UCLouvain

École polytechnique de Louvain

Improving treatment plan prediction in protontherapy with transfer learning

Author: **Thomas DES TOUCHES**
Supervisor: **John A. LEE**
Readers: **Ana Maria BARRAGAN MONTERO**, **Margerie HUET**, **Christophe DE VLEESCHOUWER**
Academic year 2020–2021
Master [120] in Biomedical Engineering

Abstract

Radiation therapy in general is involved in about half of the cancer treatments. Protontherapy is developing worldwide because the ballistics of protons helps to spare the healthy tissues around tumours better than conventional radiotherapy.

Radiation therapies need high quality treatment plans to deploy their potential but these still require several hours of human work to be generated. Most of this time is spent in contouring the target volumes and organs at risk and optimising the treatment machine parameters so that the target volume is well irradiated while minimising the dose to the healthy tissues.

This thesis is in line with ongoing research, the challenge of which is to achieve a fast and accurate fully automated planning procedure. This study explores the use of knowledge acquired on radiotherapy to improve the quality of protontherapy dose maps predictions made by a U-Net model through transfer learning. One of the proposed strategies was able to improve the validation accuracy of a model trained with 10 patients by 18% as well as the precision on most of the clinical metrics of interest. It also increased the Dice similarity coefficient by 2.5% in average, 6.3% for high doses and 4.8% for low doses.

The model benefiting from transfer learning and trained with 10 protontherapy patients reached performances similar to a model trained with 40 protontherapy patients, without transfer learning. Since building a protontherapy dataset with 40 patients takes about a month, this improvement is highly valuable in the clinical practice to speed up the deployment of automatic planning methods.

Acknowledgements

First, I would like to thank the members of the MIRO lab, as a team, who helped me throughout the year. Even with the distance, I felt warmly welcomed as a part of the team always listened during our meetings.

Precisely, I want to thank my supervisor, John A. Lee, who shared his precious experience with me and was ready to answer all my theoretical questions.

I also express gratitude to Margerie Huet, who gave a lot of her time to help me to understand the computation environment as well as the implementation of the networks and the data at my disposal.

Finally, I thank Ana Barragán, who brought me her expertise and who was a great guide from start to end.

Contents

Chapter 1

Introduction

Cancer is a major plague of the twenty-first century and radiation therapy aims at treating them by irradiating tumours with photons or particles. The irradiation damages the DNA, leading to mutations that ultimately kills tumor cells. However, it also affects the surrounding healthy cells, leading to numerous side effects that can affect the quality of life of the patients. In the case of head and neck cancers, xerostomia and dysphagia are common side effects that create difficulties to swallow. On the long term, the damages to the healthy tissues can also lead to secondary cancer. Thus, the ideal treatment would highly irradiate a tumour while sparing the healthy tissues of the patient.

Protontherapy is a promising modality of particle therapy where the X-rays from conventional radiotherapy are replaced by protons. Due to their different ballistics, protons benefit from the Bragg peak, characterised by a local high energy deposition density at their maximum range, which is defined by their initial energy, as shown in figure 1.1. Well employed, the peak increases the fraction of the dose delivered to the tumour and reduces the damage on the healthy cells. Thus, in comparison with conventional radiotherapy, protontherapy has the potential to minimise the risk of complications and offer better chances of long-term survival.

However, a protontherapy center is at least 25 times as expensive as a radiotherapy center and require much more space. Coupled with the youth of the modality, it makes it difficult to access the expertise for proton planning because there are still very few centers that have accumulated experience in protontherapy. Therefore, getting a large and high quality dataset of proton plans is harder than for conventional radiotherapy. Furthermore, an increasing number of centers are breaking ground, creating a growing need for automatic ways to capture and transfer the knowledge from radiotherapy to protontherapy planning on one hand, and from centers with protontherapy expertise to less experienced ones on the other hand.

Precisely, the expertise of a center is vastly expressed during the treatment planning, which is part of the treatment procedure, as shown in figure 1.2. The

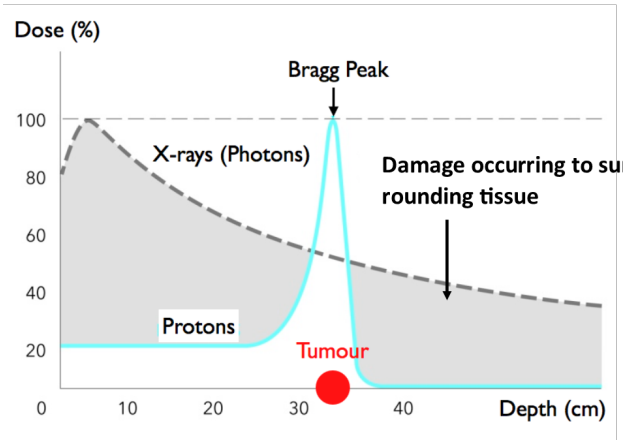


Figure 1.1: Comparison between the dose deposited by photons and protons when crossing through living tissues. The dose from photon builds up in the first centimeter before an exponential decrease. Protons deposit a lower dose before the Bragg peak, then the dose quickly rises to its maximum. Virtually no dose is deposited after the Bragg peak. Credits: [1].

complete procedure requires much work and most of it is allocated to the contouring and the planning steps. Therefore, practitioners and their teams spend hours preparing their treatment for each patient.

Humans are prone to errors and variability when they perform such tasks. Moreover, they are slow and while they spend time preparing the treatment, the disease continues to progress, reducing the chances of survival of the patients. The automation of the planning procedure could improve the general quality of the protontherapy treatments by removing the human intervention and therefore standardise the clinical practice, reduce the errors and the variability as well as ensuring the best treatment for the patients.

Research currently tackles the automation the contouring and planning tasks, this document focuses on the latter. One way to plan a treatment automatically is to first predict a dose according to the experience acquired with previous patients and then mimic it to find the parameters that will produce the dose in question These techniques, while improving each year, are still not sufficiently reliable to reach clinical practice. Here, the approach selected uses machine learning to predict doses. The goal is to train a computer to imitate good quality plans on new patients.

The predictions have to meet three main requirements, as explained in [19] First, it must be physically realistic. It is indeed crucial that the predicted dose can be delivered to the patient. Second, it must meet the clinical requirements regarding doses to the target volume and organs at risk around it. Third, the prediction must

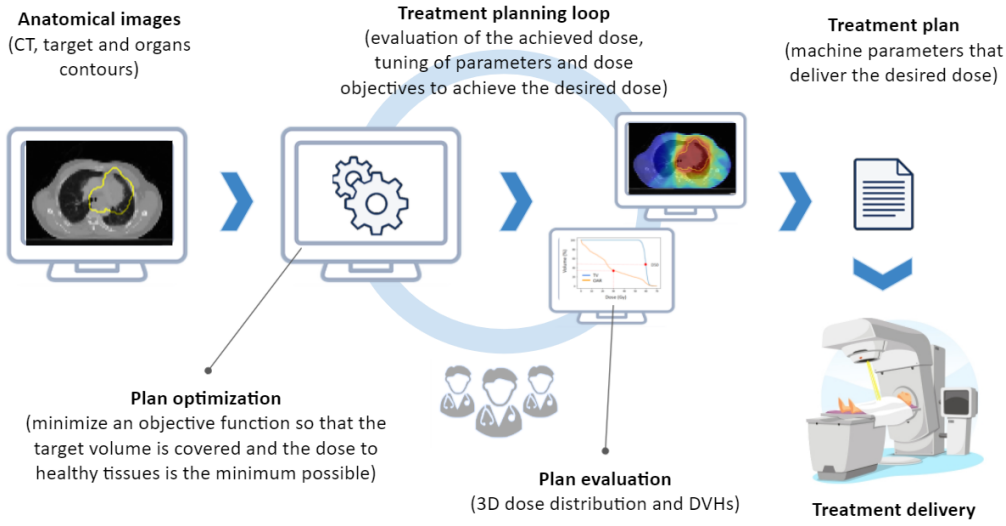


Figure 1.2: Main steps in protontherapy treatment. The first one is the imaging of the patient. With these images, a physician manually contours the target volumes and the organs at risk. Then, a physicist iteratively selects sets of beams that have to be optimised by computer each time until they get a plan that suits them. Afterwards, the treatment can be delivered to the patient. Credits: [12].

be quick enough so that it is practically usable in the clinical process.

This master thesis extends and builds upon that of Margerie Huet [19]. It addresses the transfer of expertise from experimented radiotherapy and protontherapy centers by proposing an automatic dose prediction model protontherapy that is easily shareable from centers to centers. It presents and assesses methods to transfer the learning of a neural network trained to predict conventional radiotherapy doses to a network to be trained with protontherapy treatment plans.

The document is organized as follows. Chapter 2 presents the context of conventional radiotherapy and protontherapy, with their physics, state-of-the-art treatments and planning procedure. Chapter 3 reviews the literature on the current automatic planning, machine learning, and transfer learning methods. Chapter 4, details the methods used. Chapter 5 and 6 present the results obtained and a discussion of this work. The last chapter wraps up the study with a conclusion.

Chapter 2

Context

Radiotherapy was invented more than a century ago but medical linear accelerators, which enabled modern external radiotherapy, were first used in 1953 [33]. Since then, radiotherapy has benefited from impressive improvements, whether in terms of dose sculpting, knowledge of radiobiology effects, or implementation of treatment techniques with other particles like protons. Nowadays, radiotherapy plays a crucial role in the fight against cancer.

It is involved in about half of the cancer treatments and its role is to ensure loco-regional tumor control. Radiotherapy is indeed especially important during the early stages of the disease thanks to its capability to kill the tumor before it spreads over the body by metastasis. Otherwise, complementary systemic treatments - the most popular being chemotherapy - are needed to attempt to kill these metastasis.

Radiotherapy uses ionizing radiation to treat cancer. The ionization of the DNA, which is the formation of an ion by electron removal, changes its structure. However, it is not selective: Radiation causes similar damages to cancerous and healthy cells and if it is possible to aim for a tumor, it is impossible to completely avoid the neighbouring tissues because the irradiating particles have to pass through them before reaching their target. Tumors also have microscopic insertions in the healthy tissues that need to be destroyed as well. Three scenarios can occur after the ionization: accurate repair, misrepair and inadequate repair as depicted in Figure 2.1 [12].

If a cell accurately repairs its DNA, it resumes its normal life, which is the ideal scenario for healthy cells. If the repair is inadequate, the cell dies, is inactivated or killed by the immune system. This is the goal of the tumor irradiation, but can have side effects on the healthy tissues. Finally, when a cell repairs itself but not accurately, it can survive at the cost of a genetic mutation. The mutations in the healthy cells can lead to secondary cancer in the long term, but their accumulation might also kill cell, which is useful when the cell is malignant.

Even though the radiation is not selective, there is still a massive difference

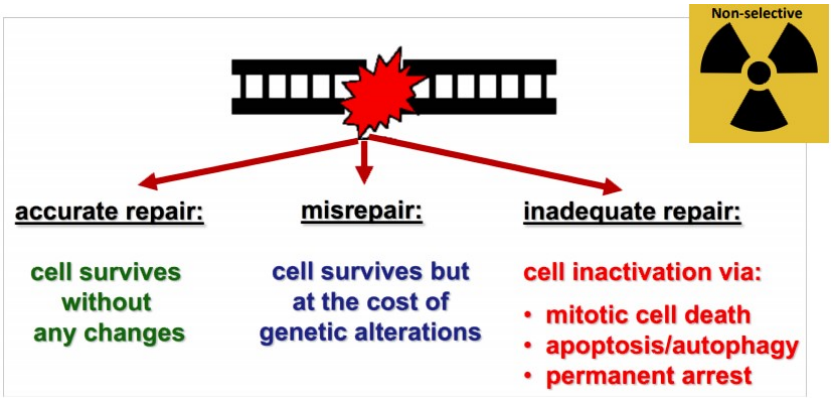


Figure 2.1: Possible scenarios after ionization of the DNA. Credits: [12]

between healthy and cancerous cells. Whereas healthy cells can generally recover from the ionization in 6 to 8 hours, cancerous cells are genetically altered by definition, and their mutations often affect their repair mechanisms. Residual errors therefore accumulate easier in malignant cells. This opens a treatment window where it is possible to control the tumor while avoiding normal tissue complications.

Knowing the principles of radiation therapy, the rest of this chapter presents the ballistics and state-of-the-art treatments for radio- and protontherapy as well as the planning procedure in protontherapy trying to be automated.

2.1 Conventional radiotherapy

2.1.1 Photon physics

Conventional radiotherapy uses highly energetic, ionizing, photons to treat cancer. They are created by linear accelerators - or linacs - that accelerate electrons to the desired energy using high tension electrodes. The electrons then hit a metal plate at the end of the linac. The absorption of the electrons by the metal produces a photon beam, typically X- or γ-rays, that is aimed at the tumor.

Photons are not electrically charged, thus they are the indirectly ionizing radiations: They do not ionize matter by themselves, but when they meet an atom, they give them enough energy to induce an ionization. According to their energy, they may act differently on the atoms they hit [12, 8].

X-rays are the least energetic photons used. They have two main types of interactions with matter, called photo-electric and Compton effects, which are similar and shown in Figure 2.2. When a photon hits an electron, it can excite or ionize their atoms by transferring them their energy. If the photon releases all its energy, is absorbed, and ejects an electron, it is the photo-electric effect. Otherwise,

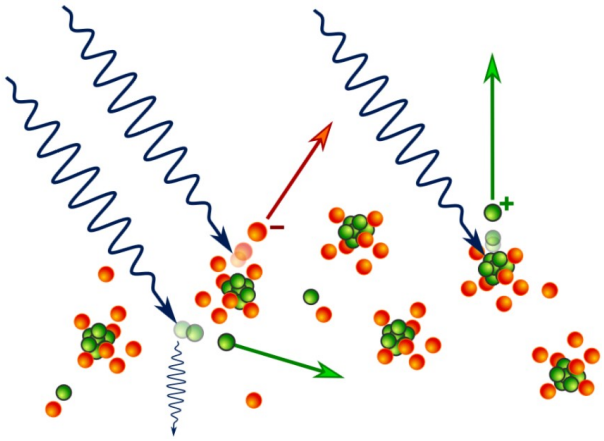


Figure 2.2: Examples of photon-matter interactions. From left to right: Compton effect, photo-electric effect and photonuclear transmutation. Credits: [28]

it can give only a part of its energy to the electron and bounce off it. This is the Compton effect, that can either excite or ionize the atom according to the energy transferred.

