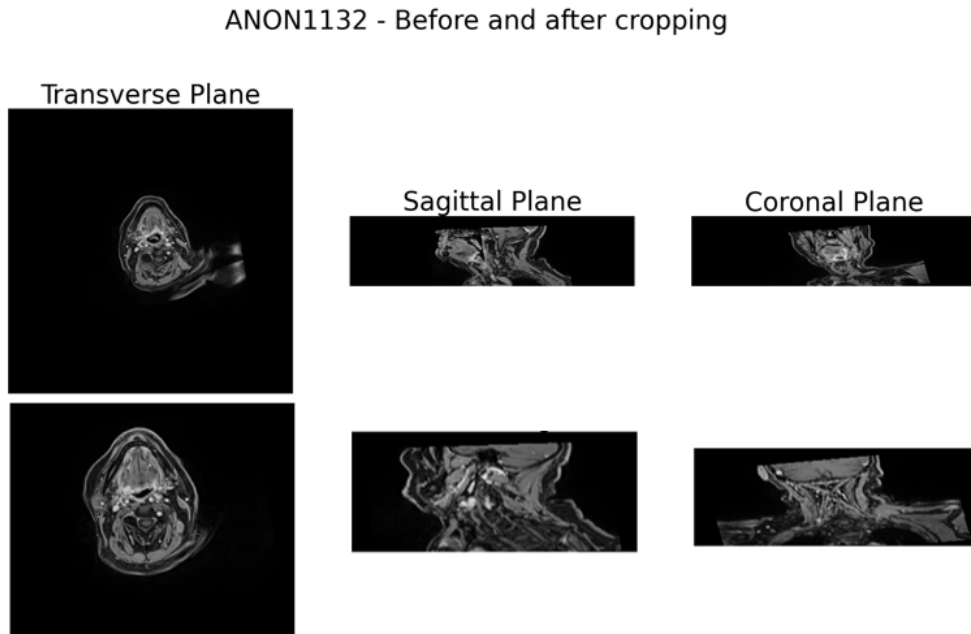


Figure 4.1: MRI of ANON1132 before and after the cropping done by nnU-Net.

The dataset fingerprint also always includes the image spacing (physical size of the voxel), imaging modality (CT scan, MRI, ...), number of classes for all images, total number of training

cases, intensity mean and standard deviation, and finally, the 0.5 and 99.5 percentiles of intensity values in the foreground. The authors also refer to the "pipeline fingerprint" in their paper, which is similar to the dataset fingerprint but for the network. Thus, it includes design choices related to the network.

The second key concept of nnU-Net is dividing all the design choices that have to be made into 3 groups of parameters:

- **The fixed parameters** are predefined parameters that remain the same between different segmentation tasks and datasets. According to Paul Jäger [57], they were decided based on Fabian Isensee’s experience in the field ². They include among others the initial learning rate, the loss function (Dice and Cross-entropy) and the network optimizer (SGD with Nesterov momentum $\mu=0.99$). See subsection 4.1.3 for more explanations about these parameters. A very important fixed parameter is the network architecture. Indeed, they chose the original version of U-Net from Ronneberger et al. [45] to be the CNN architecture. They only brought minor modifications to it, hence the name "nnU-Net", standing for "no new U-Net".
- **The rule-based parameters** are decided based on dependencies that were formulated between elements of the dataset fingerprint and design choices. These rules can just be if-then-else statements or more complex, possibly iterative procedures to find the best parameter. In some way, these dependencies formalise a general domain knowledge that experts in the field have and that was transmitted in an informal way between researchers.

For instance, intensity normalization depends on the imaging modality, image size and image intensity 0.5 and 99.5 percentiles, mean and standard deviation. By default, if the images are not CT scans, each one is normalized by subtracting its mean and dividing by its standard deviation (z-scoring). If the cropping step results in a size decrease of at least 25 %, normalization is applied only on a mask centered in the middle of the image. However, if the dataset is composed of CT scans, the 0.5 and 0.95 intensity percentiles of the foreground voxels are used for clipping, then the foreground intensity mean and standard deviation for normalization. Other examples of rule-based parameters are the patch size and batch size.

- **The empirical parameters** are the remaining parameters that did not fall into one of the two first categories. They are learnt empirically by monitoring the validation data performances. The authors managed to only have two parameters of this type. They are involved in the post-processing, so they do not induce a more intensive GPU work. These two empirical parameters are related to the model selection and connected component analysis. In fact, as explained below, nnU-Net automatically run three different networks and one of the empirical parameters selects the best one. The second empirical parameter takes care of ensembling. Since segmentation is performed voxel by voxel, one lonely voxel could be labelled as belonging to a TV or OAR whereas it is far from it, because the network made a mistake. To avoid that, nnU-Net removes, for each class, all voxels except those forming the largest connected component of the class.

After dataset fingerprint extraction and execution of the heuristic rules, nnU-Net produces three different U-Net configurations: a 2D U-Net, a (full resolution) 3D U-Net and a 3D U-Net cascade (trained only for big images). In the 3D U-Net cascade, a first U-Net learns coarse

²This could be a limitation of the method.

segmentation maps on downsampled data, then a second U-Net refines these maps by operating at full resolution. In 2D models, training and prediction are realised slice by slice. It requires less computational resources but induces a loss of context in the craniocaudal axis, which can lead to errors or weaker performance than a 3D model. After training, nnU-Net chooses the best model, realises ensembling and performs inference on unseen data, if asked to. Figure 4.2 shows an overview of the method.

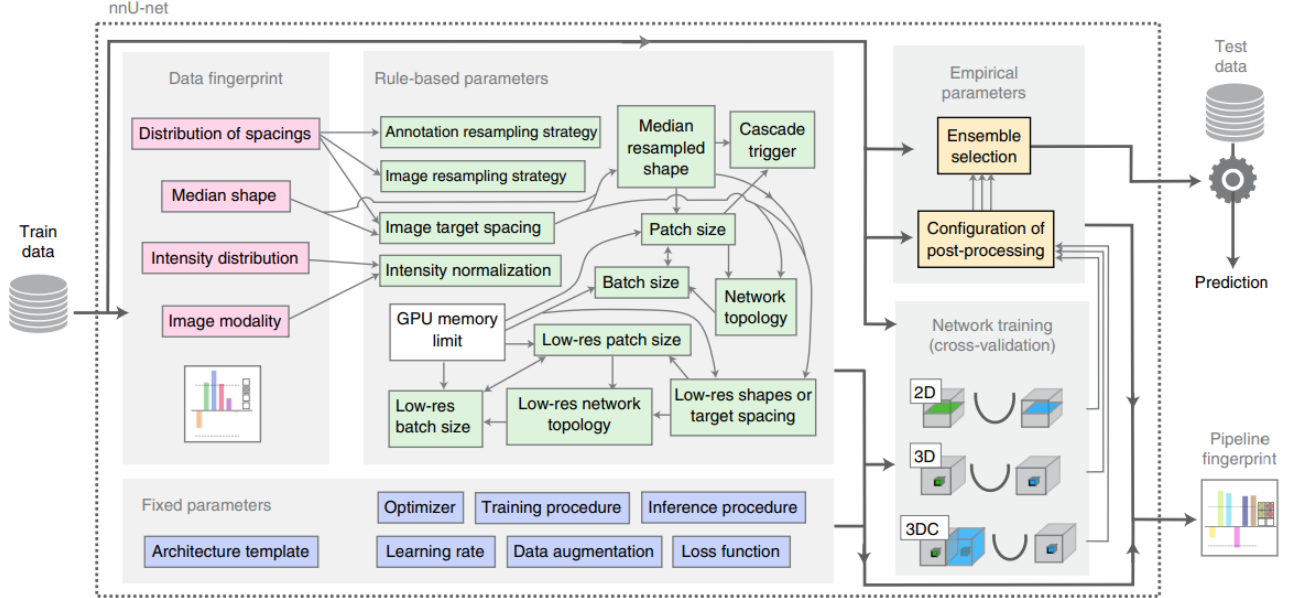


Figure 4.2: Overview of the nnU-Net method (from [55]). For a matter of visual clarity, only some of the most important parameters are represented. Dataset fingerprints are in pink, fixed parameters in blue, rule-based parameters in green, empirical parameters in yellow and interdependencies are shown by the grey arrows.

4.1.2 nnX-Net

During this project, a second network called "nnX-Net" has been developed. It is based on the original nnU-Net and allows combining the information from two images of different imaging modalities even if they are not registered. In fact, in addition to the MR images and their corresponding GTV contours, CT scans of the same patients with GTV contours were also available. In a clinical context, these two sources of information would be used by a physician to perform the contouring. However, as detailed in Section 4.2, MR images were recorded for the diagnosis of patients whereas the CT scans were acquired for the planning, meaning that several weeks had passed between the two images acquisitions. In such cases, simple registration of the images is not accurate since the geometry of the patient is too different. Thus, it was not possible to just input them to the network in two different channels. This novel network aims to address this issue in order to keep using the two modalities.

nnX-Net is actually composed of two 3D nnU-Nets that are merged at their bottlenecks. Its architecture, which gives its name to nnX-Net, is represented in Figure 4.3. A first nnU-Net (on the top) takes the MR images as inputs and a second one (at the bottom) is fed with the CT scans. Then, as in nnU-Net, convolutions and LeakyReLU are applied to go through the layers. Two loss functions are computed: one for the MR and the other one for the CT branches. The

optimizer (always SGD with Nesterov Momentum 0.9) aims to minimize the sum of these two loss functions.

At the bottleneck, the two resulting matrices from the MRI and the CT are concatenated. Combining the information from the two modalities at this step allows to share the important features the networks have learnt regarding the delineations of the GTVs without being biased by the spatial alignment between that does not exist in this case. Indeed, while going deeper in the network until the bottleneck, the spatial resolution decreases and the number of channels increases, on the contrary. After the bottleneck, two decoders operate separately and rebuild, separately, the GTV contours for each modality. The up-convolutions and skip-connections between the encoders and decoders allow to get back to the initial spatial resolution. The final result are two GTV contours (one for the MRI and one for the CT) both influenced by the other imaging modality.

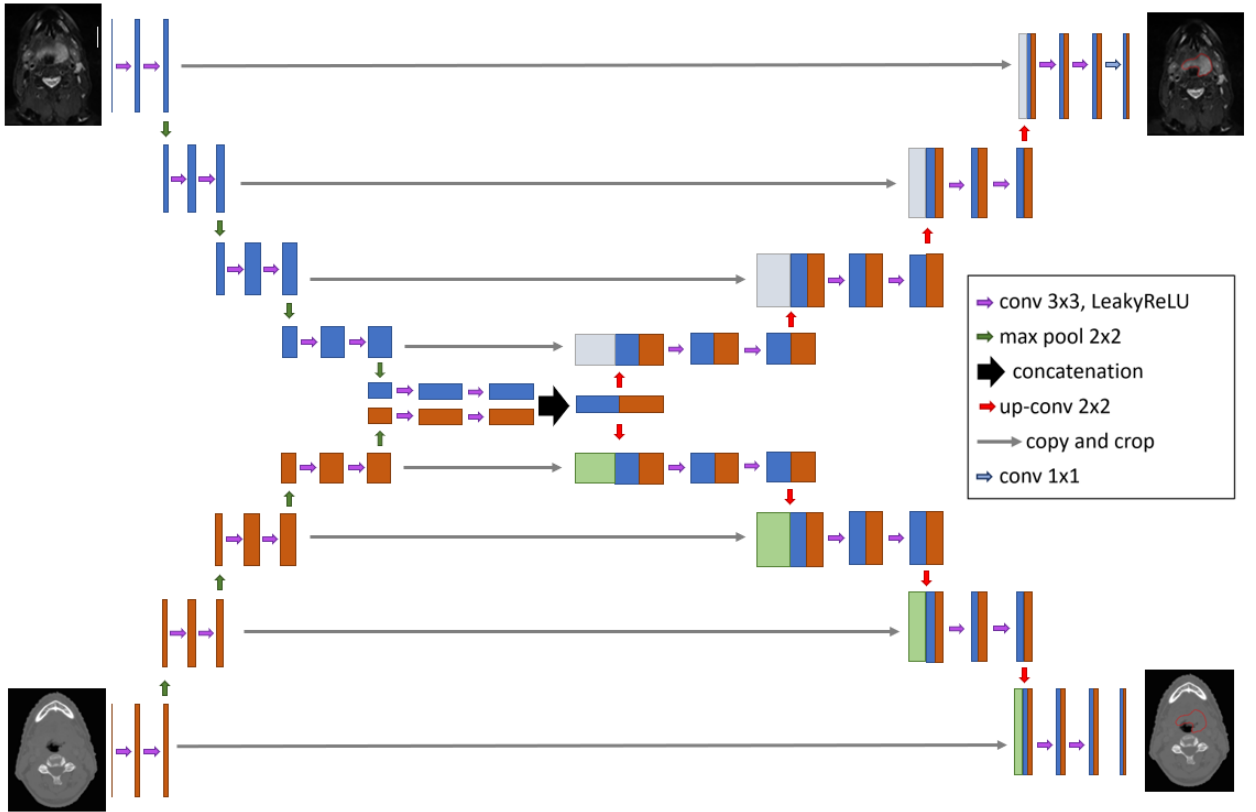


Figure 4.3: nnX-Net architecture

4.1.3 Models hyperparameters

The number of epochs is a parameter setting how many times the algorithm will go through the dataset during training. It should not be confused with the batch size, which is the number of patients that walk through the NN before the internal parameters (weights, biases and so on) are updated. In this work for all models the number of epochs was 1250. Initially, it was a fixed parameter set at 1000, but it was slightly increased to 1250 after observing that the cross-validation curve of the model could possibly pursue its decrease with a bit more epochs. Since this modification improved the Dice of 1% in one of the 3D nnU-Net runs, it was kept for the other model trainings.

In nnU-Net, the loss function is a fixed parameter. Isensee et al. chose the Dice + cross-entropy loss function. The latter is the sum of the binary cross-entropy loss and the Dice loss functions and is defined as [58]

$$L = L_{\text{BCE}} + L_{\text{Dice}} \quad (4.1)$$

$$= \sum_i y_i \log o_i + (1 - y_i) \log(1 - o_i) - \frac{2 \sum_i o_i y_i}{\sum_i o_i + \sum_i y_i} , \quad (4.2)$$

where o_i is the i^{th} output and $y_i \in \{0, 1\}$ the corresponding true label.