γ-rays are the most energetic photons used. Whereas X-rays only interacts with electrons, γ-rays are also able to interact with nuclei. When they collide with nuclei, they can cause either the production of an electron-positron pair and thus annihilation, either the release of nucleons by photonuclear transmutation, shown in Figure 2.2. These interactions deteriorates the treatment quality but γ-rays are sometimes needed to reach deep tumors because the highest the energy of photons, the longest their expected path.

Atoms naturally fill the gap left by the excited or ionized electron to recover their equilibrium. They emit either a photon (fluorescence) or an outer electron (Auger effect) to reduce reduce the energy level of their electrons.[21]

Also, ionizing photons have a low linear energy transfer (LET) which means that they need a lot of interactions to transfer all their energy to the matter they go through [15].z Therefore, they deposit their energy in a diffused manner, disseminating small damages on the DNA.

Finally, the exponential decay of the fraction of the dose with the depth, visible in the Figure 1.1, is due to the strict binary nature of photon-matter interactions: Either the photon hits an atom and transfers its, either it passes next to it without interaction. Thus, at each distance step, a fraction of photons deposit a dose and disappear. Therefore, if all photons are present when the beam enters the body, the more distance they travel, the least they are, and thus the lower the dose they deposit all together.

2.1.2 State-of-the-art treatments in radiotherapy

Now that the physics of photons were laid out, it is time to expose some state-of-the-art treatments for radiotherapy in head and neck cancer. The most advanced clinical practice is the adaptive radiation therapy, abbreviated by ART [34]. It is based on Intensity Modulated Radiotherapy (IMR) and Image Guided Radiotherapy (IGRT). Together, they lead to great conformal dose with steep gradients at the edge of the tumor. This results in high, homogeneous dose on the target volume (TV) and low dose on the organs at risk (OARs). The target volume mainly contains the tumour and the OARs are the important organs that could be damaged by the radiation, the brain stem or the esophagus for example.

However, such a gradient also means that a position error can easily lower the dose delivered to the target volume and increase the toxicity to the healthy tissues. The goal of ART is to regularly image the patient to adapt to anatomical changes occurring during the treatment, the main ones being weight loss, tumor shrinkage and tissues displacement. Doing so enables to account for positional errors accumulating throughout the treatment and to adapt the plan to keep a nice dose distribution.

Volumetric Modulated Arc Therapy (VMAT) is also a radiation therapy modality used nowadays [17]. The principle is to shoot photons in multiple angles, while the nozzle performs a continuous rotation around the patient. Therefore, they are a lot more shooting angles and thus possibilities to have a higher quality plan. Still, it brings complexity during the treatment planning and the performance of VMAT is highly dependent on the optimisation of the plan [4]. Currently VMAT enables shortened delivery times that increase patients comfort and reduce their movements during delivery [17].

Another modality is currently under investigation. Response adapted ART (R-ART, in opposition with anatomy adapted ART, A-ART) that was described earlier, goes one step further by adapting the treatment not only to anatomical changes but also to the response of the patient [6]. R-ART proposes indeed that the dose to the target volume could be increased if the tumor shows resistance to treatment or instead decreased if the side effects on the patient, like significant weight loss or surrounding normal tissues deterioration, are too heavy.

2.2 Protontherapy

2.2.1 Proton physics

Protontherapy does not use photons but protons as its name tells it. They are produced from dihydrogen molecules in a chimney at the center of a cyclotron

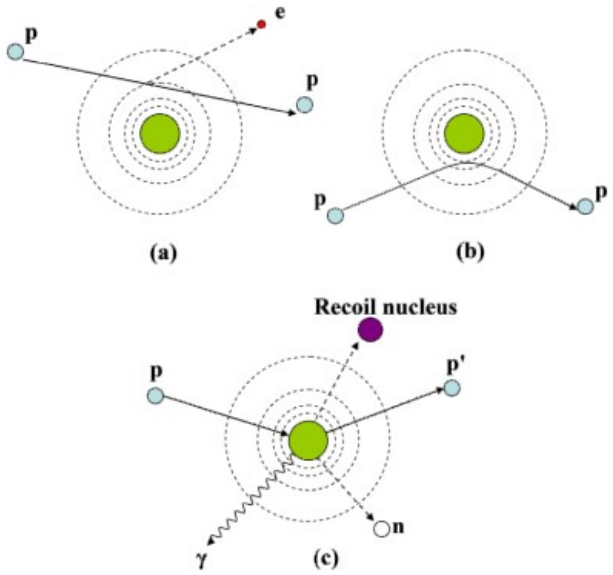


Figure 2.3: Interactions mechanisms between protons and matter they encounter. (a) Collision with bound atomic electron. (b) Elastic scattering. (c) Nuclear interaction. Credits: [23]

[12]. The chimney is composed of a hollow electrode that receives a dihydrogen flux which is dissociated in two non charged hydrogen atoms. Then, high energy electrons, emitted by the electrode, ionize the hydrogen to transform it into protons.

Unlike photons, protons then need to be accelerated to receive energy. Since they are heavier than the electrons accelerated in conventional radiotherapy, they need a longer distance to speed up. Therefore, a linear accelerator would not be practical because it would be too long. Instead, cyclotrons are used. They use a strong magnetic field to force the proton flux to draw growing circles during the acceleration. The drawback is that cyclotrons tend to be too cumbersome to rotate around the patient in the treatment room. Thus, they lie in a separate room and the proton flux is brought to the treatment room via vacuum tubes in which it is maintained in the center by magnets that steer and focus the beam. Protons would indeed interact with the surface of the tube if they hit it.

While photons have no mass and are neutral, protons are heavy, positively charged particles. Their charge makes them directly ionizing particles. Protons have three interaction mechanisms, depicted in Figure 2.3: Collision with bound atomic electron, elastic scattering and nuclear interactions.[23]

First, inelastic collisions, represented in Figure 2.3(a), are also called inelastic electromagnetic scattering because they result from the Coulomb interactions. The attraction between the positive charge of the proton and the negative charges of bound atomic electrons is the main cause of energy deposition by protons. Collisions

are divided in soft and hard categories, according to the distance with the nucleus of the target atom. When a proton crosses an atom at relatively long distance, it interacts with it as a whole. It is a soft collision because the interactions are weaker due to the distance. Such interactions can excite or ionize valence electrons. A hard collision occurs when a proton crosses the target atom at short distance, interacting mainly with a single bound electron and ejecting it with a high kinetic energy which may lead to secondary radiations. In both cases, the proton loses a part of its energy, that is transferred to the electron. Soft collisions are much more frequent than hard ones but they involve lower energy transfer. Thus, both collisions have a similar contribution to the energy deposition. Since protons are a lot heavier than electrons, they are merely deviated during the inelastic electromagnetic scattering.

Second, elastic scattering, depicted in Figure 2.3(b) can result from the Coulomb or nuclear interactions. The main difference with inelastic scattering is that, in this case, the proton, which has a positive charge, is repelled by the positive charge of the target nucleus or other nuclear forces. This scattering is called elastic because in most cases, the energy transfer between the proton and the atom is negligible because the latter is a lot heavier and has thus more inertia. If the target is an hydrogen atom, however, the energy of the incident proton can be shared in halves with the hydrogen atom, but such energy transfer occurs most times due to the nuclear interactions, covered in the next paragraph. Elastic interactions do not have a significant contribution to the dose distribution but are the main cause of angular deflections of protons.

Third, nuclear interactions, shown in Figure 2.3(c) are rare but still have a significant impact on the dose deposited by the protons because they are incorporated to the nucleus, therefore transferring all their remaining energy at once. The result of the incorporation is a transmutation of the atom. For example, a Carbon atom (C) that incorporates a proton becomes a Nitrogen atom (N) for a short instant before entering in fission. The fission can lead to a variety of outcomes, the most frequent being the expulsion of a secondary proton and neutron, leaving a recoil nucleus in place of the atom. Whereas secondary protons deposit a significant local dose like the primary ones, the neutrons have far fewer interactions and are more of a trouble for radioprotection. Nuclear reactions catch protons on their way to the target volume and attenuate the Bragg peak, explained in the next paragraph.

In contrast to photons which have an exponential decay of their relative dose with the depth, Figure 1.1 shows that protons tend to deposit their energy in a constant way before reaching a peak, called Bragg Peak, leaving no dose afterwards. The so called range of proton is due to the first type of proton interactions discussed, inelastic scattering, that occurs so often that it results in a quasi continuous loss of energy from the protons, called the stopping power. The latter is mainly defined by the electronic density and the mean excitation energy of the medium that the

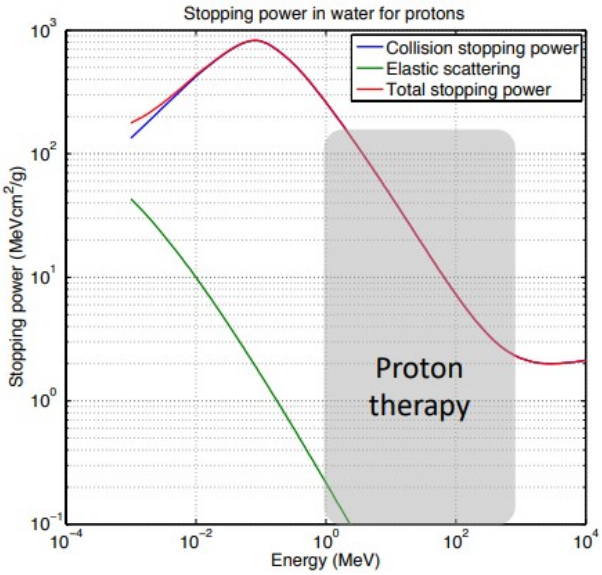


Figure 2.4: Cumulative stopping power for protons in water in logarithmic scales. The gray rectangle highlights protontherapy energies. The inelastic scattering has the major contribution. Credits: [12]

proton passes through. As shown in Figure 2.4, at protontherapy energies, the stopping power decreases as the square of the energy of the proton. Thus, the more depth the proton travels, the less energy it has, and the more the stopping power grows. As the stopping power reflects the energy that protons loose, they release most of their energy at the end of their path. Therefore, the depth of the Bragg peak is defined by the initial energy of the protons and the stopping power of the medium they pass through. Although, the peak is not sharp because the inelastic collisions are random events that lead to a variability of range between photons, called range straggling.

Finally, it is possible to use protons with various initial energies to spread out the Bragg peak over a larger target, as pictured in Figure 2.5. The main advantage of proton over photons is that they leave less dose before and almost no dose after the target volume, which reduces the risks of normal tissue complications. One last difference between photons and protons is related to the way they deposit energy. As written in the section about photon physics, they tend to deposit their energy in a diffused manner whereas other ions like carbon deposit their energy in a much compact way, leading to different damages to the DNA. That is where the radiobiological effect (RBE) comes in handy. It is the ratio between physical doses of photons and the studied radiation that would lead to the same biological effects. The RBE of photons is thus defined as one whereas that of protons is around 1.1 with a slight increase near the maximum depth [12].

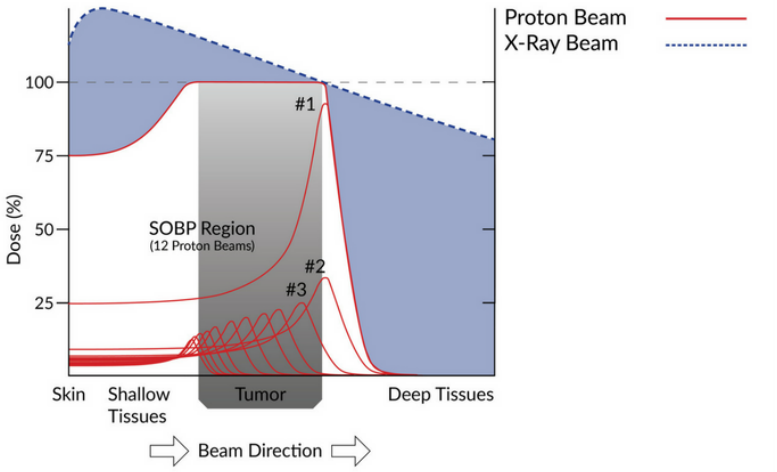


Figure 2.5: Spread out Bragg Peak resulting from the addition of the contribution of each proton beams. Credits: [29].

2.2.2 State-of-the-art treatments in protontherapy

Like conventional radiotherapy, state-of-the-art protontherapy methods are based on Intensity Modulated Protontherapy (IMPT) [16]. In this case, the innovation took advantage of the depth dimension to introduce the Spot-Scanning modality. The principle is to use a thin beam that "paints" the tumor as illustrated in Figure 2.6. As different energy implies different depth of Bragg peaks, several painting with various energies are used to cover the whole tumor slice by slice. This enables having a dose that fits the clinical target volume more tightly in comparison with regular IMPT, where the whole tumor is irradiated at once and thus where there is less liberty regarding the decoupling between the energy of the protons and their direction.

Also, future modalities are being developed. One of them is the Proton Arc Therapy (Arc-PT). It might not lead to a mean lower integral dose to the OARs than current methods but the radiobiological effects of a low, spread, dose on healthy tissues is generally better than the same dose focused on a part of them [5]. Actually, tissues are not hurt linearly with the dose they receive. Cells sustain and heal the first damages to their DNA more easily than when they accumulate, the lesser the damages per cell, the higher the chances to adequately repair each of them. Therefore, PAT might lead to a better therapeutic index than IMPT.

Finally, there is a "flash effect" that occurs for a yet unknown reason when irradiating a patient at dose rates higher than $40Gy/s$ [20], where "Gy" is the symbol

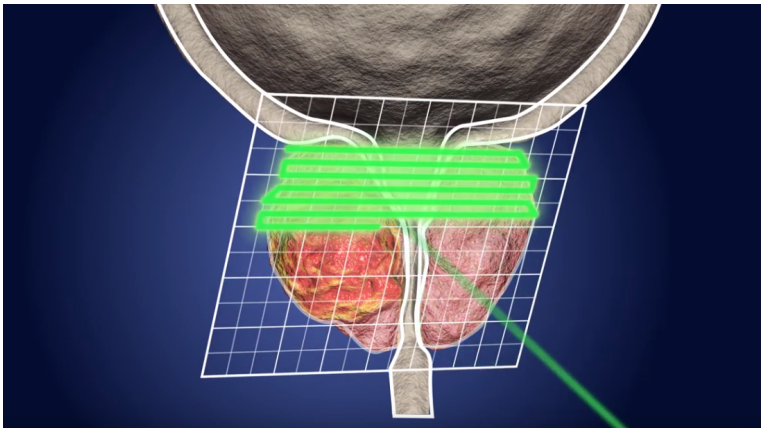


Figure 2.6: Illustration of Spot Scanning protontherapy for a prostate cancer example. Credits: [7]

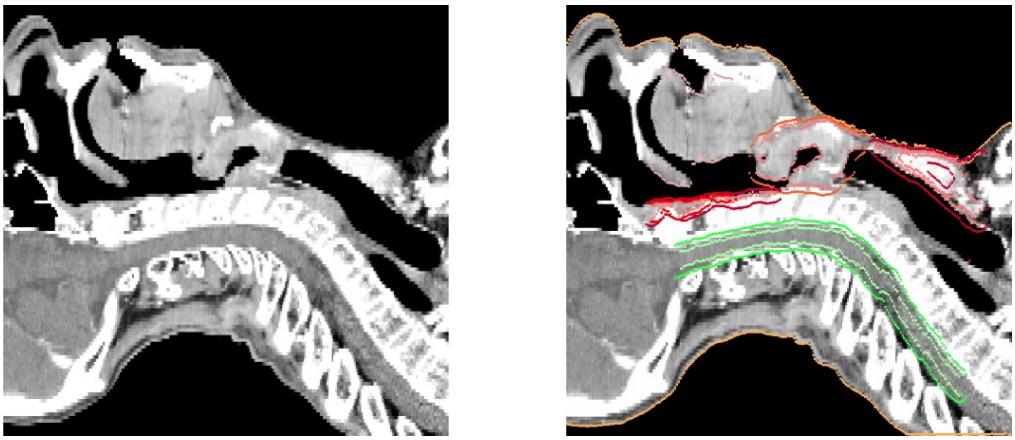
of the Gray unit and $1Gy = 1J/kg$. From this rate, healthy tissues encounter less risks for the same dose than in low rates. Yet, there are many technical challenges between this modality and clinical practice. It is still impossible to marry high dose rate, energy modulation and precision which are all needed for an efficient flash treatment. The current approach consist to get rid of the energy modulation and work with maximal energy protons that pass through the patient and not o take advantage of the Bragg Peak, leading to plans already better than VMAT. [20, 35].

2.2.3 Treatment planning procedure

Now that the proton physics and the state-of-the-art methods were detailed, the main steps of treatment planning for radiation therapy, and in particular for protontherapy, are exposed. The procedure is divided in four parts detailed in the following paragraphs.

Patient imaging A 3D representation of the patient's region of interest is produced using a CT scan. The scan is composed of several 2D slices in which each voxel has an intensity proportional to the attenuation of the X-ray by the electrons of the tissues inside it. The intensity is rescaled to the Houndsfield Unit [HU], a unit where air and water have respectively -1000 and 0 as values in normal temperature and pressure conditions. While it is not possible to place the patient identically on the couch during the imaging and delivery, it is crucial to reduce position uncertainties as much as possible during this step because they would translate to systematic errors during the treatment. Therefore contention masks and lasers are used to precisely position the patient. The result of the imaging step

is shown in the Figure 2.7a.

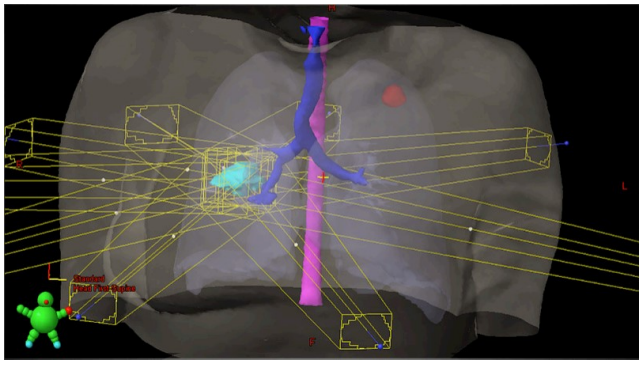


(a) CT scan slice. (b) Slice with some contours of interest.

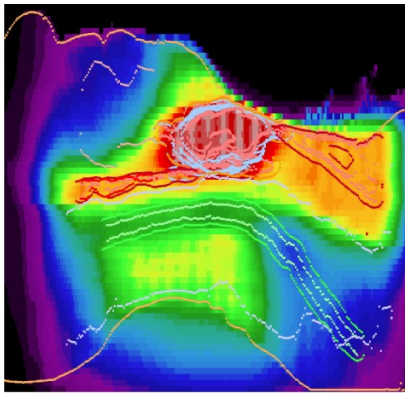
Figure 2.7: Sagittal section of a head and neck cancer patient. Credits: [12].

Contouring and dose prescription When the physician has the images of its patient, it can start manually contouring its body, tumor and some surrounding organs at risks. However, the frontier between cancerous and healthy tissues is not sharp because tumors have microscopic insertion in the healthy tissues. Thus, it is necessary to contour not only the Gross Target Volume (GTV) but also the Clinical Target Volume (CTV) that encompasses such insertions, which require medical expertise since they depend on the tumor location and the patient anatomy. Too tight margins and some cancerous cells might not receive a sufficient dose to die, too large and the toxicity to the patient is increased needlessly. A last margin is applied to the CTV, that considers the position error as well as motion during treatment. The higher the uncertainties, the larger this Planning Target Volume (PTV), emphasizing the importance of an accurate positioning system evoked in the previous paragraph. If PTV margins work fine in conventional radiotherapy, they might fail to ensure a satisfying coverage of the CTV for protontherapy treatments. Instead, robust optimisation is generally used to generate IMPT plans directly on the CTV [22]. Once the contours are done, as shown in Figure 2.7b, the physician must define a prescription for the tumor and rules to spare the organs at risk.

According to the structure of the organ, the rules may either be maximum local or mean doses. In series organs, typically in the nervous system, the priority is indeed to reduce the maximum local dose so that every healthy cell has the lowest chance of being hurt because the death of a single cell disable other cells to work properly. On the other side, in parallel organs like the lungs, it is better to reduce



(a) Treatment plan for a lung tumor.



(b) Sagittal section of a dose map for the treatment of a head and neck cancer.

Figure 2.8: Examples of treatment plan and dose map used in protontherapy. Credits: [12].

the mean dose because the death of a cell has far less impact on the others and that is the survival of the organ as a whole that matters.

Plan Generation With the image, contours and prescriptions, a plan is ready to be generated. A dosimetrist optimises the objective function, which is generally a weighted sum of all constraints on the various contours, by iteratively tuning parameters of the plan until it reaches the desired quality. Sometimes, constraints must be added, modified, or removed to guide the optimisation, which also requires expertise and may lead to plans with variable quality.

An important point of plan generation in protontherapy is the robust optimisation [11]. The Bragg peaks must indeed be robustly placed on the tumor but conventional margins are not sufficient in general to ensure an adequate coverage due to high dependency to the stopping power and the steep dose gradients. A popular robust optimisation algorithm is minimax. Its principle is to minimize the penalty of the worst case scenario, among those that are physically realisable, taking into account the intervals of possible range and setup errors. Thanks to this consideration for realisable errors only, minimax is able to be less conservative than other robust optimisation algorithms while offering the same guaranties on the CTV coverage.

The present thesis focuses on the automation of this cumbersome step, trying to predict a dose to later find a plan that would match it by inverse optimisation. A plan as well as a typical dose map are shown in the Figure 2.8

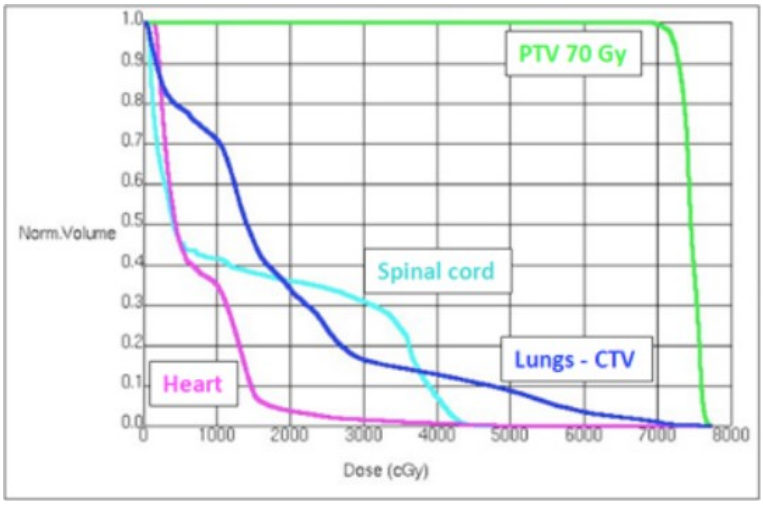


Figure 2.9: Dose Volume Histogram of a treatment plan for a lung cancer. The X-axis is the dose and the Y-axis is the fraction of volume of the regions of interest. Credits: [12].

Plan Evaluation In the clinical practice, plans are evaluated by referring to their Dose Volume Histogram (DVH) curves, as shown in the Figure 2.9. A DVH curve is a graphical representation of the fraction of the volumes, organs and targets, that receive a given dose. They determine the DX metrics, that are the minimum dose deposited in the X% hottest volume of the region of interest. The most common are the D95 and D5 that serve as lower and upper bound respectively for the target volumes, and the D2 (the maximum dose in almost all the volume) and D_{mean} (the mean dose) for the series and parallel OARs, respectively. If the plan is validated by the dosimetrist and the physician, the treatment can be delivered.

The plans are optimised with the aim of being robust to setup and range uncertainties but planning by itself does not bring any guaranty. Unfortunately, most clinical treatment planning system nowadays sacrifice the precision for the sake of speed by using analytical models of beams, even though they are known to be less reliable in heterogeneous medium, that are common ground in the human body. Robustness evaluation is an important step in the plan evaluation because it enables to check that the uncertainties stay in an acceptable range.

A precise robustness evaluation method was developed using Monte Carlo simulations [32]. At the cost of longer computation time, Monte Carlo algorithms perform a precise physical simulation of protons going through a patient. It is therefore possible to simulate the impact of range and setup errors, among others, on the quality of the plan more precisely than with analytical methods. The advantage is that precise knowledge of the errors enables to be less conservative regarding the margins on the target volumes and thus to spare more healthy tissues.

Chapter 3

Literature review

This section reviews some articles from recent literature about dose prediction in radiation therapy. First are presented the current clinical automatic planning methods, then the articles about machine learning in such context with an emphasis on U-Net, and finally transfer learning literature for deep learning architecture.

3.1 Current methods for automatic planning in radiation therapy

Head and neck cancer treatment are especially hard to plan due to the sometime large target volumes involved as well as numerous important radiosensitive organs close to them and multiple dose prescription levels [25]. To reduce human labor during treatment planning, various methods already reached clinical practice [19]. Although they do not totally automate the workflow, they still offer the possibility to quicken the process. Automatic planning methods are grouped in two categories, named Object- and Knowledge Based Planning.

Objective Based Planning (OBP) methods optimise the treatment according to a set of objectives generally defined with DVH metrics. A regular plan optimiser already enters this category as it enables the automatic optimisation that would not be feasible in reasonable time by humans. Another example of OBP is the regional optimization using planning scripts developed in [36]. It specifically detects hot and cold spots in the dose maps and modifies the beams that contribute to those local spots to reduce them. OBP methods are currently limited by the need of human intervention to define the objectives to be reached. [37].

Knowledge Based Planning (KBP), on the other side, uses sets of patients that are similar regarding their pathologies and geometries as well as the modality of treatment to infer DVH metrics on a new patient. Similar cases often lead to similar plans but KBP methods still need handmade features to match patients [31].

Whereas it does not output a dose map, KBP can be handy when selecting the treatment modality for a patient because it can infer DVH metrics for radiotherapy and protontherapy without requiring full plans that are expensive to prepare.

In both cases, current methods still non negligible human intervention to output their results which may lack spatial information. Therefore, dedicated methods to predict dose maps are currently under development.

3.2 Machine learning for dose prediction

Machine learning (ML) is a helpful tool make predictions. It is a branch of artificial intelligence that has been developing algorithms that are able to acquire knowledge from the data that is presented to them in order to perform tasks such as classification or regression making.

Deep learning (DL) is a type of ML that gained in popularity in the recent years and that relies on networks of artificial neurons, as illustrated in Figure 3.1. Neurons output the result of simple non linear mathematical operations on their inputs and, when linking the outputs of some with the inputs of others, they form a network that is able to perform complex functions. This mechanism enables the first layers to acquire relatively general knowledge on the data whereas, building on the previous layers, the last layers develop a knowledge more specific about the task at hand. The network can change its global function to perform a given task by tuning the weights attributed to the inputs of the neurons.

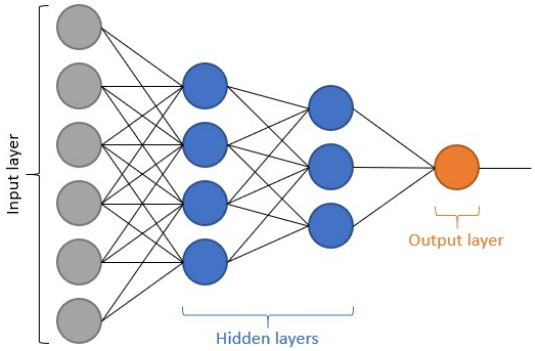


Figure 3.1: Schematic of a neural network.

ML mainly involves two types of learning, supervised and unsupervised. The former framework gives the model its inputs as well as its expected output whereas the latter only gives the inputs. Like many ML techniques, neural networks can learn by example in a supervised framework: Multiple input-output pairs are presented successively to the network, and an error is computed, which is the

difference between the true and predicted outputs. Then, an optimiser tries to minimise the error, by changing the weights that the neurons give to their inputs with a method called gradient back-propagation.

The non linearity of the neuron functions is often performed by an activation function that is applied after the weighted sum of the inputs. The most common activation function is called "ReLU" for Rectified Linear Unit and $ReLU(x) = max(0, x)$. This simple function is handful because it is quick to compute and generally sufficient.

Neural networks are composed of two main types of trainable layers: dense and convolutive. Dense layers were invented first and, as their name states, two successive dense layers have dense connections with each other, that is, a neuron of a dense layer receives inputs from every neurons of the previous layer, they are fully connected. If two successive dense layers have x and y neurons, then they are $x * y$ connections between them, each with a unique weight.

In contrast, in convolutive layers, neurons receive inputs only from a few local neurons through a convolution kernel. The kernel has predefined shape and size that determine the receptive field of a neuron. The principle is to significantly reduce size of the receptive field of the neurons, which is the number of neurons of the previous layer from which it receives inputs. Also, the weights given by the kernel to the inputs are shared within a layer. Therefore, if two successive convolutive layer have x and y neurons and a kernel of size z (with $z << x$) linking them, there are only $y * z$ connections with z weights in total between the layers.