The optimizer is an algorithm used to minimize the loss function and thus get the best weights tuning in the NN. A very popular optimizer is the Stochastic Gradient Descent (SGD). At each network iteration, this algorithm computes the gradient of the loss function, i.e., its partial derivatives with respect to the different weights:

$$\nabla L(w_1, w_2, \dots, w_n) = \begin{bmatrix} \frac{\partial L}{\partial w_1}(w_1, w_2, \dots, w_n) \\ \frac{\partial L}{\partial w_2}(w_1, w_2, \dots, w_n) \\ \vdots \\ \frac{\partial L}{\partial w_n}(w_1, w_2, \dots, w_n) \end{bmatrix} \quad (4.3)$$

Since the gradient gives the direction of steepest increase of the loss function, the weights have to be modified in the opposite direction to the gradient to minimize the loss function. The "length" of the steps the weights can make during modification is controlled by the learning rate. This parameter can be tricky to choose, because if the learning rate is too small, the algorithm will take too much time to reach the minimum, but if it is too big, it might jump over the minimum and the model will diverge.

However, as expressed in Equation 4.3, the loss function is composed on sums over all the training samples, so the loss gradient will be composed of very long sum of derivatives. Since the gradient computation has to be performed many times through the training, it would be very inefficient to go through all the training samples to compute the gradient. Therefore, only a certain amount of them, corresponding to the batch size, are chosen to compute the gradient and update the weights accordingly. SGD corresponds to the extreme case where the batch size is equal to one, thus only one patient contributes to compute the gradient. Sometimes in the literature, SGD is used to actually speak of Mini-Batches Gradient Descent, in which the number of patients in a batch is just small.

When using the SGD, at each step of the iterative process, the weights w_j are updated according to the formula

$$w_{j,i+1} = w_{j,i} - lr \frac{\partial L}{\partial w_j}(W, y, \hat{y}) , \quad (4.4)$$

where W is the weights matrix, $j = 1, \dots, n$, i corresponds to the i^{th} step and lr is the learning rate.

In nnU-Net, the optimizer is a fixed design choice, it is SGD with Nesterov Momentum of 0.9. In Momentum optimizers in general, the velocity (v) and momentum (β , $\beta \in [0, 1]$) variables are introduced. $\beta = 0.9$ is a common choice. The Momentum helps achieving faster convergence

and avoiding oscillations. The velocity is initialized at 0, and is computed at each iteration as

$$V_i = \beta V_{i-1} + (1 - \beta) \frac{\partial L}{\partial w_j}(w_j, y, \hat{y}) \quad (4.5)$$

and the weights update is slightly adapted so that

$$w_{j,i+1} = w_{j,i} - lr V_{i,j} . \quad (4.6)$$

Nesterov Momentum is a variation of Momentum in which the gradient is computed on $W - \beta V_{i-1}$ instead of W . Thus,

$$V_{i,j} = \beta V_{i-1,j} + lr \frac{\partial L}{\partial w_j}(w_j - \beta V_{i-1,j}, y, \hat{y}) \quad (4.7)$$

and the weights are updated according to

$$w_{j,i+1} = w_j - V_{i,j} .$$

4.2 Dataset properties and splits

The database includes 20 HNC patients from Saint-Luc University Hospital who underwent a CT scan and MRI as a part of their cancer staging and radiotherapy treatment planning. Different MRI sequences were acquired, among them T2 Fat-sat, a sequence that was introduced in subsection 2.3.3. It is this sequence that will be used in this work³, since it enables a good tumor visualisation. The MR images were acquired using the Siemens MAGNETOM Skyra Fit, which produces a magnetic field of 3T and some patients received Gadolinium as a contrast agent. The MRI spatial resolution are 1x1x2 mm. A slice dimension is 512x512 and the MR volume contains different number of slices depending on the patient. The CT spatial resolution is 1x1x2 mm. The size of a CT slice is also 512x512, and the number of slices also depends on the patient. All the data were previously anonymized and approval from the ethical committee of UCLouvain and Saint-Luc University Hospital was obtained to use the patients in this study.

The GTV segmentation was performed on the MR images by a radiation oncologist and the CT scan contours by other physicians. The delineated GTV volumes are in a range of 2,592 to 100,65 cubic centimeters (cc). An example of MRI, CT and their respective contours for patient ANON3435 can be found in Figure 4.4. The GTV-T is represented in both images. The GTV-N delineation was also available for the CT scan, but it will not be used in this project. Therefore, down from here "GTV" will always refer to the GTV-T.

³Actually, it was noticed afterwards that for a small number of patients, it is FLAIR MRI that was in the database.

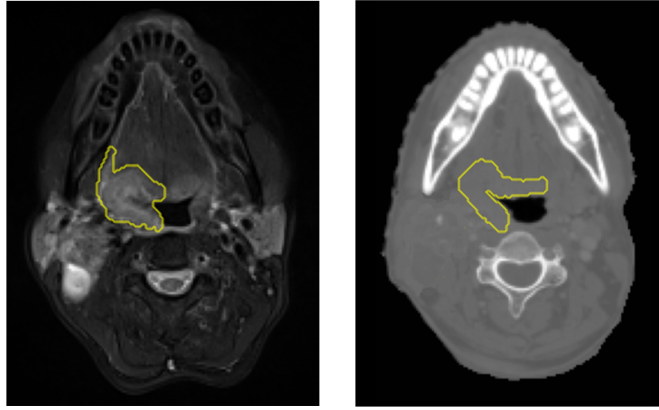


Figure 4.4: MRI and CT scan slices with GTV delineations for patient ANON3435 (slice computed at the center of mass of the tumor).

The patient data were originally in DICOM format. This format is widely used in medical imaging and also includes information about the machine that was used to perform the examination, the acquisition time, the patient's position, and so on. These information can easily be extracted from the file by using "tags" referring to specific properties. Since nnU-Net requires NumPy files as input files, the conversion from DICOM to NumPy was performed. Before the execution of the nnU-Net method, the data are converted into the NIfTI format, which is also the network output format. The NIfTI format (Neuroimaging Informatics Technology Initiative) is a format usually used to store brain MR images. A NIfTI file is simpler than a DICOM file, but still contains the essential data in a simple header followed by raw data. It is faster to load, process, and store. The NIfTI predictions were then converted into NumPy to visualise them and compute metrics using Python functions.

The 20 patients were randomly divided into training and validation sets composed of 16 and 4 patients, respectively. The nnU-Net method, in this particular case, allows assessing its performance on the validation patients as if they were test patients. Indeed, the validation patients are usually used to tune some network parameters. Yet, nnU-Net automatically chooses them thanks to the fixed and rule-based parameters, the model performance can be assessed on the validation patients. The two parameters relying on validation patients are the choice of the best performing model (2D, 3D, or 3D cascade) and the ensembling. In this work, the network choice is fixed before training. In addition, the Dice difference between the raw and post-processed results (i.e., after ensembling) was very small for all patients, so it was decided that the CNN performance could be evaluated on the validation patients.

The database does not contain a great number of patients and 4 validation patients is not enough to draw conclusions about the overall performance of the model. Therefore, the networks were independently run multiple times with different partitions of the database. 5 different splits were run with 5 different validation sets of 4 patients so that each patient was once part of the validation set, as represented in Figure 4.5.

SPLIT 0	ANON590	ANON1132	ANON3620	ANON3648	ANON242	ANON3146	ANON3207	ANON4232	ANON690	ANON866
	ANON2971	ANON4641	ANON1677	ANON2453	ANON3435	ANON3647	ANON1783	ANON2834	ANON2865	ANON4297
SPLIT 1	ANON590	ANON1132	ANON3620	ANON3648	ANON242	ANON3146	ANON3207	ANON4232	ANON690	ANON866
	ANON2971	ANON4641	ANON1677	ANON2453	ANON3435	ANON3647	ANON1783	ANON2834	ANON2865	ANON4297
SPLIT 2	ANON590	ANON1132	ANON3620	ANON3648	ANON242	ANON3146	ANON3207	ANON4232	ANON690	ANON866
	ANON2971	ANON4641	ANON1677	ANON2453	ANON3435	ANON3647	ANON1783	ANON2834	ANON2865	ANON4297
SPLIT 3	ANON590	ANON1132	ANON3620	ANON3648	ANON242	ANON3146	ANON3207	ANON4232	ANON690	ANON866
	ANON2971	ANON4641	ANON1677	ANON2453	ANON3435	ANON3647	ANON1783	ANON2834	ANON2865	ANON4297
SPLIT 4	ANON590	ANON1132	ANON3620	ANON3648	ANON242	ANON3146	ANON3207	ANON4232	ANON690	ANON866
	ANON2971	ANON4641	ANON1677	ANON2453	ANON3435	ANON3647	ANON1783	ANON2834	ANON2865	ANON4297

Figure 4.5: Database partitions - Blue=training patients, Orange=validation patients

To further assess the model performance, other HNC patients who were available but not delineated were used for inference. The predicted GTV were then qualitatively evaluated by the radiation oncologist who delineated the 20 training and validation patients.

4.3 Network Trainings

Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (FNRS). More specifically, models have been trained on the Dragon2 cluster, a computer cluster containing 2 GPUs, more precisely 2 NVIDIA Tesla V100 with 16GB HBM2 (second generation of High Bandwidth Memory).

4.3.1 nnU-Net trainings

4.3.1.1 nnU-Net 2D

First, the nnU-Net 2D model was individually trained 5 times, with the different patient splits as explained in Section 4.2. The input channel was composed of the patients MRI slices (2D matrices containing different levels of grey) and the slices of the GTV mask. This mask contains 1 or 0 associated with the corresponding pixel in the MRI. If the pixel belongs to the GTV, its corresponding value in the GTV mask is 1, if not it is 0.

For the 2D model, batch size was 12, meaning that 12 slices are fed into the network before the loss gradient is computed and the internal parameters are updated to minimize the loss. For the SPLIT 0 configuration (see Figure 4.5 for more details), run time was 07h 46m 01s, which was representative of the run times of the other splits.

4.3.1.2 nnU-Net 3D

Then, 5 3D nnU-Nets were trained with the same 5 splits. For the 3D nnU-Net, the batch size was 2. The input channel contained the patients MRI and the GTV mask, both in 3D this time. Run time was 09h 21m 44s for SPLIT 0, which was representative of the run times of the other splits.

After the 5 model trainings, the model trained with SPLIT 3 was used to perform inference on 8 patients. This model was chosen because it gave the best mean Dice on the validation patients. This time, training and inference took 09h 25m 55s.

4.3.1.3 Data augmentation

Data augmentation is a technique used to artificially increase the amount of data which is especially useful for small databases. It is performed by creating copies of the original data (in this case both the MRI and the GTV contours) that underwent transformations such as geometric transformations (image flipping, cropping, rotation, translation), kernel filters application (image blurring or sharpening), and so on.

In addition to increasing the size of the dataset, data augmentation also helps preventing overfitting. This phenomenon happens when the model learns "too well" on the data and has trouble generalizing on unseen data.

Nevertheless, one should be careful when applying data augmentation because it is not suitable for every task that a NN might be asked to learn and perform. For instance, it would not make sense to apply cropping in a dose prediction problem but since segmentation is a geometric problem, it is completely appropriate in this work.

nnU-Net always performs data augmentation, it is part of the fixed parameters. On all datasets, rotation, scaling, Gaussian noise and Gaussian blur applications, modifications of brightness and contrast, gamma augmentation and mirroring are performed. These transforms are applied to the images with different probabilities. In a new 3D nnU-Net, those probabilities were augmented in the source code and the model was run to see if it enhanced the performances compared to the 3D nnU-Net with the original data augmentation parameters.

4.3.1.4 Adding of the CT as a channel

A 3D nnU-Net was trained with both MR images and CT scans and their GTV masks as input channels. Worse results were expected since the MRI and CT scans were not registered. In fact, this training was more of an exercise to confirm this hypothesis and highlight the importance of spatial alignment between images, and to get familiar with the training of nnU-Net with multi-channel inputs.

4.3.2 nnX-Net training

nnX-Net was trained once with SPLIT0, only in 3D. It took more time (21h 38min 00s), which was expected. In fact, multithreading, which was originally included in the nnU-Net code, was removed. Multithreading is the faculty of a GPU (or CPU) to supply several threads of execution simultaneously. As mentioned, it was removed during the development of the nnX-Net code to make it easier, but could easily be reintroduced in a new enhanced version of nnX-Net.

4.4 Performance assessment

Once the predicted GTV masks were available, the results were first qualitatively analyzed by using Mango, a visualisation tool developed by the Research Imaging Institute from the University of Texas Health Science Center. Then the masks were converted to NumPy arrays to be processed and thus display graphs and compute metrics.

Three main metrics were computed for each predicted GTV of each model: the Dice, the Hausdorff distance and the Surface Dice (threshold of 2mm). They are detailed in Section 3.1.

The individual model performances were assessed, as well as the correlation between Dice and GTV volumes, HD and GTV volumes, how the Surface Dice add information to the Dice, and so on.

Chapter 5

Results and Discussions

This chapter begins with a section displaying the nnU-Net results in tables, graphs and boxplots for the different configurations that were developed in the previous Chapter. The second section presents the results of the nnX-Net training. The third section discusses these results and studies the correlation between variables such as the GTV volume and DICE. In the fourth sections, limitations of this work experiments are presented, focusing first on the limitations inherent in the CNN and then developing the inter- and intra-observer reliability. Finally, the fifth section introduces a few perspectives on what could be tried to enhance the results, and potential new methods for biomedical images segmentation in general.

5.1 nnU-Net results

In this section, results (Dice, HD, Surface Dice, and contours displayed on the MRI) for nnU-Net 2D, nnU-Net 3D and inference on non-delineated patients are presented.

5.1.1 nnU-Net 2D

The results for the nnU-Net 2D model are displayed in Table 5.1. Over the 20 patients, the mean Dice, HD and Surface Dice are 0.47 ± 0.25 , 15.95 ± 3.94 mm and 0.42 ± 0.21 , respectively. The median values for these metrics are 0.53, 16.19 mm and 0.4, respectively.