There are numerous types of convolution layers, this paragraph present some 3D-convolutions of interest. The simplest convolution layer take its input in a cubic kernel of adjacent voxels. To keep the same shape in successive layers, it is necessary to add a padding around the input to allow the kernel to overflow outside the image. Common choices are the zero-padding that adds zeros around the image or the mirror padding that reflect the image past it boundaries.

But there also exist dilated convolution for which the kernel is not composed of adjacent voxel. They introduce a parameter called the dilation coefficient that defines the space between each voxels in the kernel, as shown in Figure 3.2. A dilation coefficient of 1 correspond to a regular convolution, a coefficient of 2 means that there 1 voxel out of 2 is taken in every direction.

Artificial neural networks that comprise convolutive layers are called Convolutive Neural Networks, abbreviated by CNN. They are the main reason for the recent success of deep learning. An example of CNN architecture is displayed in Figure 3.3. As visible on this image, the last layers of CNN are often dense.

Another type of not trainable layer plays a crucial role in CNN: Max pooling. These layers simply output the maximum value in a pool on local inputs of predefined shape and size. They enable to reduce the resolution of the inputs.

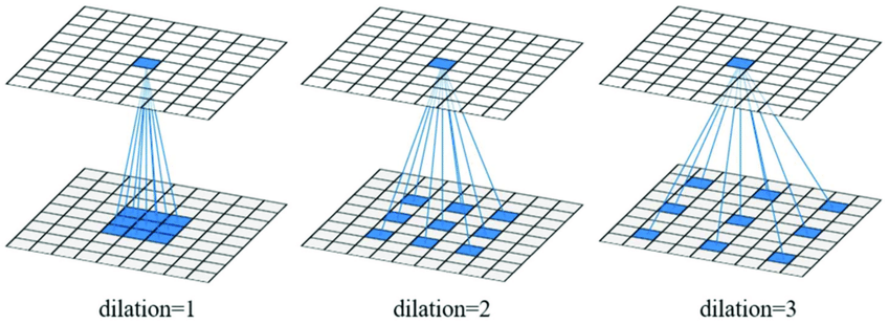


Figure 3.2: Illustration of a dilated 2D-convolution. Credits:[10].

Also, concatenation layers concatenate the output of at least 2 layers into a single output.

Finally, normalisation plays a key role in CNN. Layers that fulfill this function enable to rescale their input linearly. There once again exist several type of normalisation layers but only the instance normalisation is used in this work and therefore presented. An instance normalization layer normalizes each channel it receives by subtracting its mean and dividing by its standard deviation. Then, it re-offsets and scales the channels with trainable parameters.

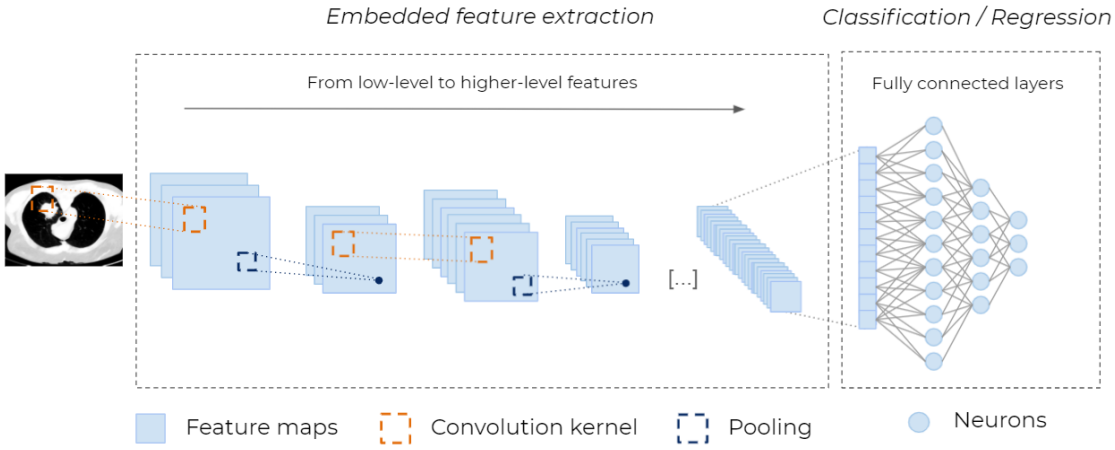


Figure 3.3: A Convolutive neural network used in medical imaging. Credits: [2].

There exists loads of CNN architectures specialized for various tasks, and one of them is the U-Net, invented by O. Ronnenberger [30]. U-Net is a family of architectures that was initially intended for image segmentation and that share the same global organisation, shown in Figure 3.4. U-Net is inspired from Auto-Encoder, which is composed of two successive parts called encoder and decoder.

The purpose of the auto-encoder is to significantly compress domain specific

data with an encoder in order to store or transmit it efficiently and then decompress it with low loss by the decoder. The encoder sequentially reduces the resolution of the input image, which may have several channels, to gather information at different scales. The decoder receives the data compressed by the encoder and increases its resolution sequentially to reproduce the input image with high fidelity.

In U-Net, the goal is not to reproduce the input after a compression but to output a map corresponding to the task in hand, the predicted dose map in the case of this document. Two major modifications differentiate U-Net from Auto-Encoder. The first one is the bottleneck, it is a third part that sometimes refines the information between the encoder and the decoder. The second one is the introduction of skip connections that bridge non successive layers. Specifically, they bring the inputs from encoder to decoder layers that share the same resolution. They also carry the error gradient the other way around during training.

U-Net quickly showed to suit for predictions both in radiotherapy [3] and protontherapy [24]. M. Huet highlighted the convergence of other architectures to the U-Net format [19] for this task. One of the architecture that have shown state-of-the-art results for dose prediction in head and neck cancer is the hierarchically densely connected U-Net (HD-UNet), a refinement of the U-Net that comprises parts of an other popular network, DenseNet [25, 19]. HD-UNet is more precise than the networks it was inspired of, makes predictions 4 times faster than DenseNet and uses 12 times less trainable parameters than U-Net.[25]

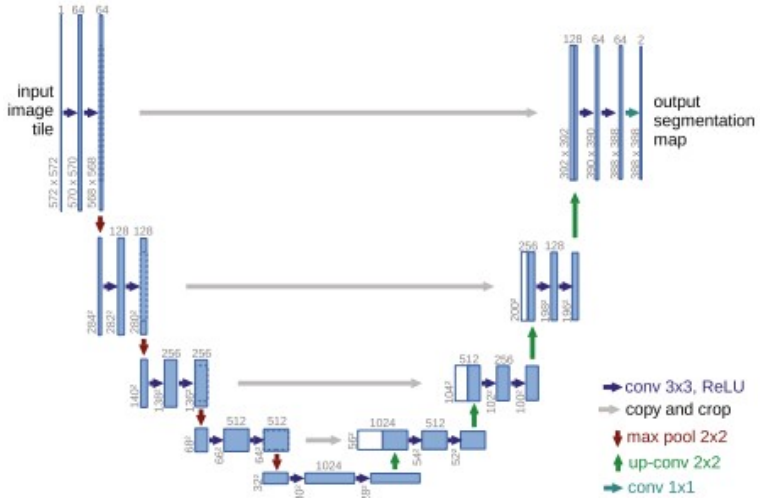


Figure 3.4: U-Net architecture, as depicted in [30].

With a proper training on a database of previously treated patients, deep learning can be used for automatic planning. A neural network, like the U-Net, could indeed be used to take as input the anatomy of a patient, to map it with

and output the optimal dose distribution the patient should receive.

3.3 Transfer learning

According to Pan and Yang[26], "transfer learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task". The main application of transfer learning is thus to learn basic knowledge about a task (source task) with a first dataset composed of abundant but low quality examples (source domain) before training the model on a second dataset of higher quality but rare data (target domain) to perform a related task (target task). It is useful when not enough high quality example are available to reliably train a model for a task.

Transfer learning has been successfully used with CNNs to compensate the common lack of medical data. For example, domain adaptation through transductive transfer learning for medical imaging tasks like brain lesion segmentation substantially improved the results of the model in [13]. In the case of transductive transfer learning, the source and target tasks are the same, only the domain changes and therefore, theoretically, no labeled data in the target domain are needed whereas there are numerous in the source domain [26].

Actually, it was applied in the context of radiation therapy dose predictions to adapt a generic model to different treatment planning practices with less than 30 patients per style [18]. To do so, all but the last layer where transferred from the generic model. The last layer was randomly reinitialised and the first half of the model was frozen, only the second half could learn to fit other practices. This is an inductive transfer learning where the source and target domains were the same but the learning tasks slightly different. This means that labeled data in the target domain are needed to induce the objective of the target learning task [26].

In the context of protontherapy, plans in conventional radiotherapy can play the role of the source domain because they are relatively similar in their physics, principles and goals. Predicting radiotherapy doses would be the source learning task whereas predicting doses in protontherapy would be the target learning task. The advantage of using radiotherapy as the source domain lies in the fact that it is older and better established than protontherapy, the target domain, more good quality plans have been generated in this modality.

Since the functions to estimate, the dose maps, are different but related in the source and target tasks (the same patient should not yield the same maps in both tasks), and because the source domain contains a lot of labeled data, this work places itself in the inductive transfer learning category. Therefore, it is plausible that learning basic knowledge about radiation therapy and on dose maps prediction might ease the learning of the model on a related task in protontherapy, where less examples of good plans were produced.

Chapter 4

Methods

This thesis explores the possibility to improve the performance of a U-Net when predicting doses in protontherapy for head and neck cancer. In this chapter are detailed the methods applied to perform such study. Precisely, the network, data, patching system, as well as training and transfer strategies are explained successively.

4.1 Data

Two datasets, one for each modality were needed to enable learning on radiotherapy and protontherapy. They were similar but they had differences that were significant when trying to match them for the transfer learning. Neural networks, like U-Net, need inputs that always keep the same format to work and this was not the case originally, as explained after the description of the datasets.

The first dataset is composed of 200 conventional radiotherapy treatment plans for head and neck cancer from the Open Knowledge-Based Planning (OpenKBP) Grand Challenge organised by the American Association of Physicists in Medicine (AAPM) [27]. The patient were treated via IMRT with 9 equispaced coplanar beams, placed at 0°, 40°, ..., 320°. The original data is available on The Cancer Imaging Archive (TCIA) [9]. For each patient, there is a CT scan, some OARs and targets. The data was cleaned and prepared to be easily usable: The CT were downsampled to 128 x 128 x 128 voxels and the contours were factored into some boolean 3D tensors where each voxel set to 1 belongs to the corresponding contour.

The dataset provides a "possible dose" mask, with the voxels that can receive a dose. It also labels the brain stem, spinal cord, right parotid, left parotid, larynx, esophagus and mandible for a total of 7 OARs. Additionally, it contains 3 PTVs of 70, 63 and 56 Gy respectively for the gross disease, intermediate-risk target, and elective target volumes. PTVs were clipped to be at least 5mm from the patient's

surface. The matrices of contours that were not provided in the original data were set to 0. Finally, each patient has its ground truth dose distributions that was generated with equal complexity and plan physics.

The second dataset is smaller and composed of 60 patients treated for bilateral head and neck cancer by protontherapy. The plan were generated by M. Huet for [19] in RayStation (Raysearch Laboratories, AB) for pencil beam scanning delivery. All plans used the same couch and beam configurations: four beams with gantry angles of 60^o, 120^o, 240 and 300^o. The couch angle was of 10^o for the two first beams and 350^o the two last beams. The contours used for the planning are detailed in the next paragraph. The plans were generated using Monte Carlo dose calculation and minimax optimisation, presented in [11], on the CTV volume with 21 scenarios (2.6% proton range error and 4mm systematic error in the three space directions). To maximise the consistency of the generated plans, they were manually generated following a protocol, in the same center, by the same dosimetrist.

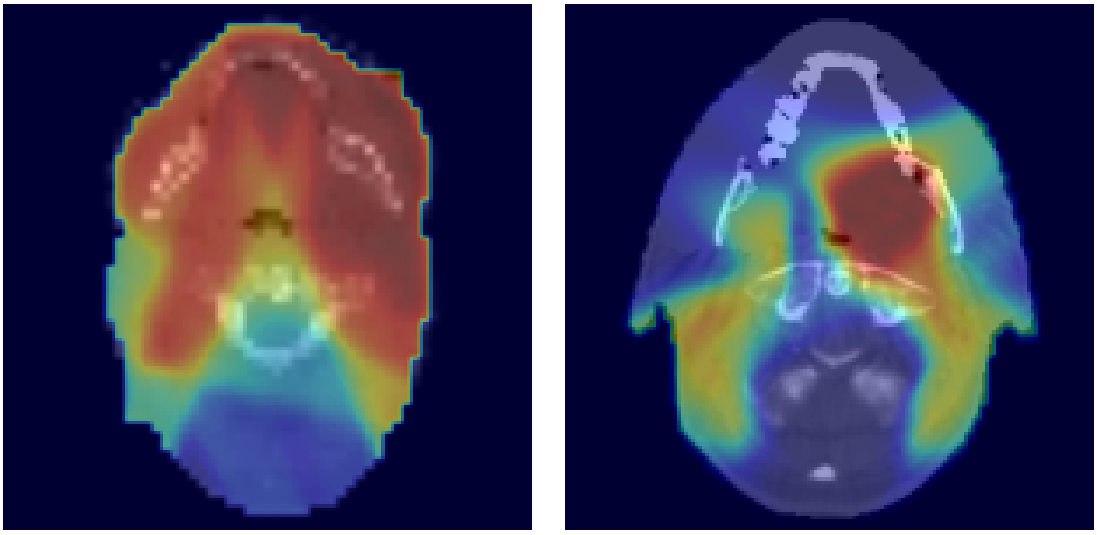
A major difference between the datasets resides in the labeled contours. The protontherapy patients presents a contour of their body. Also, they have 13 OARs: brain stem, upper esophagus, glottic area, oral cavity, left and right parotid glands, inferior, middle and superior pharyngeal constrictor muscles, spinal cord, left and right submandibular glands and the supraglottic larynx. There are only two target volumes in this dataset with 54.25 and 70Gy of prescription on the elective nodal volumes and the tumor, respectively.

Figure 4.1 display doses from the radiotherapy and the protontherapy datasets, and Table 4.1 summarises the differences in terms of information between both datasets. It is necessary to find a way to match them in order to fit the inputs with the network and perform an adequate transfer learning. To do so, only similar channels were kept. There is one CT scan per patient in both cases thus they were matched. The first dataset had a "possible dose mask" that was close to the "body contour" of the second.

For the Target Volumes, even though there is a difference between photon and protontherapy, it is possible to order them by value ($54.25Gy < 56Gy < 63Gy < 70Gy$) and thus to keep them in a single channel no matter their number by simply keeping the numerical value of their prescriptions instead of binary masks.

However, giving a scalar value to OARs to order them makes no sense. Therefore it is necessary to split them in channels. Six of the seven channels of the photon dataset could be matched with those of the proton dataset: Brain stem, spinal cord, left and right parotid glands matched with their homonyms whereas the esophagus and larynx of the radiotherapy dataset matched with the upper esophagus and the supraglottic larynx of the protontherapy dataset, respectively.

In total, nine channels were chosen as input for the network and more details about the transfer from radiotherapy to protontherapy are given in the section 4.5.



(a) Source: Radiotherapy (b) Target: Protontherapy

Figure 4.1: Example of dose distributions in the source and target datasets

Table 4.1: Number of channels present for each patient in both datasets and those kept as network inputs, grouped by category. *channels were merged due to the possible numerical ordering.

Channels	CT scan	external contours	OARs	TVs
Photon	1	1	7	3
Proton	1	1	13	2
Network inputs	1	1	6	1*

4.2 Network architecture

The U-Net architecture was selected for its great performances in the radiation therapy dose prediction tasks, as was detailed in the literature review.

It is organised in 3 successive parts: encoder, bottleneck and decoder, for a total of 28 primary layers. The network is indeed hierarchical and composed of 2 types of layers: simple, and composite. As their names suggests, simple layers have no internal layers whereas composite layers, also called blocks, are composed of a few internal layers. Blocks and layers are detailed alongside the parts they belong to in the next paragraphs. Figure 4.2 shows the detailed architecture of the network and the few next paragraphs explains it with more precision.

The encoder is a succession of convolution blocks and max pooling that alternate 4 times. It takes a 3D image with size $96 * 96 * 64$ and the 9 channels presented in Table 4.1 as input. Convolution blocks refine the information whereas max pooling

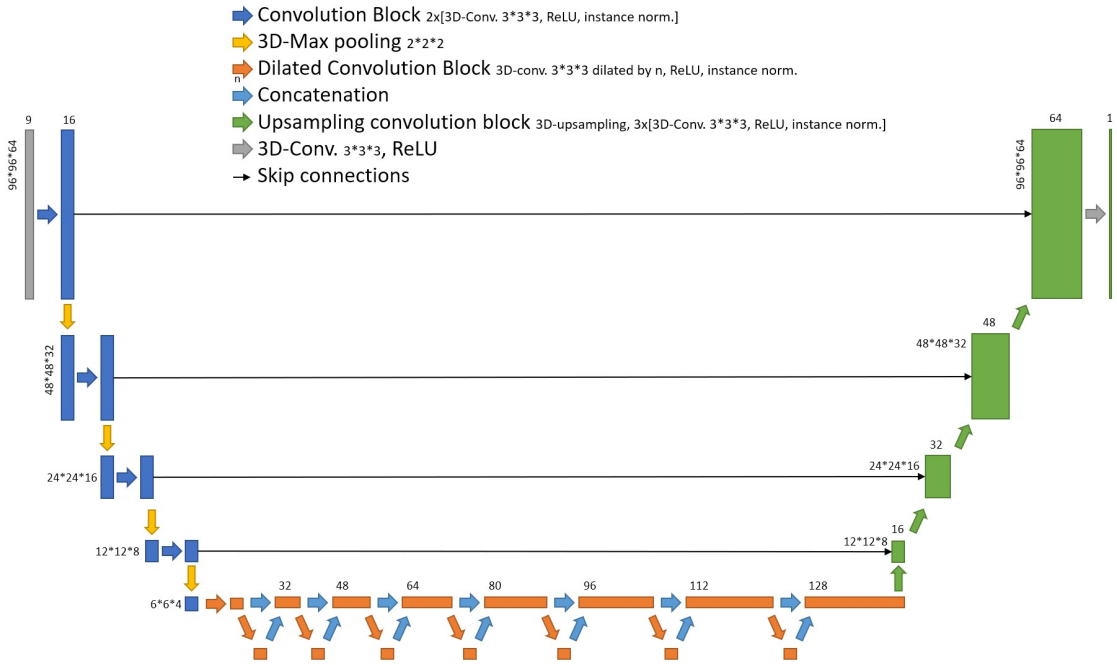


Figure 4.2: Detailed architecture of the U-Net used in this study.

layers with cubic pools of size 2 reduce its resolution to output a $6 * 6 * 4 * 16$ $(length * width * depth * channels)$ tensor.

The convolution blocks are made of 4 layers: 2 3D-convolution layers, each followed by an instance normalization layer. 3D convolution layers used in the encoder have a cubic kernel of size 3 and output 16 filters that pass through a "ReLU" activation.

The bottleneck is composed of 8 dilated convolution blocks separated by concatenation layers. It inputs the $6 * 6 * 4 * 16$ tensor given by the encoder and produces new channels to reach an output shape of $6 * 6 * 4 * 128$.

The dilated convolution blocks output a $6*6*4*16$ tensor each. The 8 successive blocks had dilation coefficients of 1, 2, 5, 9, 1, 2, 5, 9, respectively. For the rest, the convolution blocks of the bottleneck have the same parameters as those of the encoder. The first block receives the output of the encoder whereas the others receive a concatenation of the outputs from the previous blocks of the bottleneck. The bottleneck outputs the concatenation of the results of all its convolution blocks.

Finally, the decoder is fed with the enriched information from the bottleneck and produces the predicted dose map with shape $96 * 96 * 64 * 1$. It is made of 4 successive upsampling convolution blocks plus a 3D-convolution block as last layer.

Upsampling convolution blocks are composed of an upsampling layer and 3 3D-convolution layers each followed by an instance normalization layer. The

upsampling layer simply doubles the spatial resolution in every dimension by copy of the values of the voxels. These 3D convolution layers differ from those of the encoder by the number of channel they work with. They output $16 * n$ channels, where n is the index, in the decoder, of the block they belong to: 16 in the first block, 32 in the second, etc.

The final 3D-convolution block also differs only from those of the encoder by the sole channel it outputs, which is the dose map.

4.3 Patching system

Patching is the action of cutting the images of a patient into pieces of smaller size, called images patches. It is useful for various reasons. First, it enables to reduce the size of the input and output data, therefore the complexity of the network and to fit the image patches of all patients into the GPU memory. Second, it enables to predict different sizes of input and output because when patches have been passed to the network, the outputs can be reassembled to form the complete dose map. Patching is thus two sided, it can cut the inputs into patches of (96 x 96 x 64) voxels to fit the network and reassemble the outputs to predict the whole dose.

During the training of the model, patching was done according to a particular sample probability, before the presentation of each patient. The sample probability is computed by counting the number of voxels in the target volume for every patches. Image patches that contain a part of the CTV are thus more likely to be taken. Therefore, the network would rarely see the exact same inputs because it was shown different regions of the same set of patients each time, increasing its robustness to position variance. During the tests, patching was done with a stride of 32 voxels for each dimension, that is, the patching system fed the network with successive image patches that were shifted by 32 voxels each time to cover the whole original image. The complete map was then reassembled and in the regions where different prediction overlapped, their mean was taken.

4.4 Training

Regarding training, the first step was to teach a model to predict dose maps of the radiotherapy modality. The traditional method was used, it consists of the presentation of input-output pairs to the network so as to let an optimiser modify the weights of the neurons each time to reduce the errors made by the model. The Adam optimiser from Keras/Tensorflow was used in this work. The errors were measured by the mean squared error (MSE) function, which computes the mean of the square of the difference between the ground truth and the predicted dose on

the whole volume.

From the 200 patients, 144 were part of the training set whereas 36 and 20 were in the validation and tests sets. The training set is the one presented during learning to the network. The validation set is used after each presentation of the whole training set, to validate that the model learned useful knowledge about the task and did not just overfit the training data. Overfitting is the action, for a model, of learning to fit too closely to a particular dataset, so that it fails to reliably generalise its learning with new data, because it has extracted information about residual variation like measure noise. The goal of the training is actually to perform as well as possible on the validation set. Finally, the role of the test set is to assess the performance of the network once the learning is done but it was not used for the model for radiotherapy in this study because it is not the model of interest.

Starting from a Glorot uniform distribution initialisation presented in [14], the source model was trained for 150 epochs, where an epoch is the presentation of a random patch of every patient in the training set.

Then, no matter the transfer learning strategies adopted, the target models, working on the protontherapy dose predictions, were validated using 5-fold cross validation. After taking 10 of the 60 patients for the test set, the remaining patients were split in 5 groups of 10. For each of the five folds, a different group was chosen as validation set and the four other groups were the training set. A variation was also employed were only one of the four groups, a different one for each fold, was used in order to train the model with only 10 patients. The protontherapy models were always trained with 150 epochs like the radiotherapy model.

Two first target models were trained with 40 and 10 patients without any transfer learning to have a baseline against which to compare the different transfer strategies presented in the next section.

4.5 Transfer learning from radiotherapy to protontherapy

This section details both strategies of transfer learning explored. The first strategy involved a partial transfer of the knowledge whereas the second involves a complete one. In both cases, the transfer took place after the 150 epochs of training of the source model and before the beginning of the training of target models.

4.5.1 Partial transfer

As explained in the section 3.3, the first layers of a network tend to be general whereas the last layers are more specific to the problem. Based on that principle,

this first strategy consist in the transfer of the first layers. The layers that were transferred were also frozen, which means that their trainable weights were fixed. Various number of layers, ranging from only the first one to all of them were tested to assess the optimal fraction of the model to transfer. With this method, a model where all layers were transferred can no longer learn because all its parameters are frozen. The layers that were not transferred were randomly initialised. The figure 4.3.b. shows a schematic of the transfer strategy.

4.5.2 Fine tuning

The second strategy also takes advantage of the "general to specific" trend with depth. This time, all layers were transferred from the radiotherapy model to the protontherapy model but only the shallow ones were frozen. The last layers, unlike the partial transfer approach, were not randomly initialized, but the weights from the source model were used as initialisation. This strategy is also called warm-start optimisation. The last layers would thus be later fine tuned during the optimisation. Similarly to the first strategy, various number of frozen layers were tested to assess how many need to be retrained for the new prediction task. A representation of the strategy is shown in the figure 4.3.c.

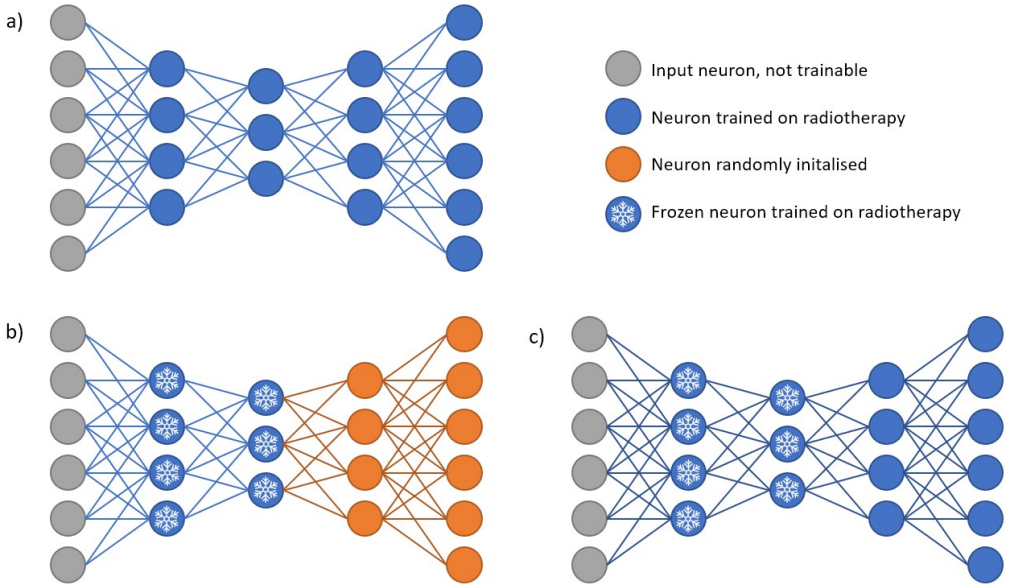


Figure 4.3: Schematic of the transfer learning methods with the state of the networks just after the transfer. a) Model trained on the radiotherapy dataset. b) First transfer strategy: The first layers of the radiotherapy model are transferred to the protontherapy model and frozen. c) Second transfer strategy: The complete model is transferred and the first layers are frozen.

4.6 Evaluation

For each variation of the transfer learning strategies, and each fold, the weights that led to the best validation error were selected to further assess the target models. Since the weights were chosen with respect to the validation data, their evaluation was performed on another dataset, the test set, that is composed of 10 cases not seen during training nor validation in order to avoid bias. Many variation were produced, thus only some of them, with good learning curves, were evaluated.

The first step was therefore to obtain readable learning curves. The separate mean of both the training and validation curves was taken across the five folds of the cross-validation procedure. Then a moving average with 10 successive epochs was computed to get smoother curves. The resulting curves were plotted together to highlight their differences.

Mean errors on some regions of interest (OARs and TVs) were computed for every selected variations, folds, and test patients. Beforehand, the predicted doses were linearly scaled to have the D50 on the CTV70 equal to 70Gy. Box plots of the errors for the various strategies were drawn and compared.

Using the same scaling, the DVH curves of the prediction on the test patients were plotted alongside the ground truth doses to have a visual appreciation of the quality of the prediction for every region of interest and dose levels. Box plots of D2 metrics of some important serial OARs like the spinal cord and brain stem, Dmean of some parallel OARs, the parotid glands and oral cavity, and D95 on the CTVs were placed side by side to compare the strategies.

Finally, Dice similarity coefficients (DSC) of the prediction from the various methods were computed and compared. Precisely, taking two ensembles A and B, the Dice coefficient measures the ratio between the number of element they have in common and those they do not share. It gives this equation:

$$Dice(A, B) = \frac{2 * \#(A \cap B)}{\#A + \#B}$$

Where the factor 2 accounts for the fact that element are present in both A and B and thus are counted 2 times in the denominator. Therefore a DSC of 0 means that no elements are shared whereas a coefficient of 1 means that the ensembles share all their elements and a Dice of 0.5 means that half the elements are shared.