Patient	Dice	HD (mm)	Surface Dice
ANON242	0.35	18.38	0.28
ANON590	0.56	13.34	0.43
ANON690	0.57	12.17	0.37
ANON866	0.06	5.83	0.02
ANON1132	0.19	15	0.19
ANON1677	0.73	15.96	0.67
ANON1783	0.6	12	0.53
ANON2453	0.2	21	0.27
ANON2834	0.38	17.46	0.29

ANON2865	0.61	19.57	0.46
ANON2971	0.42	20.54	0.21
ANON3146	0.09	16.61	0.14
ANON3207	0.82	15.78	0.76
ANON3435	0.74	18.68	0.65
ANON3620	0.8	13.96	0.7
ANON3647	0.5	22.05	0.44
ANON3648	0.17	10.2	0.27
ANON4232	0.76	18.03	0.71
ANON4297	0.23	16.4	0.32
ANON4641	0.68	15.97	0.62

Table 5.1: Dice, HD and Surface Dice scores for the 20 patients - nnU-Net 2D

Figure 5.1 shows the difference between the ground-truth GTV volumes and the predicted GTV volumes.

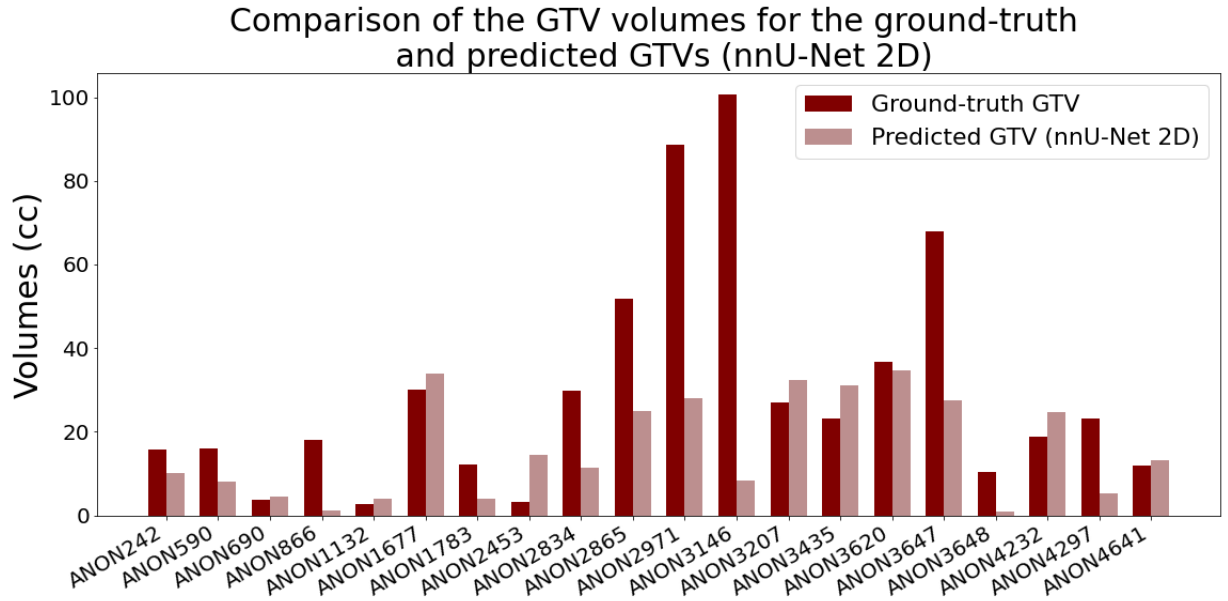


Figure 5.1: Ground-truth and predicted GTV volumes for nnU-Net 2D

Since the nnU-Net 3D model performed better, the display of the 2D results will not go further.

5.1.2 nnU-Net 3D

5.1.2.1 General results

The raw metrics results for the 3D nnU-Net network are gathered in Table 5.2.

Patient	DICE	HD (mm)	Surface DICE
ANON242	0.51	22.02	0.48
ANON590	0.85	11.22	0.88
ANON690	0.83	9.8	0.89
ANON866	0.53	12.41	0.37
ANON1132	0.4	23.28	0.3
ANON1677	0.78	15.23	0.74
ANON1783	0.66	12.96	0.62
ANON2453	0.31	16.55	0.35
ANON2834	0.69	18.38	0.63
ANON2865	0.71	19.36	0.51
ANON2971	0.46	20.71	0.27
ANON3146	0.21	14	0.19
ANON3207	0.85	15.68	0.83
ANON3435	0.81	14.18	0.75
ANON3620	0.77	17.97	0.63
ANON3647	0.56	21.45	0.47
ANON3648	0.16	22	0.23
ANON4232	0.84	13.15	0.88
ANON4297	0.64	20.04	0.53
ANON4641	0.58	10.1	0.58

Table 5.2: Dice, HD and Surface Dice scores for the 20 patients - nnU-Net 3D

The mean Dice, HD and Surface Dice for the 20 patients using nnU-Net 3D are 0.61 ± 0.21 , 16.52 ± 4.25 mm and 0.56 ± 0.23 , respectively. The median Dice, HD and Surface Dice are 0.65, 16.12 mm and 0.56, respectively. The boxplots in Figure 5.2 represents these median scores, and allows to visualize the metrics distributions.

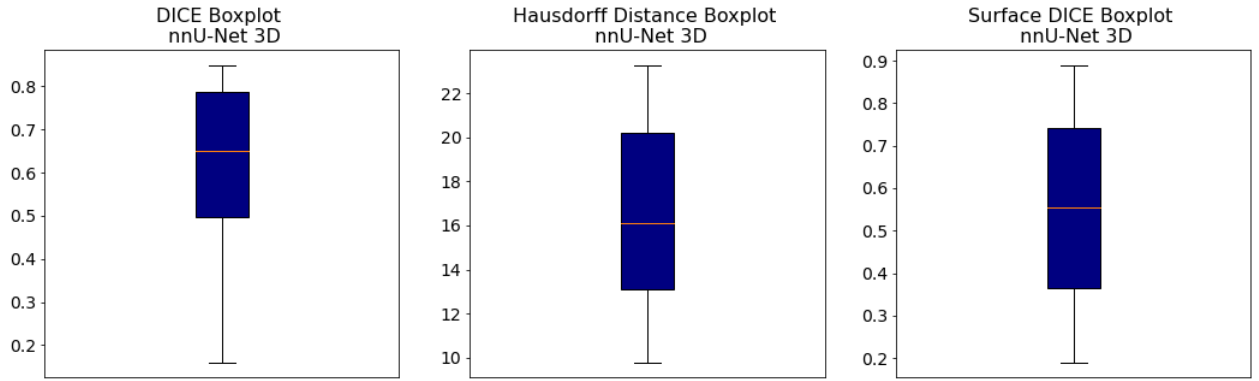


Figure 5.2: Boxplots for Dice, HD and Surface Dice - nnU-Net 3D

In Figure 5.3, the GTV volumes distribution is represented, as well as the corresponding Dice values. In the discussions subsection, the relation between these two quantities is discussed. The patients are sorted by ascending Dice scores.

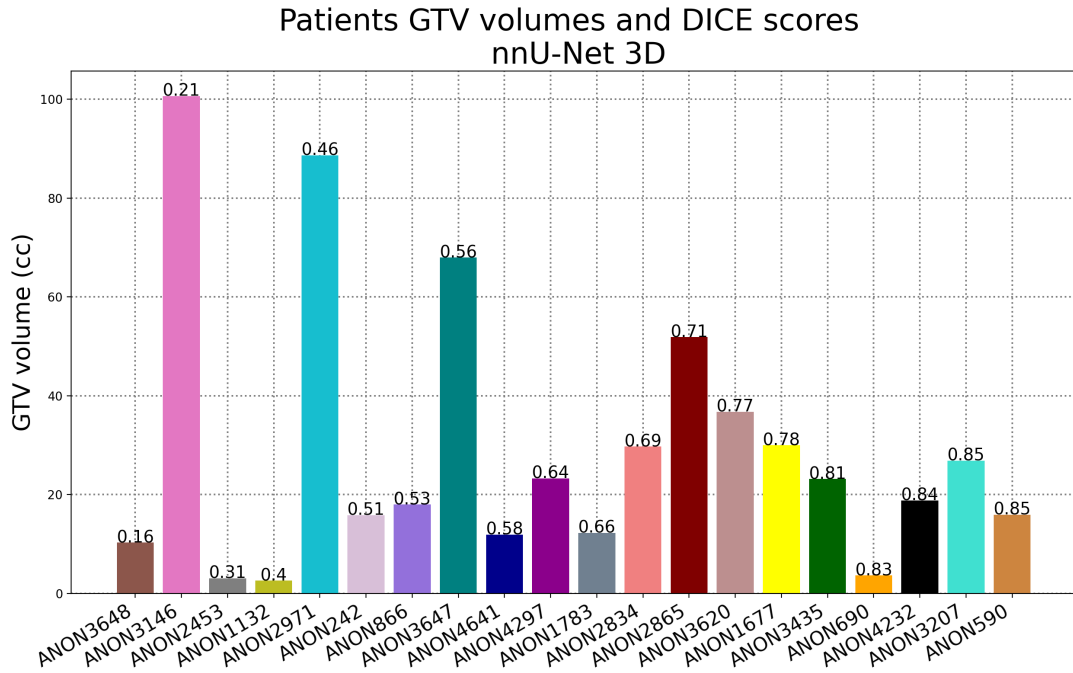


Figure 5.3: Patient's GTV volumes in cc and Dice scores using nnU-Net 3D (sorted by ascending Dice)

Figure 5.4 displays the difference between the volumes of the ground-truth and predicted GTVs.

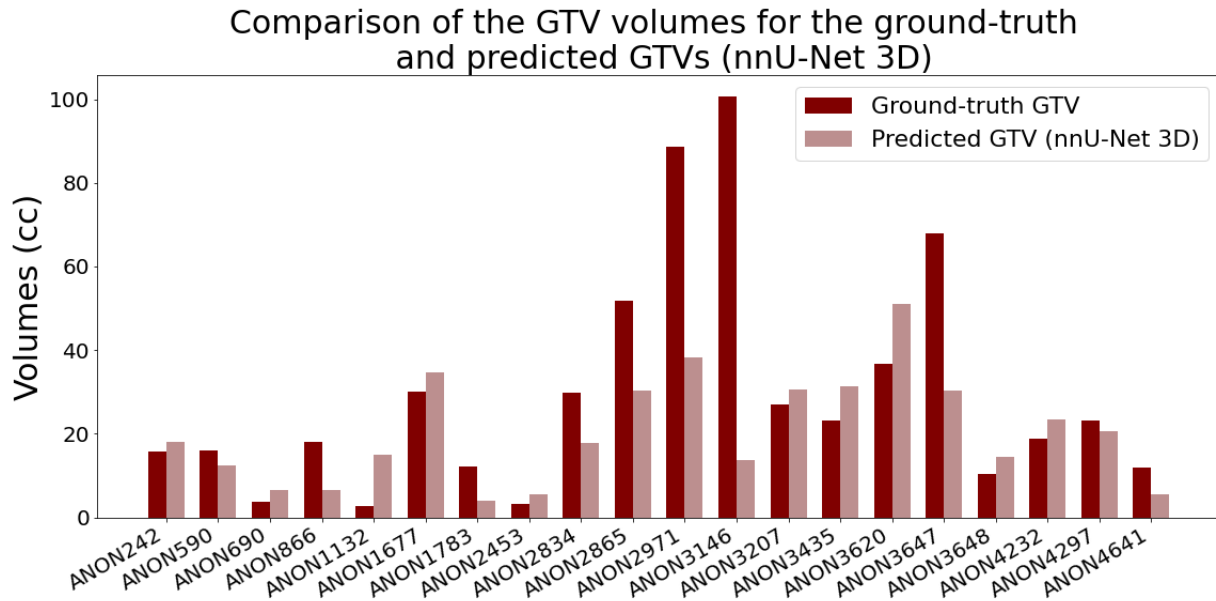


Figure 5.4: Ground-truth and predicted GTV volumes for nnU-Net 3D

The 3 following figures allow to visualize the contours on 3 example patients. The ground-truth GTV contours are represented in yellow, and the predicted contours in red. The slices that are shown correspond to the center of mass of the ground-truth GTV. Figure 5.5 displays an example of bad prediction (Dice = 0.16), Figure 5.6 an example of moderate result (Dice = 0.58) and Figure 5.7 an example of very good prediction (Dice = 0.85). These images are available for all patients in Appendix A.

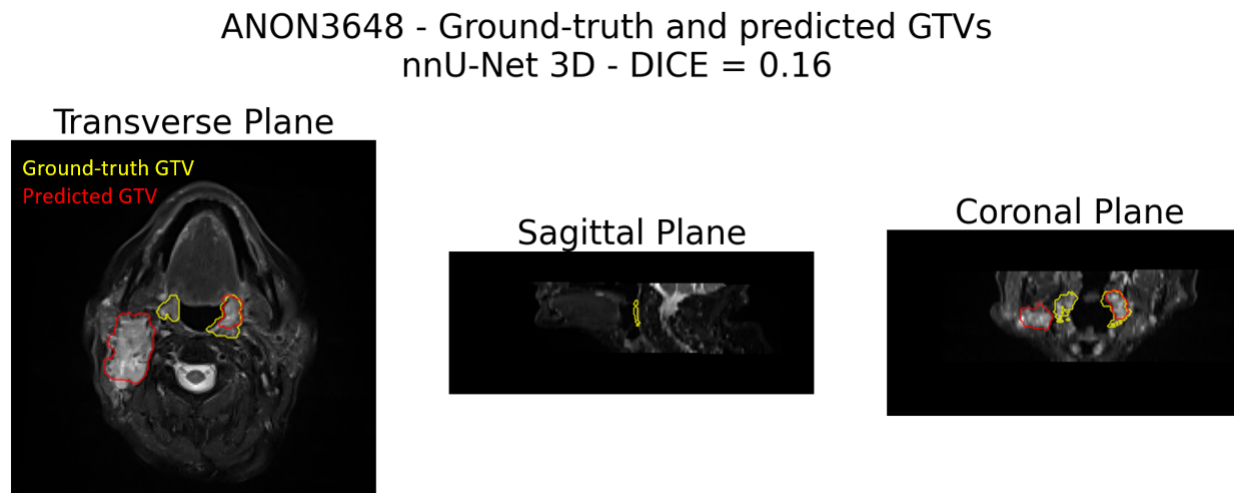


Figure 5.5: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3648.