Dice similarity coefficients do not work with continuous values but ensembles, thus the dose maps were split in isodoses masks, each mask covering the voxels that that surpass a given percentage of the prescription, for example a mask would cover every voxel which dose is higher than 56% of the prescription. for each prediction, 110 masks were computed that covered percentages of the dose ranging from 0 to 110, and these masks were compared through the Dice coefficient with those of the

ground truth dose. The result is a curve of similarity between the prediction and the truth as a function of the percentage of the prescription.

Chapter 5

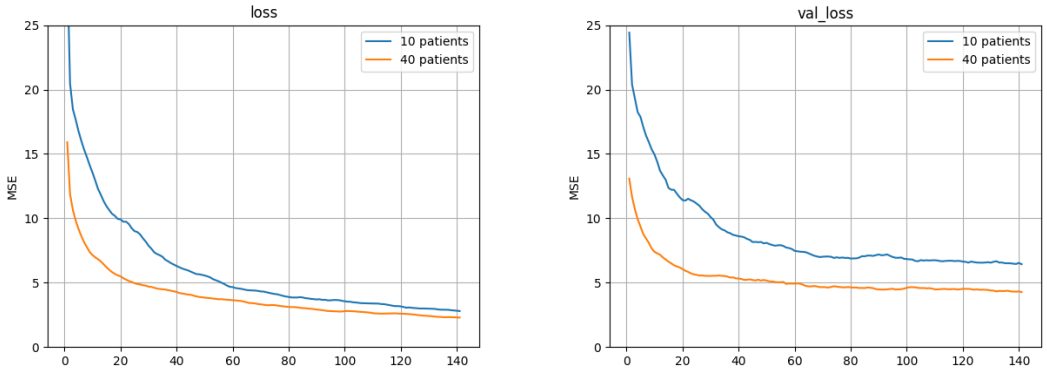
Results

This chapter presents the results obtained after following the methods that were presented earlier. Section 5.1 exposes the validation of the models and the selection of the variations. Section 5.2 assesses the models regarding their performances with clinical metrics on the target volumes and organs at risk. Finally, section 5.3 compares the predicted and the ground truth dose maps through Dice similarity coefficients and dose volume histograms.

5.1 Model validation for different transfer strategies

The first results relate to the models that were trained on the protontherapy dataset from scratch. As depicted in the Figure 5.1, two baseline models were trained with either 10 or 40 patients, they are named *baseline 10* and *baseline 40*, respectively. The training loss of *baseline 40* has a steeper decrease but is eventually caught up by that of *baseline 10* at the end of the training. On the other hand, *baseline 40* keeps a better lower validation, even when both models have stabilised, as visible in Table 5.1

The learning curves enable to appreciate the impact of the transfer strategies on the performances of the models during the learning. Figure 5.2 shows the influence of the partial transfer variations on a model trained with 40 patients, called *partial 40* regarding the training and validation accuracy alongside the curves of *baseline 40*. The models with 26, 27, and 28 layers frozen presented a clear worst validation loss (5.99 24.52, and 35.25, respectively versus 2.29 for the baseline) at the end of the training. In contrast, all the other models had losses similar to the baseline model after the training (+0.05 for the training and +0.25 for the validation) but presented training and validation losses reduced of around 0.77 and 0.36 after 20 epochs, with only slight differences between them.



(a) Training accuracy

(b) Validation accuracy

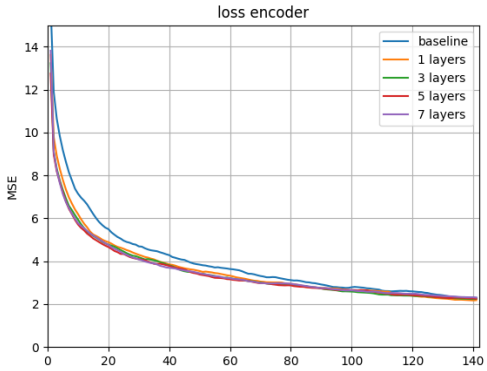
Figure 5.1: Training and validation learning curves of the baseline models (without transfer learning) and trained with 10 or 40 protontherapy patients.

To highlight differences between the variations around the partial transfer strategy, it was applied on models to be trained with only 10 patients. The learning curves of these models, called *partial 10* are shown in Figure 5.3. The models with the last 3 layers not frozen showed a mean reduction of 0.32 from 2.79 for the training loss and of 0.29 from 6.43 for the validation. Like with *partial 40*, the models involving a freezing of the last 3 layers presented clearly decreased performance in comparison with the baseline, with a validation loss increased by 1.48 with 26 frozen layers, 20.97 with 27 and 28.52 with all 28 layers frozen.

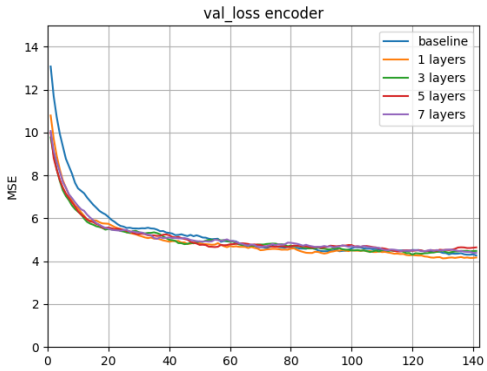
Variations of the second strategy of transfer learning, fine tuning, which involves a transfer of every layers and a freezing of some, were evaluated in the same manner on models trained with 10 patients. The learning curves of the variations of *fine tuning 10* are displayed in Figure 5.4. Once again, freezing the last 3 layers deteriorates the accuracy in comparison with the baseline. Specifically, the baseline model had a validation accuracy of 6.43 whereas models with 26, 27, and 28 frozen layers increased it by 1.51, 19.26, and 28.72 respectively. Also, the other models present losses reduction of 0.51 for the training and 0.57 for the validation, in average at the end of the training.

In total, more than 50 variations were developed: 17 variations with different numbers of layers transferred for both *partial 10* and *partial 40* plus 18 variations for the *fine tuning 10* and 1 for the *fine tuning 40*. It was thus decided to select only a few of them for further evaluation. The models considered are mostly those trained with 10 patients in this thesis and if the number of patients involved in the training is omitted when referring to a model, it is implied that 10 were used.

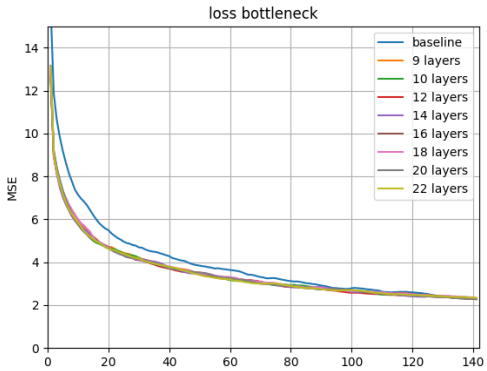
To select the models, Figure 5.5 shows the training and validation of the best



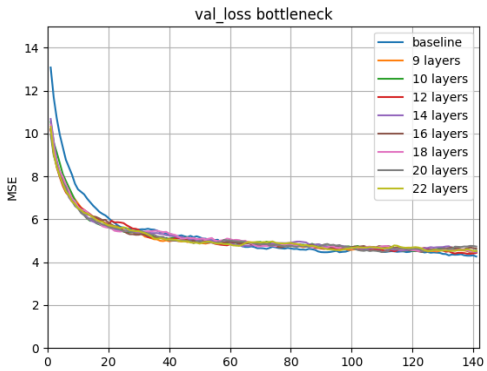
(a) Training, TL in the encoder



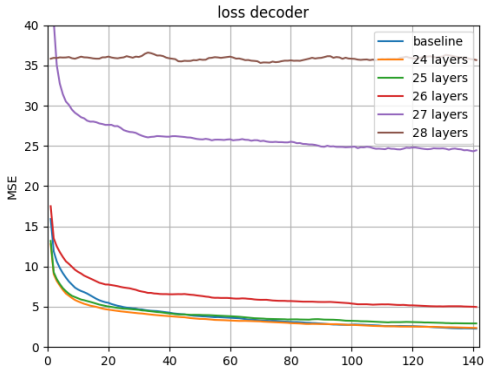
(b) Validation, TL in the encoder



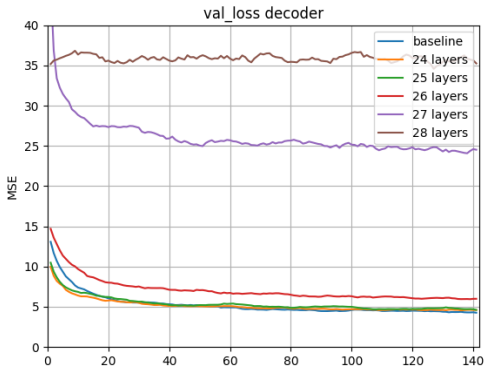
(c) Training, TL up to the bottleneck



(d) Validation, TL up to the bottleneck

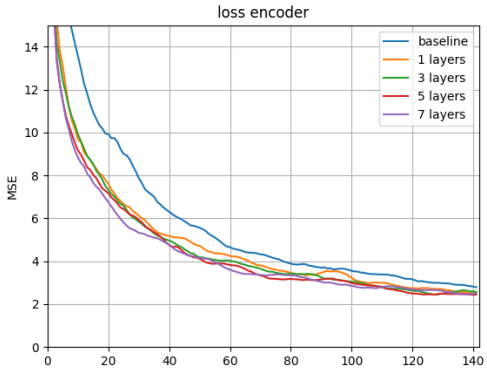


(e) Training, TL up to the decoder

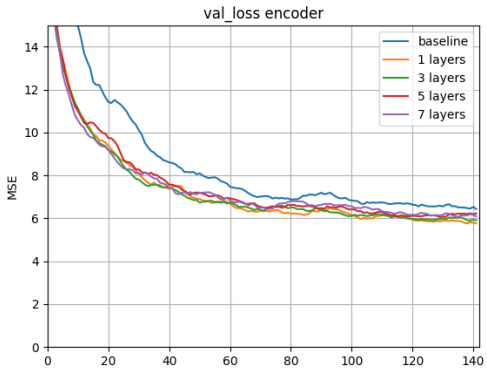


(f) Validation, TL up to the decoder

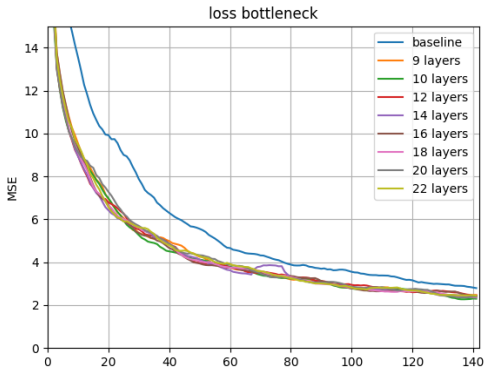
Figure 5.2: Training and validation MSE learning curves of the baseline model and variations around the *partial* transfer strategy, trained with 40 PT patients. Legend: XX layers is the number of layers transferred and frozen, starting from the input layer.



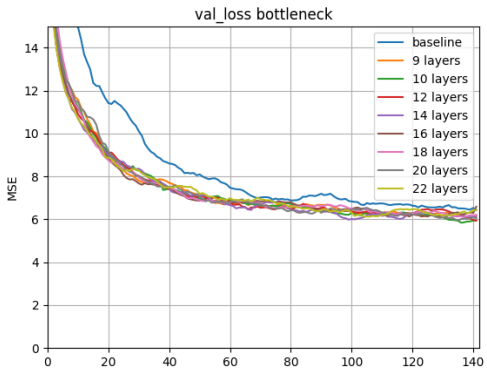
(a) Training, TL in the encoder



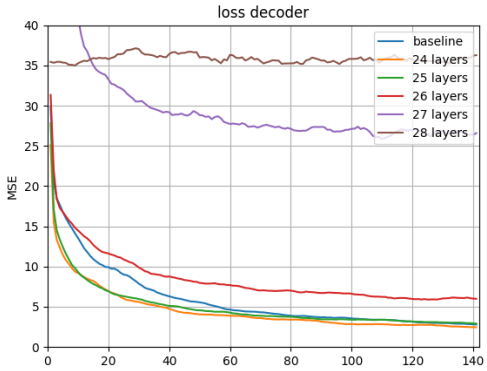
(b) Validation, TL in the encoder



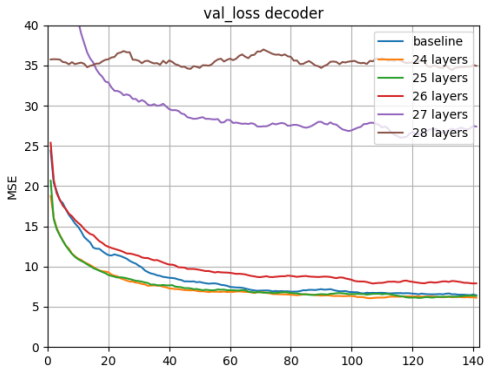
(c) Training, TL up to the bottleneck



(d) Validation, TL up to the bottleneck

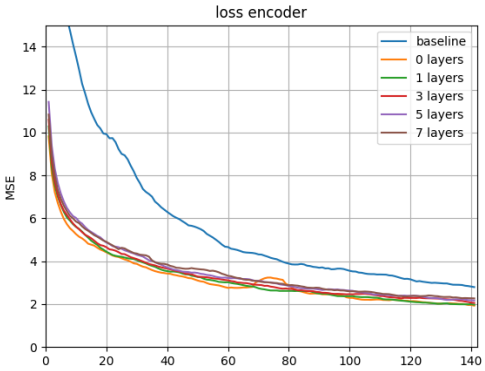


(e) Training, TL up to the decoder

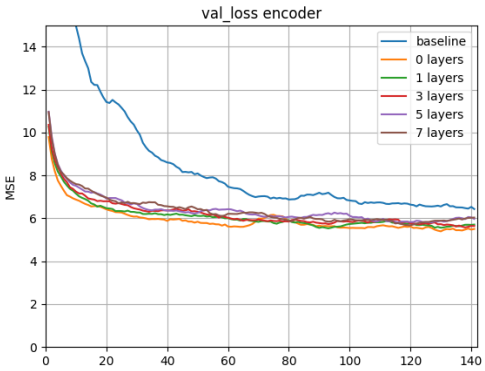


(f) Validation, TL up to the decoder

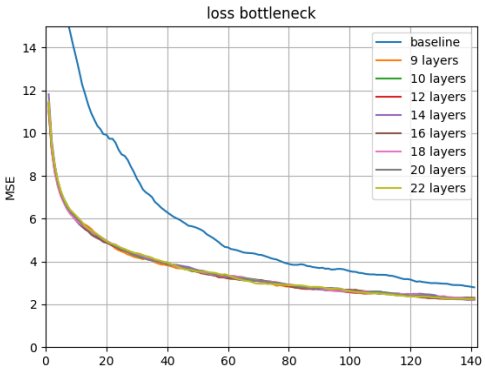
Figure 5.3: Training and validation MSE learning curves of the baseline model and variations around the *partial* transfer strategy, trained with 10 PT patients. Legend: XX layers is the number of layers transferred and frozen, starting from the input layer.