ANON4641 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.58



Figure 5.6: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON4641.

ANON590 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.85



Figure 5.7: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON590.

In order to compare them and see how well they complete each other, the Dice and Surface Dice scores are plotted in Figure 5.8 for the 20 HNC patients. The Dice and Surface Dice corresponding to the same patient are connected with a black line.

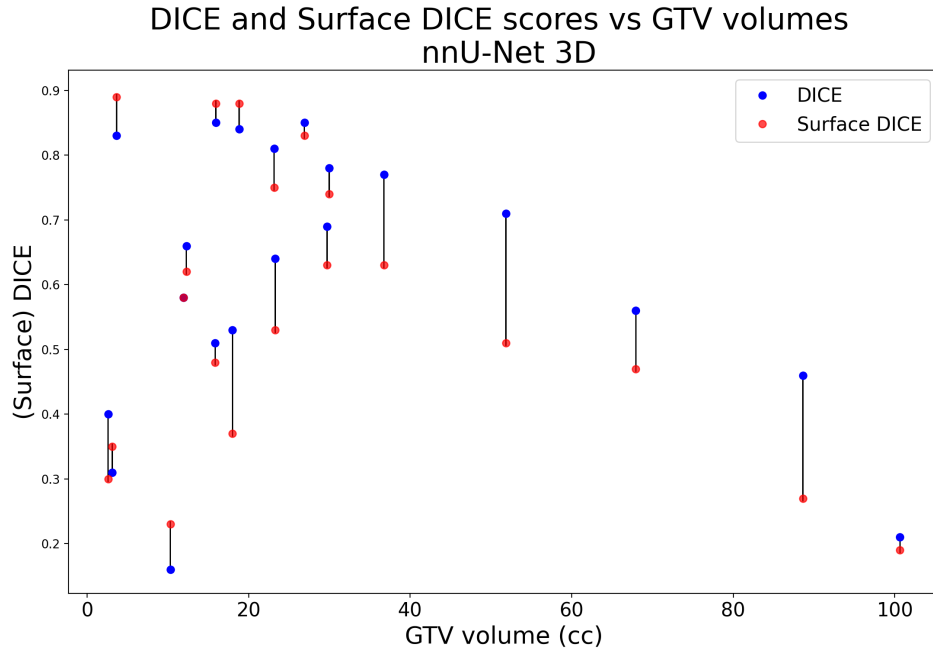


Figure 5.8: Patients Dice and Surface Dice scores ordered by GTV volumes. The black lines connect the two metrics corresponding to a patient.

Figure 5.9 represents the HD for the different patients with respect to their GTV volumes.

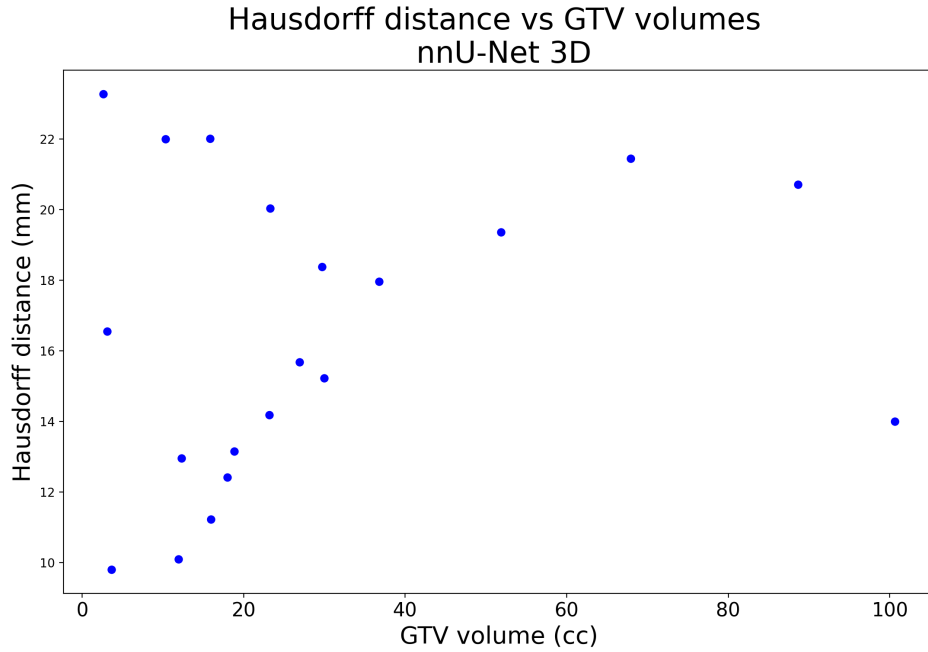


Figure 5.9: GTV volumes and Dice scores using nnU-Net 3D

In order to identify potential outliers that might bias the HD score, the 95 % Hausdorff Distance was also computed. It is displayed along with the total HD in Table 5.3.

Patient	HD (mm)	HD 95% (mm)	Patient	HD (mm)	HD 95% (mm)
ANON242	22.02	22.02	ANON2971	20.71	20.71
ANON590	11.22	10.77	ANON3146	14	14
ANON690	9.8	9.79	ANON3207	15.68	15.68
ANON866	12.41	12.24	ANON3435	14.18	14.17
ANON1132	23.28	23.28	ANON3620	17.97	17.97
ANON1677	15.23	15.23	ANON3647	21.45	21.44
ANON1783	12.96	12.96	ANON3648	22	22
ANON2453	16.55	16.3	ANON4232	13.15	12.88
ANON2834	18.38	18.38	ANON4297	20.04	20.02
ANON2865	19.36	19.36	ANON4641	10.1	10.1

Table 5.3: Hausdorff distance and 95 % Hausdorff distance for all patients (nnU-Net 3D) - HD 95% is in bold when it is different from HD.

5.1.2.2 Inference patients

As explained in the previous chapter, the original database contained MR images that were not delineated. Seven of them were selected to make inference with the nnU-Net 3D network trained with SPLIT 3. The predicted contours on these new patients were then qualitatively assessed by a radiation oncologist. The MR images with predicted contours are available in Appendix B.

For patient ANON3820, the network included air cavities in the GTV contours, which is wrong. However, the MR image was blurry and the physician confessed that because of this poor image quality, it would have been hard for him to contour the patient properly. This might explain why the network was mistaken.

On the contrary, for ANON3454 the network performed a very good contouring. There may be a slight overestimation of the GTV but it was still totally acceptable for a physician.

For ANON2839, the contours were too far in the anterior direction. In fact, the tongue was partly included whereas it should not.

For patient ANON2963 as well as for ANON3820, the network included in the contours part of bones and air surrounding the tongue. The lower part of the GTV was correctly contoured whereas the upper part (at the cephalic limit) showed some errors.

Regarding ANON3027, similarly to patient ANON2963, the lower part of the GTV is very well contoured whereas closer to the skull, the contour stops abruptly. The radiation oncologist would have gone a bit higher.

Like patient ANON2839, the contour for ANON3331 is also too anterior. Like patients ANON3027 and ANON2963, it stops too early close to the skull. However, in the posterior side the anatomic barriers are well respected.

According to the physician, ANON4138 was a very hard case. In the predicted GTV, the

anatomical barriers are well respected but the image was of very bad quality. Actually, the physician himself said that he would not have been capable of delineating the tumor based on this MRI only.

One might note that for ANON2963 and ANON3331, the MR images were not T2 Fat-Sat but FLAIR. Despite this difference, the model showed robustness since the lower and posterior predicted GTV were close to the volume physician would have delineated.

To summarize, the inference on these 7 patients showed acceptable results. The lower part tumor was most of the time well delineated, even if sometimes the predicted GTV was too anterior, encompassing the tongue whereas it should not. The network showed weakness regarding the upper part of the tumor. In fact, while going closer to the skull, it tended to underestimate the GTV and to stop too soon. This could happen because the tumors invading higher part of the head and neck are not well represented in the database, thus nnU-Net did not learn much on this type of cases. The physician also highlighted the fact that, globally, the network did not confound the GTV-P and GTV-N, which is a good point.

5.2 nnX-Net results

Due to a lack of time and computer cluster availability, no results have been obtained regarding nnX-Net at the end of this thesis. Actually, the global structure of this new method is ready, since the network training works, but the outputs are empty. Only a few minor updates and adaptations of the code probably have yet to be done to get the first results.

5.3 Discussions

In the following subsections, the results presented above are analyzed. These interpretations should be considered carefully since the database is comprised of only 20 patients, which is not a great sample to run a statistical analysis.

5.3.1 nnU-Net 2D vs 3D

nnU-Net 2D performed worse than nnU-Net 3D regarding the Dice and Surface Dice. This is probably because of the loss of context in the craniocaudal direction as explained in subsection 4.1.1.2. Maybe this effect is even worse in HNC because this region of the body contains a lot of different anatomical structures that are stacked in a small region. Thus it can change rapidly while going up. The mean HD is better in nnU-Net 2D, but the median HD is better in nnU-Net 3D, so there is probably an outlier in the HD for nnU-Net 3D. Figure 5.10 compares the performance of the 2D and 3D models for the same slice of patient ANON590. For this patient, the Dice score was 0.56 for nnU-Net 2D and 0.85 for nnU-Net 3D. According to the Figure, nnU-Net 2D underestimates the GTV volume for this patient. Figure 5.1 shows that this underestimation could be a trend for the 2D model. Indeed, it is the case for 11 out of 20 patients. Figure 5.4 shows a similar trend for the 3D model, although the differences are less pronounced.

ANON590 - Ground-truth and predicted GTVs nnU-Net 2D and 3D



Figure 5.10: Comparison between the ground-truth and predicted contours by nnU-Net 2D and 3D for patient ANON590. The displayed slice corresponds to the center of mass of the ground-truth tumor.

5.3.2 nnU-Net 3D

Provided the size of the database, the results are quite good for nnU-Net 3D (mean Dice and Surface Dice of 0.61 and 0.56, medians of 0.65 and 0.56), excepted for the HD which is big (16.12 mm). The mean Dice is even close to the results obtained by Bielak et al. [51] (training on 33 patients) in which the Dice scored up to 0.65 (see Chapter 3).

Figures 5.8 and 5.9 show that for growing GTV volumes above 50 cc, the Dice, HD and Surface Dice start worsening. This might be explained by the fact that great GTV volumes are not very represented in the database: only 4 patients have GTV volumes bigger than 50 cc. As a consequence, the network might experience difficulties learning how to perform segmentation of these great volumes.

For very small GTVs (under 11 cc), nnU-Net 3D also produces bad results (3 Dice scores of 0.16, 0.31, and 0.4 for patients ANON3648, ANON2453 and ANON1132, respectively) despite the fact that these small GTVs are more represented. However, one might remind the fact that the DICE coefficient is not always appropriate to assess the model's performance on small GTVs (see section 3.1.1). Thus, to properly evaluate these patients, one should look at another metric. The resulting HD for these patients were respectively 22 mm, 16.55 mm, and 23.28 mm. Therefore, the mean HD being 16.12 mm, the contours for ANON2453 might not be so bad in the end.

Figure 5.8 also shows the interest of not only using the Dice to assess the segmentation accuracy. Several patients have acceptable or good Dice scores but far less good Surface Dice scores that should raise awareness on the segmentation quality. For example, the pair Dice-Surface Dice for ANON2865 is 0.71-0.51, and patients ANON866, ANON2971, ANON3620, and ANON4297 show drops in a range of 0.11-0.19 between their Dice and Surface Dice scores. When looking at the predicted contours for these patients, displayed in Figure 5.11 one might notice indeed that, globally, the predicted GTV volumes are close to the ground-truth GTVs, but that the network struggles with bumps and irregularities in the contours, explaining the poor Surface Dice results. This is just not the case for ANON2971, but this patient already had a bad Dice of 0.46.

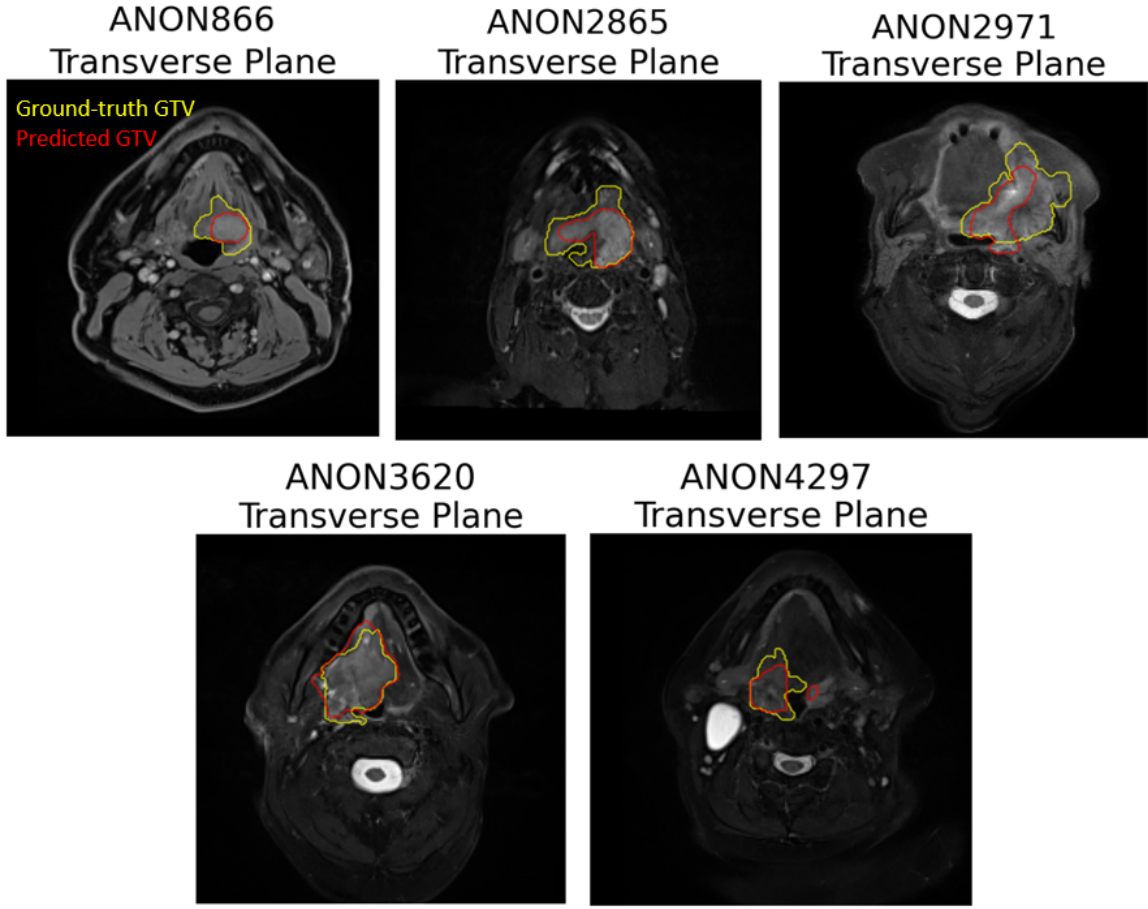


Figure 5.11: Patients showing a drop between the DICE and Surface DICE

It is interesting to observe that while the Dice can be much better than the Surface Dice, the opposite does not happen. Some patients (e.g., ANON590, ANON690) have indeed a greater Surface Dice than Dice, but the difference is never more than 0.07. This could be explained by the fact that according to its formula, a good Surface Dice implies contours that are close and of similar lengths, thus the volumes they delineate must be similar. On the contrary, a good DICE implies that the predicted volume intersects, for most of it with the ground-truth volume, but despite this volume intersection its contours might be very irregular, decreasing $|G \cap P|$ and increasing $|P|$, leading to a decrease of the Surface Dice.

In terms of clinical use, if the GTV is not too small, the Dice can definitely be used to get a first estimation of how well the model performed. If it is acceptable, computing the Surface Dice afterwards could give more information on how much the contours should be modified by a physician. For a small GTV, the Dice could be computed too, but should be completed with another metric such as the HD.

The 95% HD was computed and compared to HD in Table 5.3. The aim was to identify outliers that might bias how the distance between the ground-truth and predicted volumes is great. As a reminder, the HD is obtained by taking the maximum of the distances from the points of one set to the nearest point of the other set whereas 95% HD is computed by taking the maximum among the distances belonging to the 95% percentile. In the context of a semi-automated delineation method, in which the physician could for example adjust the contours

given by the network, this could be useful. In fact, if a patient has a big HD but that the 95% HD is significantly smaller, it might indicate that there is just an artefact in the delineation that could be easily corrected. This kind of big difference was not observed as shown in Table 5.3, since for 12 patients there was no difference between HD and HD 95%, and for the 8 remaining patients the difference stayed within a range of 0.01-0.45 mm.

Figure 5.4 shows that for 10 patients, the GTV volume is overestimated and for the 10 others it is underestimated. Since the difference in the underestimation is bigger than the overestimation, the 3D network seems to underestimate the GTV volumes.

Regarding the 3D model with the CT scans as second channel, less good results were obtained (mean Dice of 0.56 over 3 validation patients versus 0.7 for the model fed with MR images only). It was expected since the images were not registered.

As far as the 3D nnU-Net with more data augmentation is concerned, a poorer performance was observed too (mean Dice of 0.68 over 3 validation patients instead of 0.7 with the original data augmentation parameters). It can be explained by the fact that nnU-Net performs a data augmentation operation on an image with a certain probability and replace the image with its transformed version in the dataset if the transformation happens. Since the probabilities of transformation occurrences were doubled, the network probably suffered from a lack of original images.

Finally, it is worth highlighting the fact that, except for ANON3648, nnU-Net 3D did not confound the GTV-T and GTV-N (when there was one).

5.4 Limitations in the approach

The methods used in this work suffer from limitations related to different aspects of the experiments. There are indeed limitations and uncertainties associated with Deep Learning techniques, they are developed in the first following subsection. Limitations are also induced by the data, more precisely by their quality and quantity. These aspects are explained in subsection 5.4.2.

5.4.1 CNN limitations

There are drawbacks inherent to the use of Convolutional Neural Networks. Their training takes much time and requires much data and powerful computational resources. A major flaw of nnU-Net in particular is that it does not include a measure of the uncertainty with the network predictions. Different techniques have been proposed to quantify the uncertainty [59], among which the Monte Carlo dropout [60]. In this method, as represented in Figure 5.12, neurons of the network are randomly "switched off" at both training and testing time, meaning that their outputs are set to 0. Multiple predictions are generated with different combinations of "on" and "off" neurons. A prediction distribution is obtained and can be analyzed to evaluate how much is the network sensitive to modification in its architecture and thus how sure of its prediction the network is. More precisely, after several passes in the network with dropout activated in inference time, the mean of all predicted values is used as the final prediction, and the standard deviation is used as a surrogate of the uncertainty.

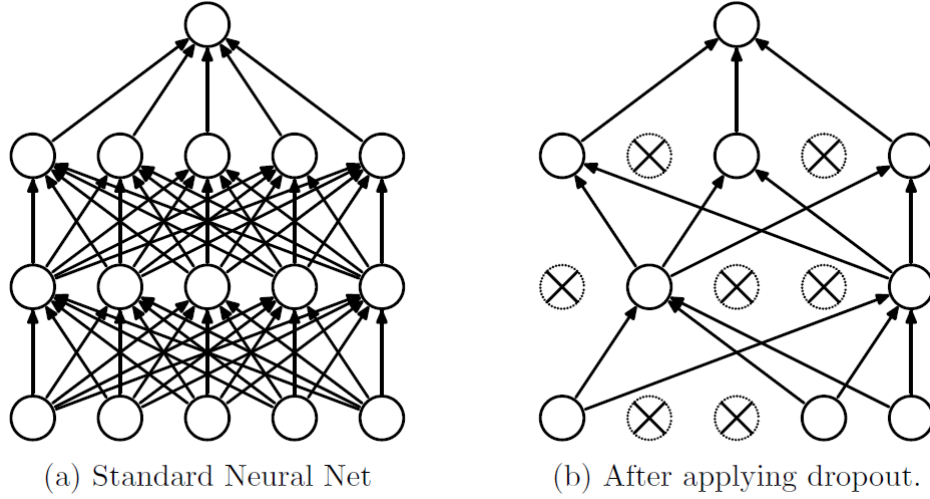


Figure 5.12: Illustration of the Monte Carlo Dropout method to evaluate the uncertainty linked to a network

5.4.2 Data quantity and user-dependant variability

Since the NN learns how to perform a task on a dataset, the latter has a major impact on the network performance. In February 2021, Barragán-Montero et al. [61] published a paper in which they assess the impact of data quality and quantity on the results produced by DL algorithms within the framework of dose prediction in esophageal cancer. They conducted 2 experiments, the first one aiming to quantify the effects of user-dependent variability and the second one the effects of the database size. The Mean Absolute Error (MAE) was used to quantify the models performances. MAE is defined, between a ground-truth mask X and a predicted mask Y containing n elements, as

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} .$$

The first experiment showed that the MAE was smaller for an homogenized database (patients were re-delineated and the treatment planning was recreated with a class-solution protocol), compared to the original heterogeneous database. In addition, the boxplots of the doses in different volumes of interest were narrower for the homogenized database, which is the sign of a more robust model. In Experiment 2, networks were trained with increasing databases sizes. This increase slightly improved the performance of the models trained with the homogenized database: the global MAE decreased of 0.13 Gy between the network trained with 26 patients and the one trained with 56 patients. However, for the models trained with the original database, the error increased with the augmentation of the database size (global MAE increase of 0.2 Gy). The authors concluded that a small database can lead to good performance and highlighted the importance of data homogeneity, prior to quantity.

In this work, the data can be considered as homogeneous, since the same radiation oncologist performed the delineation on all the MR images. However, this is also a limitation of the method, since this physician could have made mistakes. There can be what is called intra-observer variability in the dataset, i.e., differences in a repeated task performed by the same observer.

Regarding the data quantity in this thesis, the database size is definitely small. Increasing it

in order to get a broader representation of the GTV volumes and anatomical locations (in this case especially tumors localized in the upper skull) could definitely help the network learn on more different cases and thus improve the results. Active learning could also be integrated to the method to improve it. It is a type of ML in which the algorithm chooses by itself what kind of data should be labelled and input in the training set in order to enhance the performance. See [62] for more details.

One might also highlight the fact that the MR images quality in the database were not always good. In fact, about 25 % of them were blurry. This is inherent to MRI, since the acquisition time is long, it can be difficult for the patient not to move. In addition, these MR images are diagnostic images, so the purpose is to detect a disease and not especially obtaining a very high quality image. Segmentation studies performed on CT scans do not encounter this problem since the CT scans that are used are planning CT scans which are designed to be images with high resolution and quality. Some Unsupervised Learning Methods denoising MR images exist and could be used [63], as well as adding extra images as supplemental information, especially the CT scan. An analysis to quantify the blurriness of the image and put it in perspective with the performance could also be conducted. This could be done using Fast Fourier Transform for instance.

5.5 Perspectives

5.5.1 Extra input information

A first idea to improve the results would be to introduce the PET scan in the method. This imaging modality brings metabolic information that could complete the anatomical information offered by MRI and help the network localizing the tumor. As developed in Chapter 3, some papers [50] showed that including PET scans improved the segmentation results. PET scans could thus be added as an input of nnX-Net, along with the MR images for instance. In addition to the PET scan, information given by the endoscopy should be included, as well as knowledge about the patient’s general health. For instance, if a patient has HPV, it is more likely that the cancer will have spread in lymph nodes, and thus there will be a GTV-N.

5.5.2 nnX-Net

First of all, the remaining adaptations should be carried out so that the model is completely ready to produce results. As previously mentioned, adding the PET scan as an input to nnX-Net could lead to better results. The method could also be improved by reintroducing multithreading to enhance the computational efficiency.

5.5.3 Other DL techniques

5.5.3.1 Attention Networks

For patient ANON3648, as displayed in Figure 5.5 the network included the GTV-N in the predicted GTV, which is a mistake. A solution to avoid this confusion would be to use Attention techniques. In ML, Attention is a method in which more awareness is raised on some parts of the inputs than on other parts [64]. Transformers, which are described in the following section, are built upon self-attention mechanisms.

5.5.3.2 Transformers

A segmentation technique that was not used on HNC yet but that is worth mentioning because of the growing interest in the field is the use of transformers. These models, originally used in Natural Language Processing where they achieved state-of-the-art performances, have gained attention in computer vision tasks.

In October 2021, Hatamizadeh et al. [65] proposed a combination of the U-Net architecture and the transformers. They identified that a weakness of Deep CNN is the fact that their capability to learn long-range spatial dependencies is limited by the locality of their receptive fields. In order to solve this problem, they replaced the encoder part of the U-Net with a transformer, which is capable of effectively capturing multi-scale information. They kept the decoder part as in U-Net, and the skip connections between the latter and the encoder. This architecture is depicted in Figure 5.13.

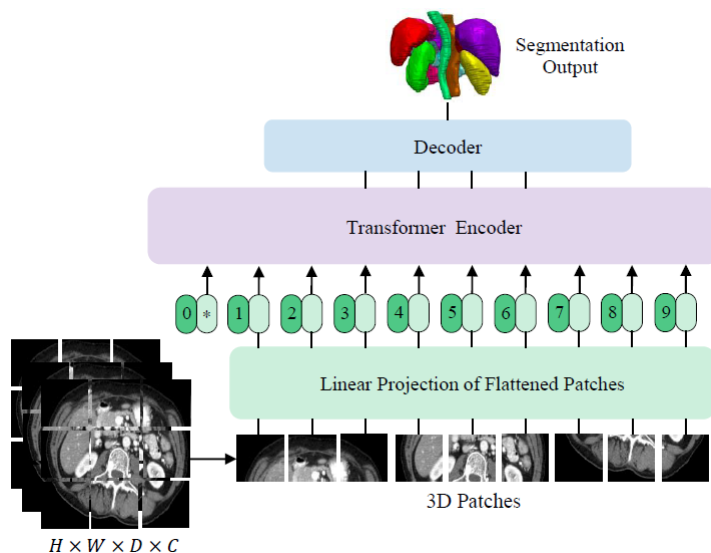


Figure 5.13: UNETR architecture (from [65])

The transformers are better at capturing long-range dependencies because of their self-attention property. This mechanism allows them to identify the important features of an input on a long-range and assess bigger weights to the input parts that they consider to be more important in the task context.

UNETR achieved a global Dice score of 0.899 at the BTCV Competition (multi-organ segmentation in the abdomen) and even set new state-of-the-art in some tasks. This transformer-based model could thus be one of the first in a new class of segmentation methods.

Chapter 6

Conclusion

The aim of this thesis was to assess the performance of nnU-Net on the MRIs of 20 patients suffering from HNC. nnU-Net was chosen because it proposes an innovative method of setting up the design choices and determining the network hyperparameters. The 2D version of nnU-Net gave less good results than the 3D network, probably due to the lack of context in the craniocaudal direction. The main observations regarding the 3D nnU-Net can be summarized as follows:

- Overall, Dice and Surface Dice scores were good provided the database size (mean of 0.61 ± 0.21 and 0.56 ± 0.23 and median of 0.65 and 0.56, respectively)
- Mean HD was big compared to other studies conducted on the same type of task.
- In inference patients, the results were clinically acceptable on most of them. The network showed more difficulties with the anterior parts of the tumors and with GTVs located in the upper skull.

Therefore, despite the small size of the database, nnU-Net showed good results. This method, which had already shown good results on international segmentation contests databases, proved that it is also quite robust for small ones.

Moreover, a new version of nnU-Net, called nnX-Net was proposed. This network allows inputting CT scans and MRIs in the network in order to get the best of the two imaging modalities even if they are not registered. Trying the combination MRI/PET scan would be interesting too. This network could not be completely finished due to a lack of time and computing resources, but it could definitely be further improved by allowing the input of more imaging modalities or reintroducing multithreading to make it faster.

Better results could be obtained by increasing the size of the database and, more importantly, making sure that the more GTV volumes, localization and shapes as possible are represented. The networks could also be improved. Indeed, techniques to assess the uncertainty in nnU-Net and nnX-Net should be implemented. Attention mechanisms could also be added to make sure that the network is focused on the GTV.