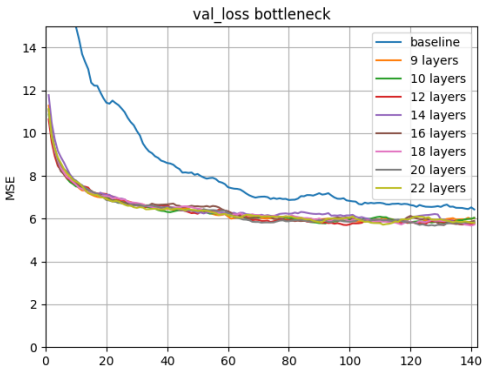
(a) Training, TL in the encoder



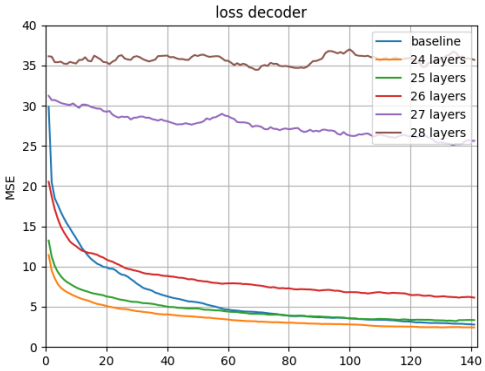
(b) Validation, TL in the encoder



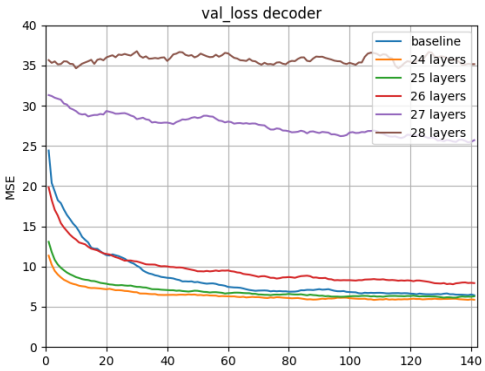
(c) Training, TL up to the bottleneck



(d) Validation, TL up to the bottleneck

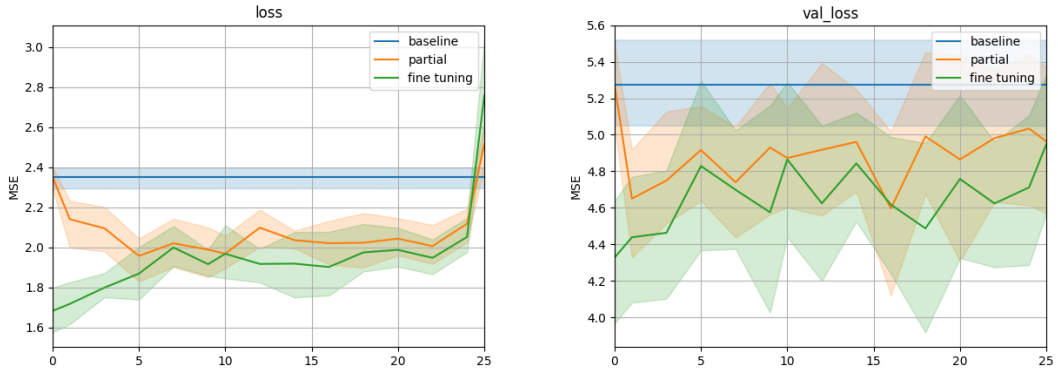


(e) Training, TL up to the decoder



(f) Validation, TL up to the decoder

Figure 5.4: Training and validation MSE learning curves of the baseline model and variations around the *fine tuning* transfer strategy, trained with 10 PT patients. Legend: XX layers is the number of frozen layers, starting from the input layer.



(a) Training accuracy

(b) Validation accuracy

Figure 5.5: Mean of the training and validation loss function MSE across folds, taken on the models that performed the best on the validation set during training with 10 patients, for the partial transfer and fine tuning methods, as a function of the number of layers transferred and/or frozen. The mean MSE of the baseline models are also displayed.

Table 5.1: Mean across folds of the validation loss function MSE of the selected models for both strategies with optimal parameters and trained with 10 or 40 patients as well as the baseline models.

Transfer method	10 patients $\mu \pm \sigma$	40 patients $\mu \pm \sigma$
Baseline	5.274 ± 0.265	3.456 ± 0.202
Partial transfer "14 layers"	4.961 ± 0.335	3.563 ± 0.236
Fine tuning "0 layer"	4.327 ± 0.384	3.044 ± 0.149

models reached during each fold of the training as a function of the number of layers involved in the variations. The models considered as the bests of their respective folds are those that got the lowest validation MSE during the learning.

The *fine tuning* strategy showed a clear highest accuracy on the training and validation sets with no freezing at all (that is, the model with "0 layers" legend). On the other hand, the results of the *partial* strategy were more mitigated. Looking at the validation accuracy in Figure 5.5a, 16 layers seems optimal but after a quick exploration of other evaluations metrics, the variation with 14 layers transferred from the radiotherapy model and frozen looks better and thus was selected. Table 5.1 indicates the validation accuracy of the models selected for further evaluation.

5.2 Clinical evaluation of selected best models for different transfer strategies

This sections gathers measurements made regarding clinical values of interest. They are grouped according to the type of volume they were taken on, beginning with Target Volumes and ending with Organs At Risks.

The following figures present box plots computed on the data from all folds and test patients. Boxes contain the data between the lower and upper quartiles, and the medians are represented as the horizontal lines in them. Also, whiskers encompass the data points that are not in the interquartile, excluding the outliers. The latter are the data points that are more than 1.5 interquartile range away from the upper and lower quartiles. The points plotted outside the whiskers are thus the outliers.

5.2.1 Target Volume

Figure 5.6 shows the errors between the D95 taken on both CTVs present in the plans of the protontherapy dataset. The *fine tuning* strategy (without freezing) improved the precision of the D95 metric prediction by 60% on both CTVs in comparison with *baseline 10*. It is even more robust than *baseline 40* with improvements of 40% on CTV54.25 and 31% on CTV70. The *partial* model (with 14 transferred layers) increased the precision on CTV70 by 18% in comparison with *baseline 10*.

5.2.2 Organs at risk

OARs were separated in serial and parallel organs to measure the error on their D2 or Dmean metrics made by the models. As important serial organs, the D2 of the spinal cord and brain stem were measured and displayed in Figure 5.7. The *partial* strategy presented a 15% decrease on the range of error on the D2 of the brain stem whereas *fine tuning* presented a 35% decrease on that error, reaching scores similar to that of *baseline 40*. However, none showed a clear improvement on the spinal cord.

On the parallel organs side, the Dmean of the parotid glands and the oral cavity is shown in Figure 5.8. Here, *partial* showed no benefit and even presented an increase of the error range on the right parotid by 41%. The *fine tuning* indicates a reduction of the error range on the left parotid gland by 15% with respect to both *baseline* models. Also, a 15% reduction is observed on the oral cavity in comparison with *baseline 10* but the resulting metrics still cover a range 15% larger than the *baseline 40* model.

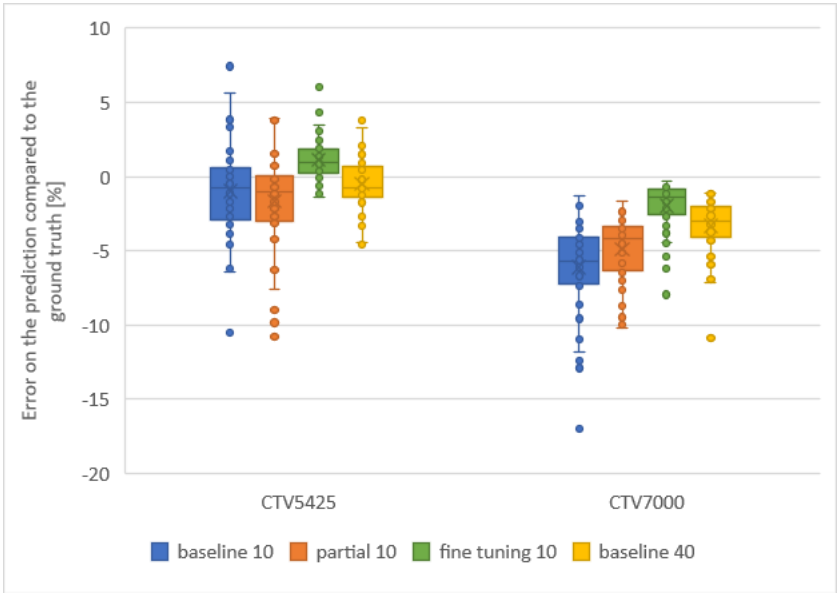


Figure 5.6: Box plots for the error (prediction - truth) in D95 for CTV 54.25Gy and CTV 70 Gy for the baseline, partially transferred, and fine tuned models. The error is presented as a percentage of the maximum prescription (70Gy).

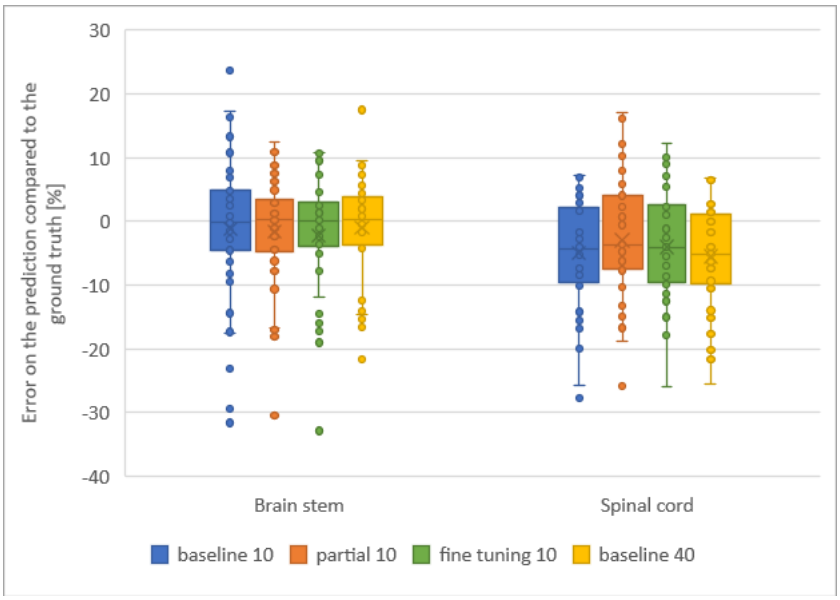


Figure 5.7: Box plots for the error (prediction - truth) in D2 for the brain stem and the spinal cord for the baseline, partially transferred, and fine tuned models. The error is presented as a percentage of the maximum prescription (70Gy).

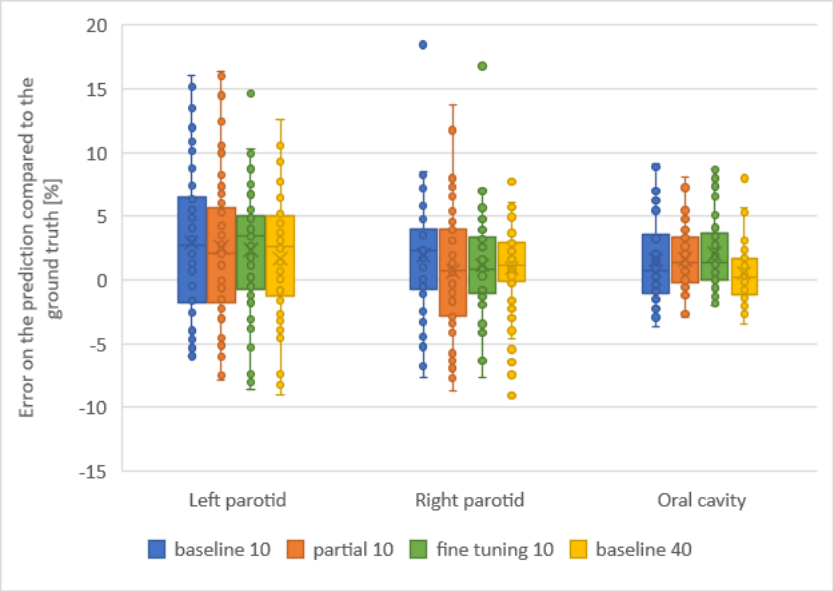


Figure 5.8: Box plots for the error (prediction - truth) in Dmean for the parotid glands and the oral cavity for the baseline, partially transferred, and fine tuned models. The error is presented as a percentage of the maximum prescription (70Gy).

5.2.3 Dice similarity coefficients

Dice similarity coefficients were measured for every folds of every models on every test patients. Figure 5.9 shows the DSC distributions of the models trained with 10 patients to give an appreciation of their precision at different dose levels. *Partial* presented and increase of 1.1% and *fine tuning* of 2.5% on the DSC scale on average, in comparison with the *baseline*.

Another comparison of Dice coefficients is shown in Figure 5.10, it shows a clear reduction of the difference between the DSC of the models trained with either 10 or 40 patients when transfer learning is used. Whereas *baseline* presented a difference of 3.08% on average, *partial* and *fine tuning* presented a difference of only 1.77% and 2.42%, respectively.

5.2.4 DVH curves and dose distribution

Figure 5.11 offers a visual appreciation of the DVH reached by each model on an typical test patient.

Finally, Figure 5.12 gives a comparison of a ground truth dose for a patient and the doses predicted by *baseline 10*, *partial 10* and *fine tuning 10*.