Future research conducted on the topic will probably involve state-of-the-art methods such as nnU-Net and promising networks like transformers. AutoML techniques and Active Learning also show great potential in enhancing these models efficiency. Hopefully, this research will lead to applications that can be used in clinical routine to automate the cumbersome task of segmentation. This would ensure good patient care and safety and save a lot of physician's time, that they will be able to dedicate to other tasks.

Appendix A

nnU-Net 3D results

ANON242 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.51

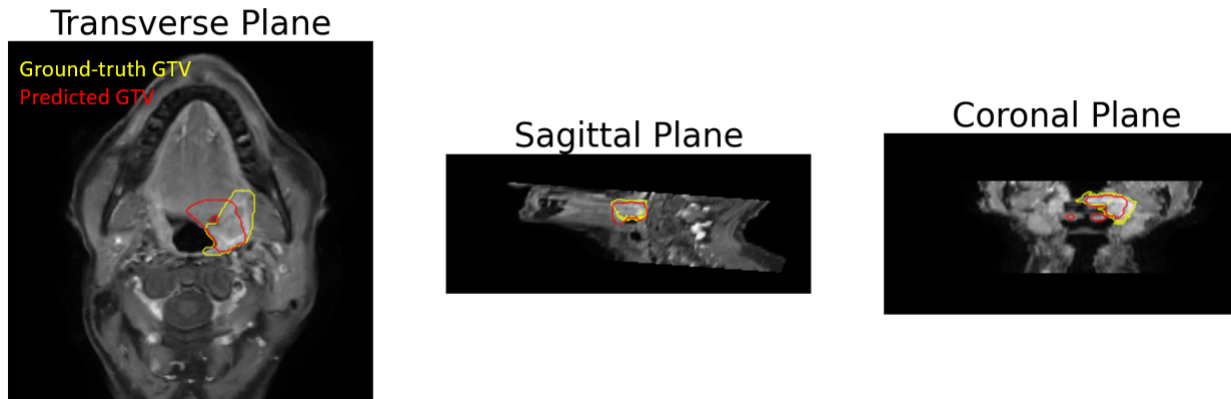


Figure A.1: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON242

ANON590 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.85



Figure A.2: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON590

ANON690 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.83

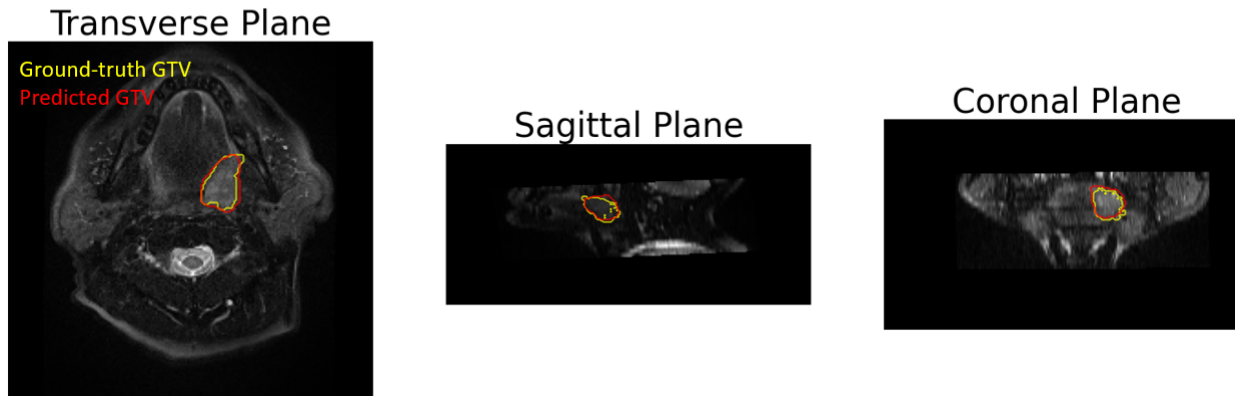


Figure A.3: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON690

ANON866 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.53

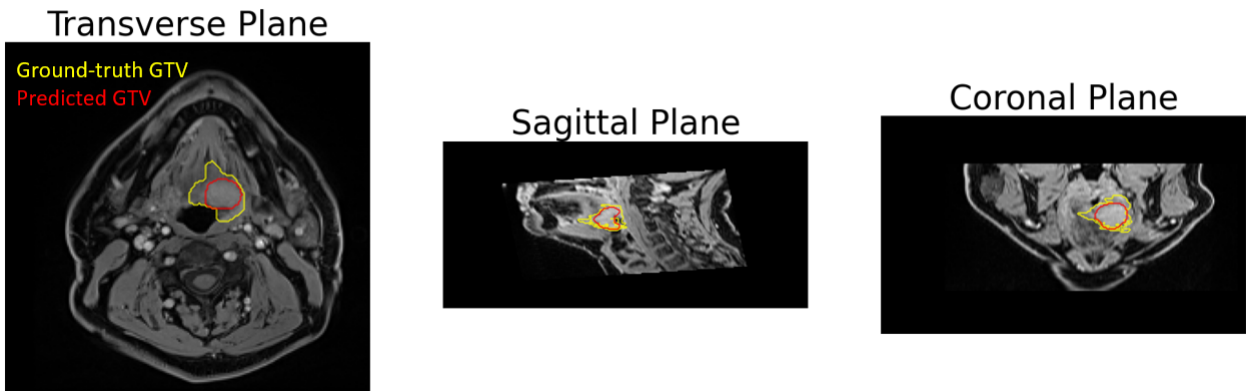


Figure A.4: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON866

ANON1132 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.4

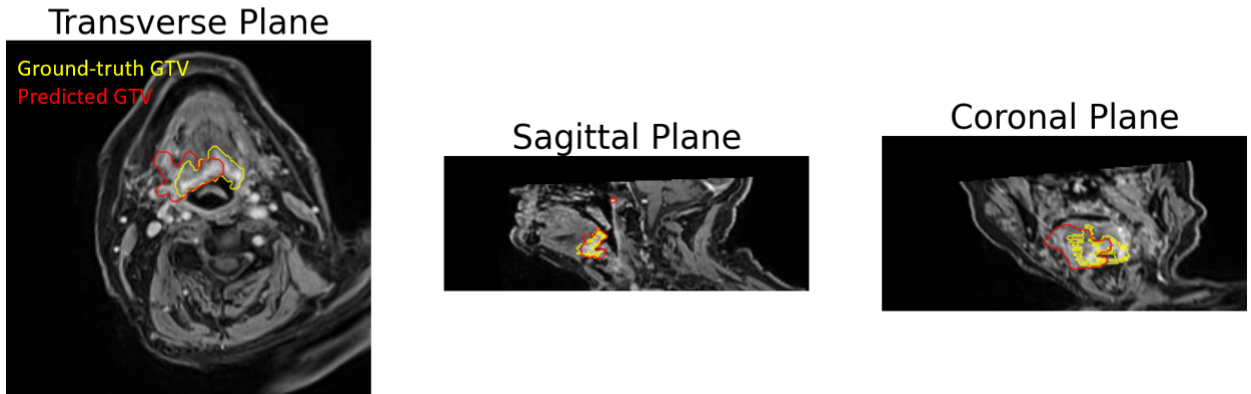


Figure A.5: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON1132

ANON1677 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.78

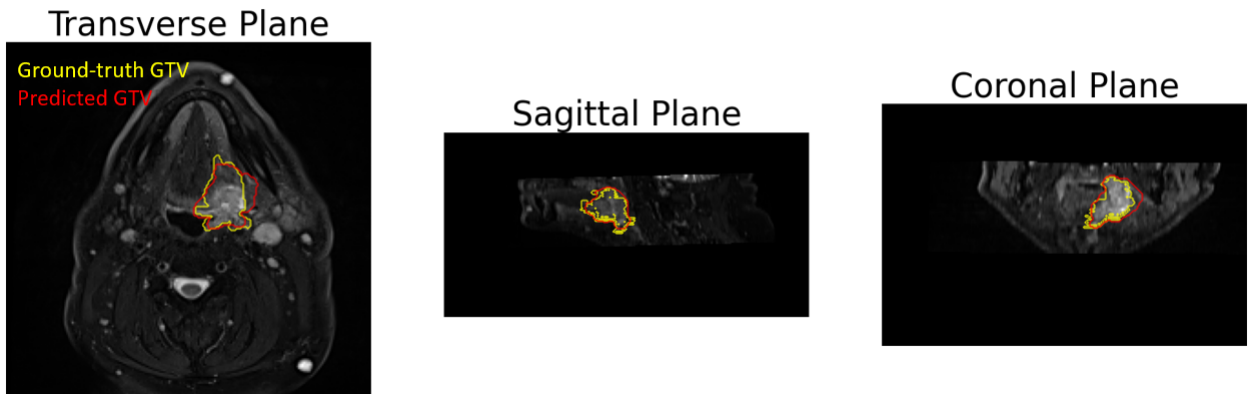


Figure A.6: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON1677

ANON1783 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.66

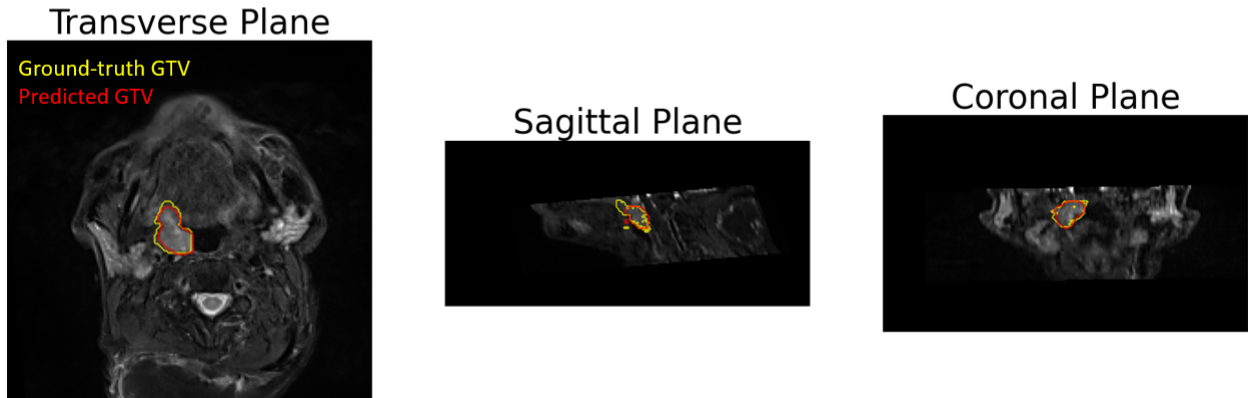


Figure A.7: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON1783

ANON2453 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.31

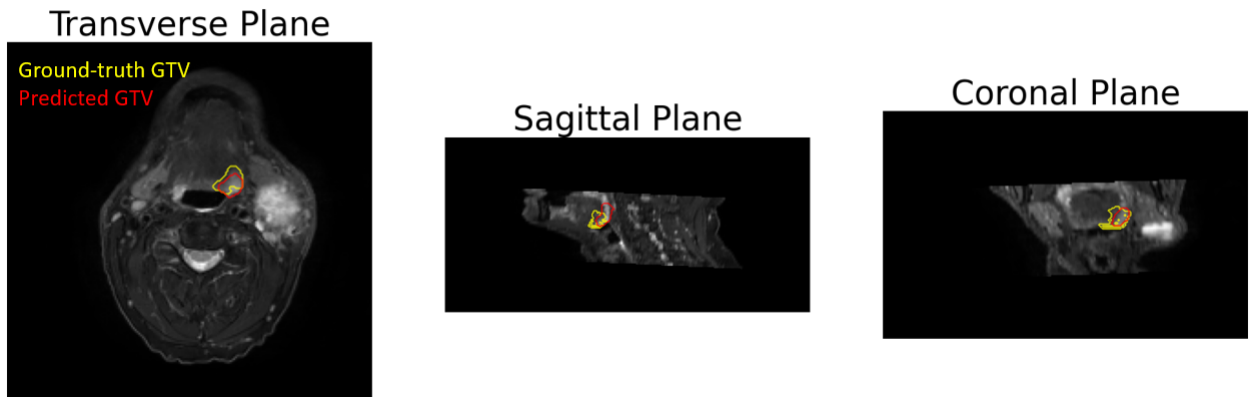


Figure A.8: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON2453

ANON2834 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.69



Figure A.9: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON2834

ANON2865 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.71



Figure A.10: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON2865

ANON2971 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.46

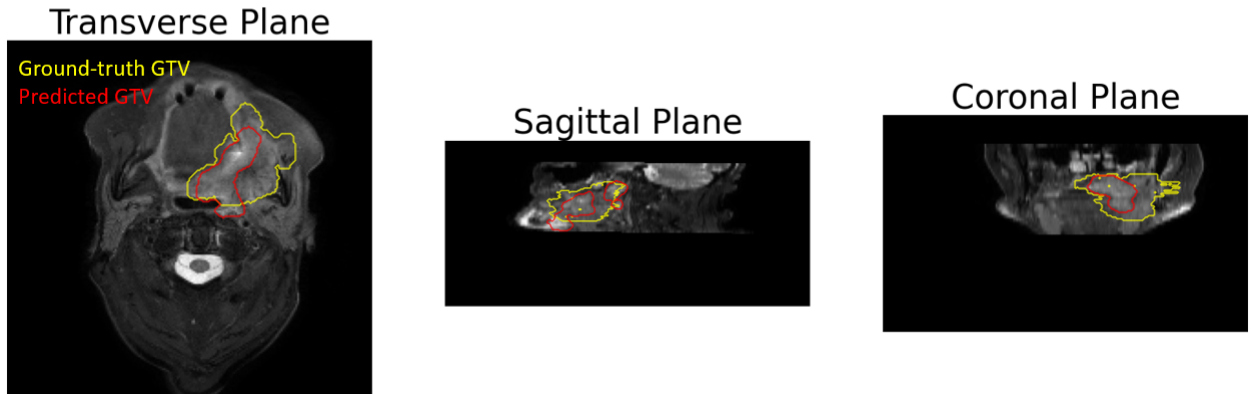


Figure A.11: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON2971

ANON3146 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.21

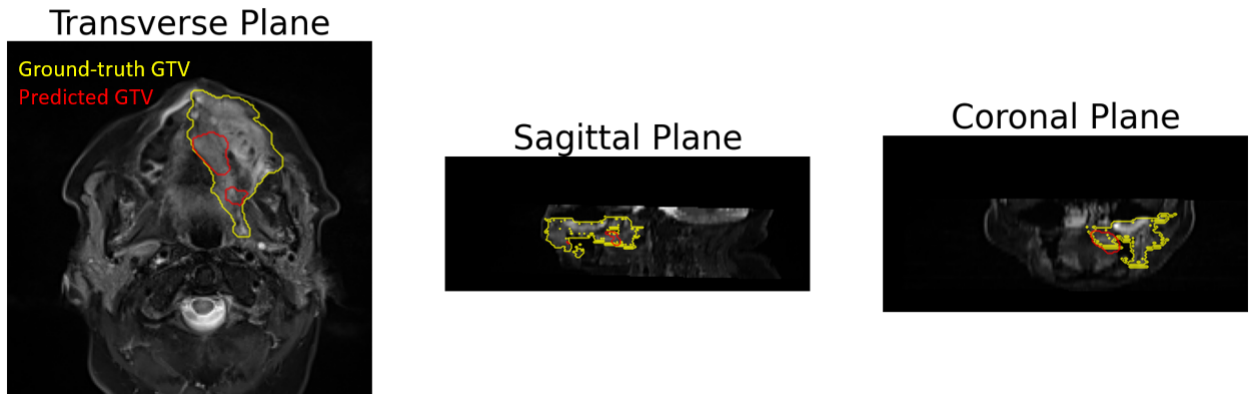


Figure A.12: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3146

ANON3207 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.85

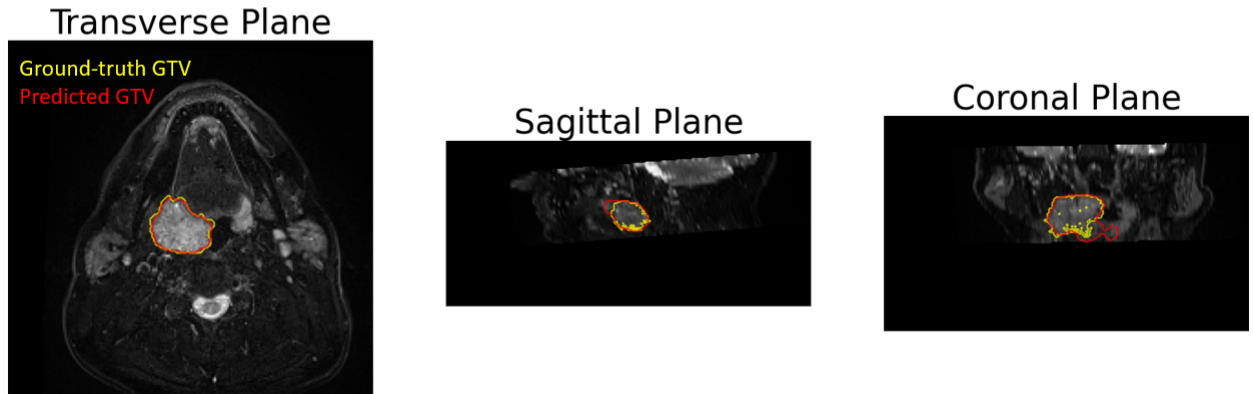


Figure A.13: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3207

ANON3435 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.81



Figure A.14: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3435

ANON3620 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.77

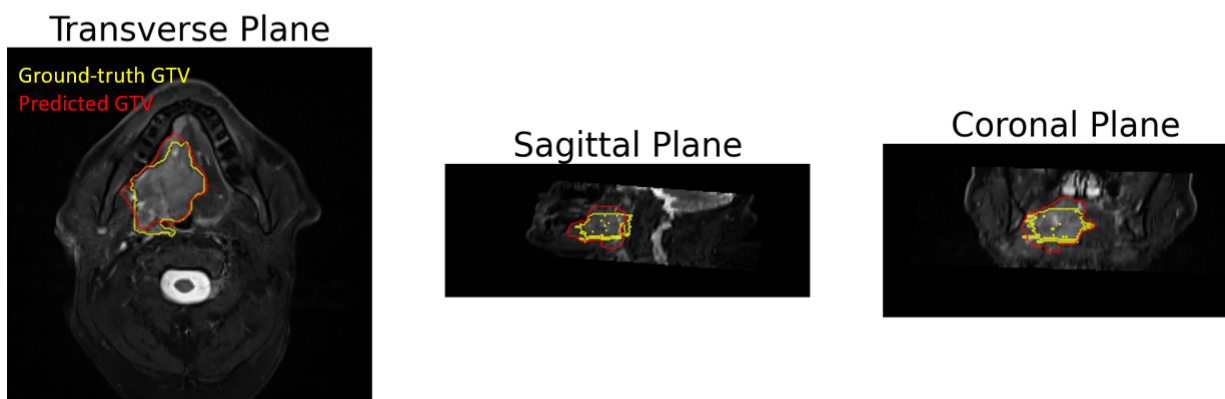


Figure A.15: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3620

ANON3647 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.56

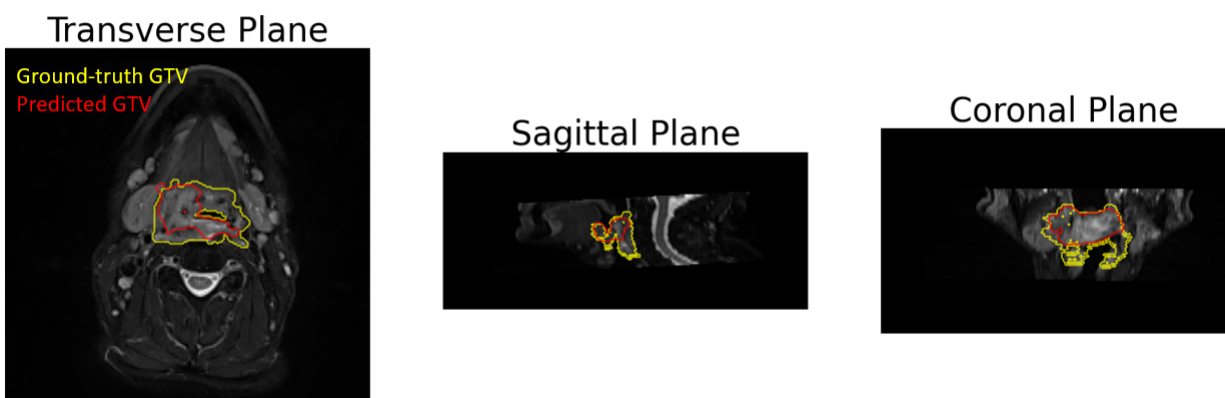


Figure A.16: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3647

ANON3648 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.16



Figure A.17: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON3648

ANON4232 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.84

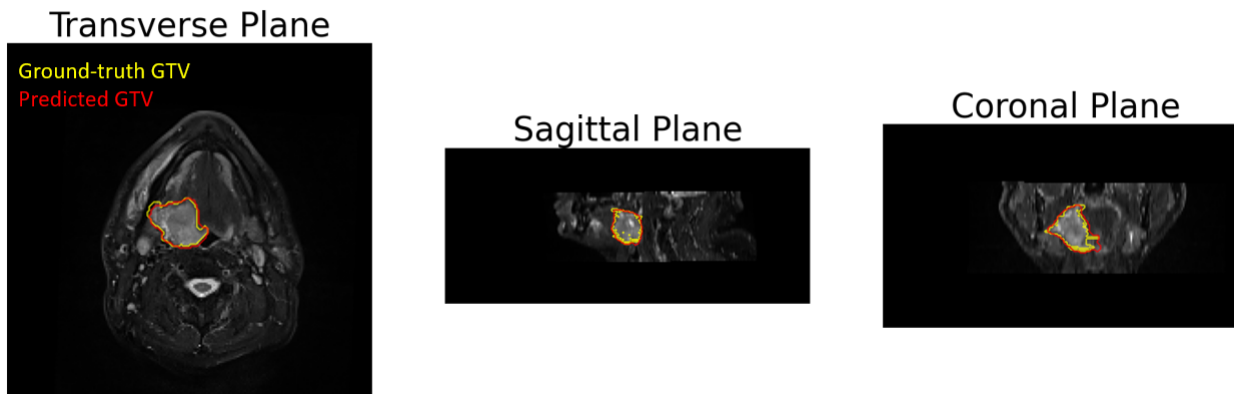


Figure A.18: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON4232

ANON4297 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.64

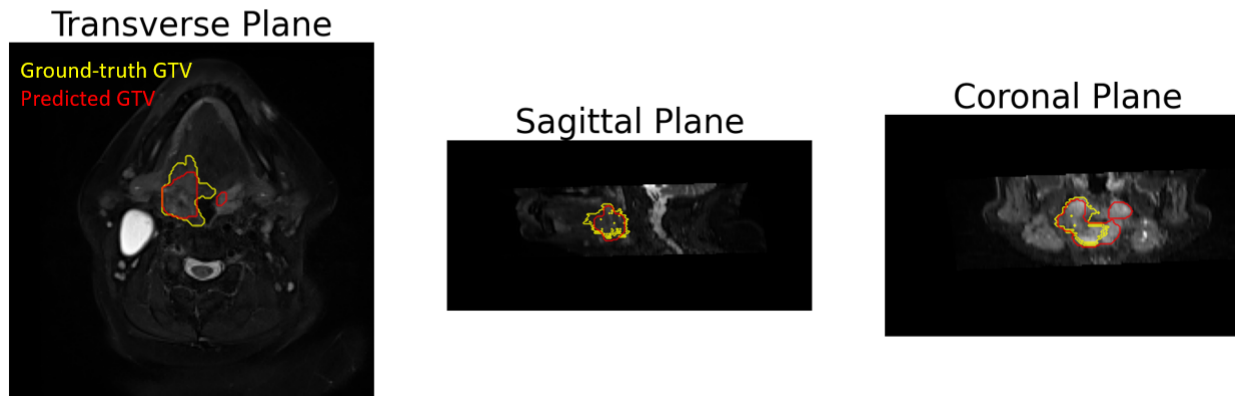


Figure A.19: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON4297

ANON4641 - Ground-truth and predicted GTVs
nnU-Net 3D - DICE = 0.58



Figure A.20: Ground-truth (yellow) and predicted (red) GTV contours at the center of mass of the ground-truth GTV for patient ANON4641

Appendix B

nnU-Net 3D - Inference results

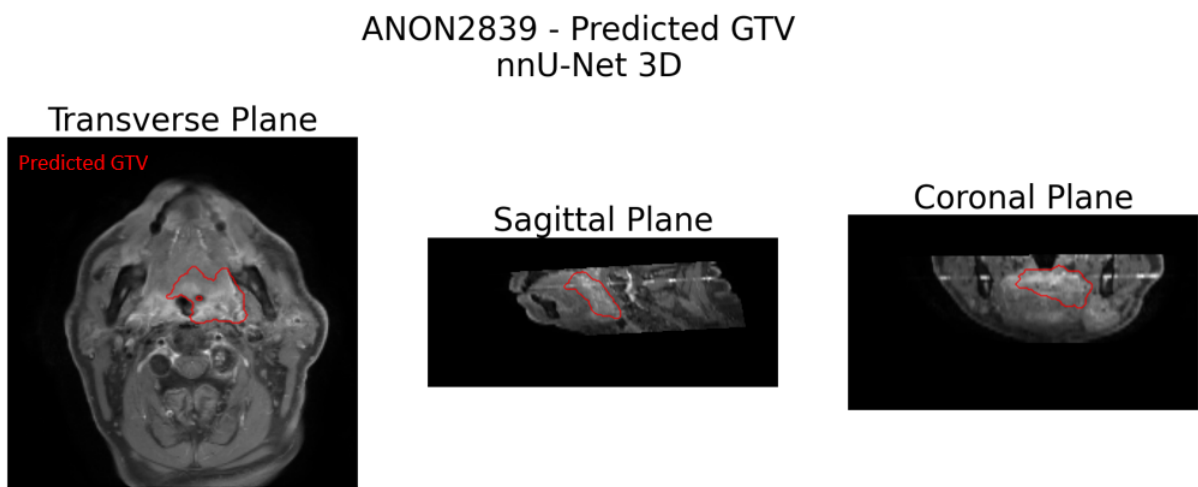


Figure B.1: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON2839

ANON2963 - Predicted GTV
nnU-Net 3D



Figure B.2: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON2963

ANON3027 - Predicted GTV
nnU-Net 3D



Figure B.3: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON3027

ANON3331 - Predicted GTV
nnU-Net 3D



Figure B.4: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON3331

ANON3454 - Predicted GTV
nnU-Net 3D



Figure B.5: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON3454

ANON3820 - Predicted GTV
nnU-Net 3D



Figure B.6: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON3820

ANON4138 - Predicted GTV
nnU-Net 3D



Figure B.7: Predicted GTV contours at the center of mass of the ground-truth GTV for patient ANON4138

Bibliography

- [1] Freddie Bray et al. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians* 68.6 (Sept. 2018), pp. 394–424. DOI: 10.3322/caac.21492. URL: <https://doi.org/10.3322/caac.21492>.
- [2] Pim Moeskops et al. “Automatic Segmentation of MR Brain Images With a Convolutional Neural Network”. In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1252–1261. DOI: 10.1109/TMI.2016.2548501.
- [3] Tomohiro Kajikawa et al. “A convolutional neural network approach for IMRT dose distribution prediction in prostate cancer patients”. In: *Journal of Radiation Research* 60.5 (July 2019), pp. 685–693. DOI: 10.1093/jrr/rrz051. URL: <https://doi.org/10.1093/jrr/rrz051>.
- [4] K.C. Santosh et al. “Deep Neural Network for Foreign Object Detection in Chest X-Rays”. In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 2020, pp. 538–541. DOI: 10.1109/CBMS49503.2020.00107.
- [5] C Kontaxis et al. “Towards adaptive IMRT sequencing for the MR-linac”. In: *Physics in Medicine and Biology* 60.6 (Mar. 2015), pp. 2493–2509. DOI: 10.1088/0031-9155/60/6/2493. URL: <https://doi.org/10.1088/0031-9155/60/6/2493>.
- [6] Belgian Cancer Registry. *Cancer burden in Belgium 2004-2017*. Brussels, 2020.
- [7] J Betts et al. *Anatomy and Physiology*. Apr. 2013.
- [8] Bhawna Gupta, Newell W. Johnson, and Narinder Kumar. “Global Epidemiology of Head and Neck Cancers: A Continuing Challenge”. In: *Oncology* 91.1 (2016), pp. 13–23. DOI: 10.1159/000446117. URL: <https://doi.org/10.1159/000446117>.
- [9] Eléonore Longton. “Target Volumes and Introduction to Head and Neck Radiotherapy”. In: *WRDTH3160 - Computed dosimetry* (UCLouvain, Academic year 2020-2021).
- [10] Rajamanickam Baskar et al. “Cancer and Radiation Therapy: Current Advances and Future Directions”. In: *Int J Med Sci* 9 (2012), pp. 193–199. DOI: 10.7150/ijms.3635.
- [11] M. Salam (ed.) *The Principles and Practice of Urology. Volumes 1 and 2*. 2nd. Jaypee Brothers, 2013.
- [12] King-Hay Yang. *Basic Finite Element Method as Applied to Injury Biomechanics*. 2018. ISBN: 978-0-12-809831-8.
- [13] T. Landberg et al. “2. Volumes”. In: *Reports of the International Commission on Radiation Units and Measurements* os-32.1 (1999), pp. 3–20. DOI: 10.1093/jicru/os32.1.3.
- [14] Beth Beadle and Carryn Anderson. “CTV Guidance for Head and Neck Cancers”. In: *International Journal of Radiation Oncology* (2020). DOI: 10.1016/j.ijrobp.2017.12.264.

- [15] V. Grégoire et al. “Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines”. In: *Radiotherapy and Oncology* 126.1 (Jan. 2018), pp. 3–24. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2017.10.016. URL: <https://doi.org/10.1016/j.radonc.2017.10.016>.
- [16] C. L. Brouwer et al. “CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines”. In: *Radiother Oncol* 117.1 (Oct. 2015), pp. 83–90.
- [17] Philip Mayles, Alan Nahum, and J. Rosenwaeld. *Handbook of Radiotherapy Physics : Theory and Practice*. CRC Press, 2007. ISBN: 97807503086016.
- [18] Jean-Philippe Dillenseger, Elisabeth Moerschel, and Claudine Zorn. *Guide des technologies de l’imagerie médicale et de la radiothérapie*. 2016. ISBN: 978-2-294-75201-8.
- [19] Nicolas Michoux. In: *WRDGN3120 - Principes, techniques et contrôles de qualité en imagerie médicale* (Academic year 2020-2021).
- [20] John P Ridgway. “Cardiovascular magnetic resonance physics for clinicians: part I”. In: *Journal of Cardiovascular Magnetic Resonance* 12.1 (Nov. 2010). DOI: 10.1186/1532-429x-12-71. URL: <https://doi.org/10.1186/1532-429x-12-71>.
- [21] In: *International Journal of Stroke* 14.3 (July 2018), pp. 270–281. DOI: 10.1177/1747493018784478. URL: <https://doi.org/10.1177/1747493018784478>.
- [22] Girish Katti, Syeda Ara, and Dr Shireen. “Magnetic resonance imaging (MRI) - A review”. In: *Intl J Dental Clin* 3 (Mar. 2011).
- [23] David Poole, Alan Mackworth, and Randy Goebel. *Computational Intelligence: A Logical Approach*. Jan. 1998. ISBN: 978-0-19-510270-3.
- [24] Tom M Michell. *Machine Learning*. 1997. ISBN: 0-07-042807-7.
- [25] Mohammad Hemmat Esfe et al. “A well-trained artificial neural network for predicting the rheological behavior of MWCNT–Al₂O₃ (30–70%)/oil SAE40 hybrid nanofluid”. In: *Scientific Reports* 11.1 (Aug. 2021). DOI: 10.1038/s41598-021-96808-4. URL: <https://doi.org/10.1038/s41598-021-96808-4>.
- [26] Patrick Amar, Francois Kepes, and Vic Norris. *Proceedings of The 2012 Evry Spring School on Modelling Complex Biological Systems in the Context of Genomics*. May 2012.
- [27] Junxi Feng et al. “Reconstruction of porous media from extremely limited information using conditional generative adversarial networks”. In: *Physical Review E* 100.3 (Sept. 2019). DOI: 10.1103/physreve.100.033308. URL: <https://doi.org/10.1103/physreve.100.033308>.
- [28] Xin He, Kaiyong Zhao, and Xiaowen Chu. “AutoML: A survey of the state-of-the-art”. In: *Knowledge-Based Systems* 212 (Jan. 2021). DOI: 10.1016/j.knosys.2020.106622. URL: <https://doi.org/10.1016/j.knosys.2020.106622>.
- [29] Hooman H. Rashidi et al. “Machine learning in health care and laboratory medicine: General overview of supervised learning and Auto-ML”. In: *International Journal of Laboratory Hematology* 43.S1 (July 2021), pp. 15–22. DOI: 10.1111/ijlh.13537. URL: <https://doi.org/10.1111/ijlh.13537>.

- [30] Eliza Rebeca Tomodan and Catalin Daniel Căleanu. “Bag of Features vs Deep Neural Networks for Face Recognition”. In: *2018 International Symposium on Electronics and Telecommunications (ISETC)*. IEEE, Nov. 2018. DOI: 10.1109/isetc.2018.8583846. URL: <https://doi.org/10.1109/isetc.2018.8583846>.
- [31] Ana M. Barragán-Montero et al. “Artificial intelligence and machine learning for medical imaging: A technology review”. In: *Physica Medica* 83 (Mar. 2021), pp. 242–256. DOI: 10.1016/j.ejmp.2021.04.016. URL: <https://doi.org/10.1016/j.ejmp.2021.04.016>.
- [32] Abdel Aziz Taha and Allan Hanbury. “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool”. In: *BMC Medical Imaging* 15.1 (Aug. 2015). DOI: 10.1186/s12880-015-0068-x. URL: <https://doi.org/10.1186/s12880-015-0068-x>.
- [33] Varduhi Yeghiazaryan and Irina Voiculescu. “Family of boundary overlap metrics for the evaluation of medical image segmentation”. In: *Journal of Medical Imaging* 5.01 (Feb. 2018), p. 1. DOI: 10.1117/1.jmi.5.1.015006. URL: <https://doi.org/10.1117/1.jmi.5.1.015006>.
- [34] Tomaž Vrtovec et al. “Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods”. In: *Medical Physics* 47.9 (), e929–e950. DOI: <https://doi.org/10.1002/mp.14320>. eprint: <https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14320>. URL: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14320>.
- [35] Jordan Wong et al. “Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning”. In: *Radiotherapy and Oncology* 144 (Mar. 2020), pp. 152–158. DOI: 10.1016/j.radonc.2019.10.019. URL: <https://doi.org/10.1016/j.radonc.2019.10.019>.
- [36] James C. Korte et al. “Cascaded deep learning-based auto-segmentation for head and neck cancer patients: Organs at risk on T2-weighted magnetic resonance imaging”. In: *Medical Physics* (Nov. 2021). DOI: 10.1002/mp.15290. URL: <https://doi.org/10.1002/mp.15290>.
- [37] M. Kosmin et al. “Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer”. In: *Radiotherapy and Oncology* 135 (2019), pp. 130–140. ISSN: 0167-8140. DOI: <https://doi.org/10.1016/j.radonc.2019.03.004>. URL: <https://www.sciencedirect.com/science/article/pii/S0167814019301082>.
- [38] Jasjit Suri, David Wilson, and Swamy Laxminarayan. *Handbook of Biomedical Image Analysis. Volume III: Segmentation and Registration Models*. 2005, pp. 435–487. ISBN: 0-306-48607-5.
- [39] Tiezhi Zhang et al. “Automatic Delineation of On-Line Head-And-Neck Computed Tomography Images: Toward On-Line Adaptive Radiotherapy”. In: *International Journal of Radiation Oncology*Biophysics* 68.2 (2007), pp. 522–530. ISSN: 0360-3016. DOI: <https://doi.org/10.1016/j.ijrobp.2007.01.038>. URL: <https://www.sciencedirect.com/science/article/pii/S036030160700199X>.
- [40] Jinzhong Yang et al. “Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy”. In: *Practical Radiation Oncology* 4.1 (Jan. 2014), e31–e37. DOI: 10.1016/j.prro.2013.03.003. URL: <https://doi.org/10.1016/j.prro.2013.03.003>.
- [41] Szabolcs Urban and Attila Tanács. “Atlas-based global and local RF segmentation of head and neck organs on multimodal MRI images”. In: *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*. IEEE, Sept. 2017. DOI: 10.1109/ispa.2017.8073577. URL: <https://doi.org/10.1109/ispa.2017.8073577>.

- [42] Guanghui Cheng et al. “Multi-atlas-based segmentation of the parotid glands of MR images in patients following head-and-neck cancer radiotherapy”. In: *Medical Imaging 2013: Computer-Aided Diagnosis*. Ed. by Carol L. Novak and Stephen Aylward. SPIE, Feb. 2013. DOI: 10.1117/12.2007783. URL: <https://doi.org/10.1117/12.2007783>.
- [43] Kieran Wardman et al. “The feasibility of atlas-based automatic segmentation of MRI for H&N radiotherapy planning”. In: *Journal of Applied Clinical Medical Physics* 17.4 (July 2016), pp. 146–154. DOI: 10.1120/jacmp.v17i4.6051. URL: <https://doi.org/10.1120/jacmp.v17i4.6051>.
- [44] Tomaž Vrtovec et al. “Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods”. In: *Medical Physics* 47.9 (July 2020). DOI: 10.1002/mp.14320. URL: <https://doi.org/10.1002/mp.14320>.
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: <http://arxiv.org/abs/1505.04597>.
- [46] Carlos E Cardenas et al. “Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks”. In: *Physics in Medicine & Biology* 63.21 (Nov. 2018). DOI: 10.1088/1361-6560/aae8a9. URL: <https://doi.org/10.1088/1361-6560/aae8a9>.
- [47] Özgün Çiçek et al. “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by Sebastien Ourselin et al. Cham: Springer International Publishing, 2016, pp. 424–432.
- [48] Carlos E. Cardenas et al. “Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach”. In: *International Journal of Radiation Oncology*Biophysics* 109.3 (Mar. 2021), pp. 801–812. DOI: 10.1016/j.ijrobp.2020.10.005. URL: <https://doi.org/10.1016/j.ijrobp.2020.10.005>.
- [49] Jordan Wong et al. “Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning”. In: *Radiotherapy and Oncology* 144 (Mar. 2020), pp. 152–158. DOI: 10.1016/j.radonc.2019.10.019. URL: <https://doi.org/10.1016/j.radonc.2019.10.019>.
- [50] Jintao Ren et al. “Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation”. In: *Acta Oncologica* 60.11 (July 2021), pp. 1399–1406. DOI: 10.1080/0284186x.2021.1949034. URL: <https://doi.org/10.1080/0284186x.2021.1949034>.
- [51] Lars Biela et al. “Convolutional neural networks for head and neck tumor segmentation on 7-channel multiparametric MRI: a leave-one-out analysis”. In: *Radiation Oncology* 15.1 (July 2020). DOI: 10.1186/s13014-020-01618-z. URL: <https://doi.org/10.1186/s13014-020-01618-z>.
- [52] Konstantinos Kamnitsas et al. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation”. In: *Medical Image Analysis* 36 (Feb. 2017), pp. 61–78. DOI: 10.1016/j.media.2016.10.004. URL: <https://doi.org/10.1016/j.media.2016.10.004>.

- [53] Zhe Guo et al. “Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network”. In: *Physics in Medicine & Biology* 64.20 (Oct. 2019). DOI: 10.1088/1361-6560/ab440d. URL: <https://doi.org/10.1088/1361-6560/ab440d>.
- [54] Michael Baumgartner et al. “nnDetection: A Self-configuring Method for Medical Object Detection”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer International Publishing, 2021, pp. 530–539. DOI: 10.1007/978-3-030-87240-3_51. URL: https://doi.org/10.1007/978-3-030-87240-3_51.
- [55] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18.2 (Feb. 2021), pp. 203–211. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z.
- [56] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: 42 (Dec. 2017), pp. 60–88. DOI: 10.1016/j.media.2017.07.005. URL: <https://doi.org/10.1016/j.media.2017.07.005>.
- [57] Paul Jäger. “nnU-Net: Automated Design of Deep Learning Methods for Biomedical Image Segmentation”. In: Data Science Seminar (DKFZ, Germany, Nov. 4, 2020). 2020.
- [58] Michal Drozdal et al. *The Importance of Skip Connections in Biomedical Image Segmentation*. 2016. arXiv: 1608.04117 [cs.CV].
- [59] Jakob Gawlikowski et al. “A Survey of Uncertainty in Deep Neural Networks”. In: *CoRR* abs/2107.03342 (2021). arXiv: 2107.03342. URL: <https://arxiv.org/abs/2107.03342>.
- [60] Yarín Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML].
- [61] Ana M. Barragán-Montero et al. “Deep learning dose prediction for IMRT of esophageal cancer: The effect of data quality and quantity on model performance”. In: *Physica Medica* 83 (Mar. 2021), pp. 52–63. DOI: 10.1016/j.ejmp.2021.02.026. URL: <https://doi.org/10.1016/j.ejmp.2021.02.026>.
- [62] Hamed H. Aghdam et al. “Active Learning for Deep Detection Neural Networks”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.
- [63] Marc Moreno López, Joshua M. Frederick, and Jonathan Ventura. “Evaluation of MRI Denoising Methods Using Unsupervised Learning”. In: *Frontiers in Artificial Intelligence* 4 (June 2021). DOI: 10.3389/frai.2021.642731. URL: <https://doi.org/10.3389/frai.2021.642731>.
- [64] Hanchao Li et al. *Pyramid Attention Network for Semantic Segmentation*. 2018.
- [65] Ali Hatamizadeh et al. *UNETR: Transformers for 3D Medical Image Segmentation*. 2021. arXiv: 2103.10504 [eess.IV].