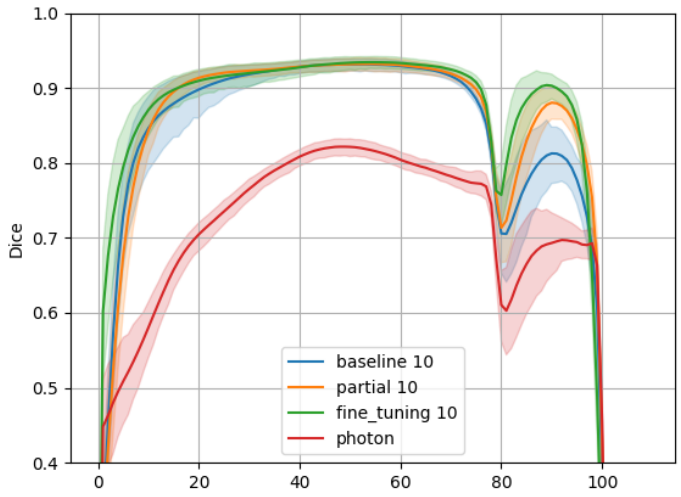
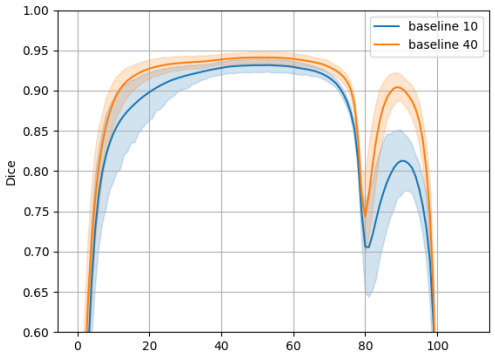
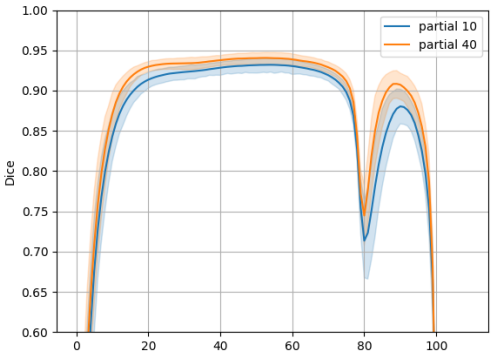


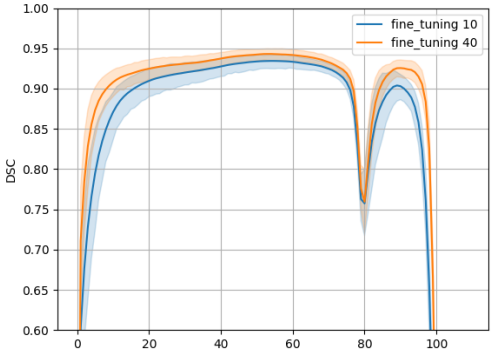
Figure 5.9: Dice similarity coefficients between the ground truth doses and the predictions of the baseline, partially transferred, and fine tuned models trained with 10 patients as well as the model trained on the photon dataset, as a function of the prescription percentage. The lines represent the means and the clouds the 95% confidence intervals.



(a) Without transfer

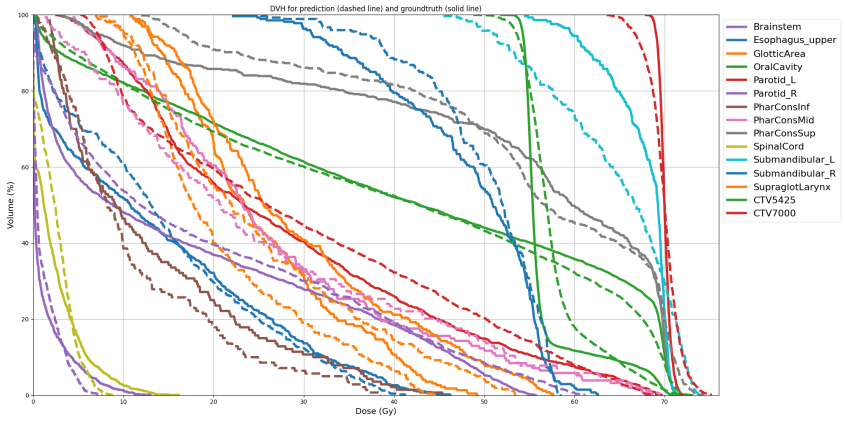


(b) Partial transfer

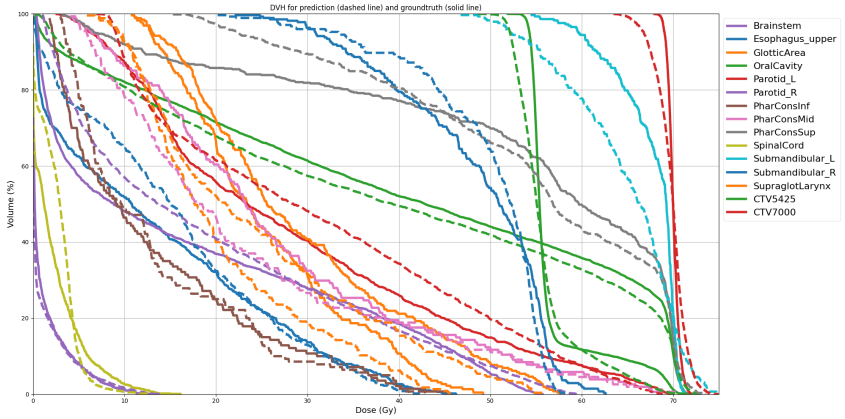


(c) Fine tuning

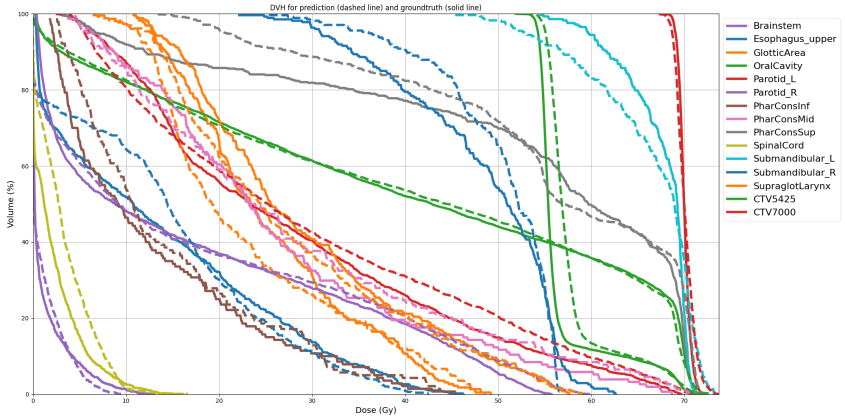
Figure 5.10: Comparison of the Dice coefficients of the baselines partially transferred, and fine tuned models trained on 10 or 40 patients, as a function of the prescription percentage. The lines represent the means and the clouds the 95% confidence intervals.



(a) Baseline.

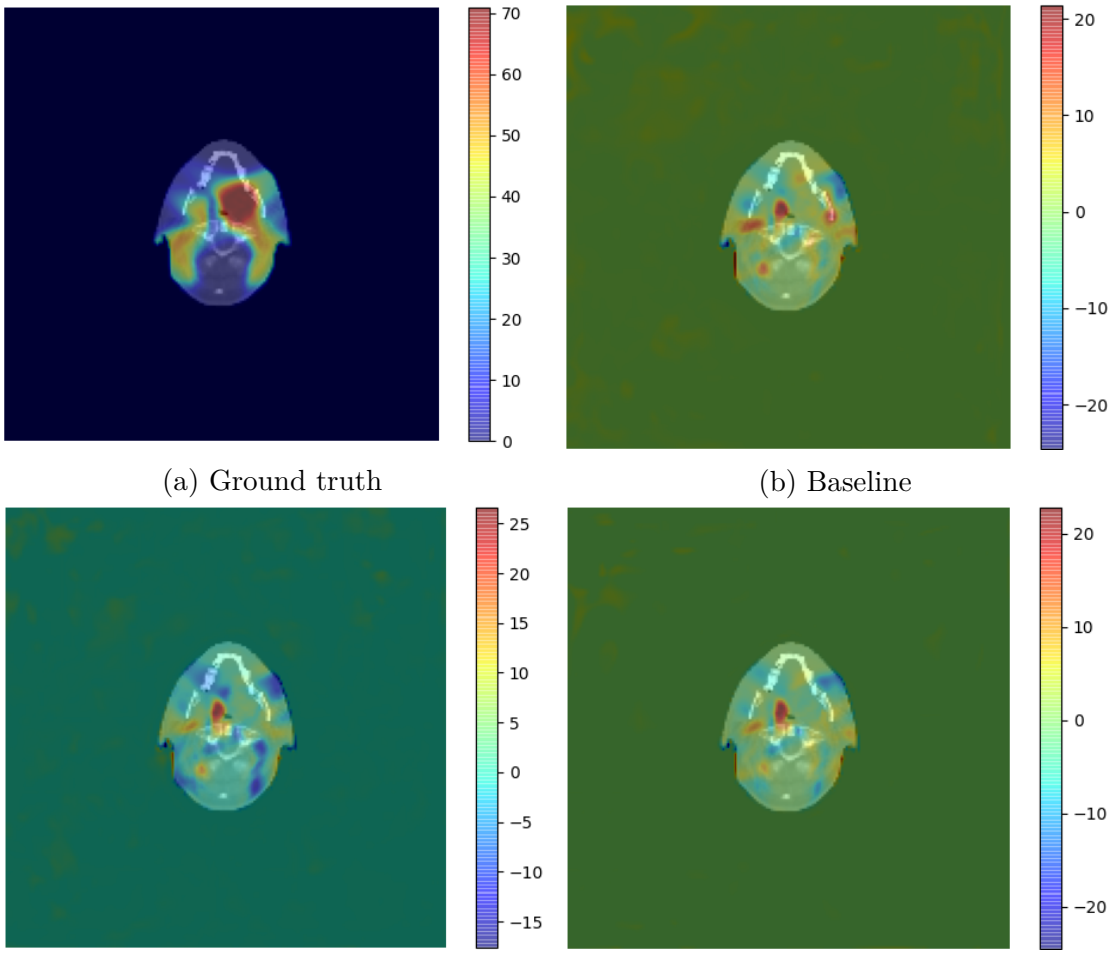


(b) Partial transfer.



(c) Fine tuning.

Figure 5.11: Comparison between the DVH of the predictions from the 3 models (dashed lines) trained with 10 patients and the ground truth DVH (solid lines) on a typical example of the test set.



(a) Ground truth

(b) Baseline

(c) Partial

(d) Fine tuning

Figure 5.12: Comparison between a slice of the ground truth dose distribution on a patient and the same slice of the dose maps predicted by *baseline 10*, *partial 10* and *fine tuning 10*.

Chapter 6

Discussion

The previous chapter reported the results of the experiments conducted during this study and this chapter is dedicated to their discussion. At first, the results are interpreted. Then, some limitations of this study are exposed. Finally, ideas of futures perspectives are expressed.

6.1 Results interpretation

6.1.1 Model validation

The learning of *baseline 10* required about 55 epochs for the training loss MSE to fall below 5 and that of *baseline 40* only required about 25. This difference of learning speed, visible in Figure 5.1a, is explained by the number of examples presented to the models. *Baseline 40* saw 40 patients per epoch, thus seeing roughly 1000 examples to reach a training MSE of 5 whereas *baseline 10* needed only around 550 examples. Under this view point, *baseline 10* learns faster, which is normal since it needs to imitate a smaller variety of cases. After 60 epochs, *baseline 10* slowly catches up *baseline 40*.

However, *baseline 40* keeps a better validation accuracy than *baseline 10* throughout the procedure, as visible in Figure 5.1b. Both models seems to have reached their optimum and *baseline 40* gets a better validation score, reflecting its better ability to generalise its knowledge of the task. This was also expected due to the wider variety of cases used to train it in comparison with *baseline 10*.

Next, Figures 5.2, 5.3 and 5.4 show the impact of the transfer strategies on the learning procedure. While the results with *partial 40* are slight, they already show better training and validation learning curves, especially for the first epochs. Regarding the decoder, it is clear that at least the last 3 layers need to be free to adapt to the new task in order not to deteriorate the contribution of the transfer,

no matter the strategy employed nor the size of the training set. Although most of the variations brought benefits in terms of learning, no readable difference was observed between the learning curves of the variations around both strategies.

Figure 5.5 confirms the similarity between most variations, especially for the *partial* strategy, bringing an average reduction of about 0.3 and 0.4 on the training and validation MSE, respectively. For both metrics, its variations had alike performances except for the extreme values, that is when nearly no weights were transferred or when most layers were frozen. Also, whereas the *partial* strategy enables a clear reduction of the errors during training, the difference in validation accuracy at the end of the process is less clear as its 95% confidence interval often crosses that of the *baseline* model.

On the other side, the *fine tuning* still showed well improved validation score at the end of the training, as confirmed by Table 5.1. The *fine tuning* strategy tends to show lower error and variability on both training and validation sets when less layers were frozen, maximising its performance with no freezing at all.

As visible in Table 5.1, good transfer learning variations had a good impact when only 10 training patients were used. *Partial* with 14 frozen layers reduced the average validation MSE by 0.313 and *fine tuning* without freezing, reduced the same metric by 0.947. With 40 patients, the *partial* transfer strategy of 14 layers was not able to improve the validation accuracy whereas *fine tuning* reduced it by 0.412.

The transfer and freezing of the first half of the network reduced the mean validation MSE by 5.93% and the transfer of all layers without freezing achieved a reduction of 17.96% in comparison to the baseline when the models were trained with 10 protontherapy patients. With 40 patients, only the fine tuning method showed an improvement of 11.92% on the validation accuracy.

6.1.2 Clinical evaluation

Regarding the DVH metrics, *partial* had a clear positive impact on the prediction of the D95 of the CTV70 (-18% on the errors) as well as on the brain stem (-15%). On the spinal cord, known as an important and on which the errors were relatively large, it only reduced the mean error of the D2 metric, therefore centering the distribution better around 0 without reducing the range of errors. Its impact on the parallel organs is poorer, with an increase of 41% of the error on a parotid and results similar to those of *baseline 10* on the other parotid and the oral cavity.

Partial improved the DSC by 1.1% in average when 10 training patients were used, but the increase in especially remarkable for the higher doses, where the baseline model struggled the most. The DSC indeed stepped up by 4.8% for doses ranging from 75% to 100% of the maximum prescription.

Looking at Figure 5.10, one can see that the *partial* transfer learning method

is able to compensate a lack of protontherapy training patients. Whereas there was a difference of 3.08% in DSC between the models trained with either 10 or 40 patients without transfer learning, it drops to 1.77% when it is employed. The *partial* transfer learning method thus reduces the negative impact of a lack of patient and the *fine tuning* method would probably bring similar results.

The *fine tuning* strategy often worked better than the *partial* strategy. It reduced the error on the D95 of CTVs by 60% and on the Dmean of the left parotid gland by 15%, therefore outperforming *baseline 40*. It also improved the precision on the D2 of the brain stem, this time reaching performance similar to, but not better, than *baseline 40*. Finally, it placed itself in between the *baseline* models trained with 10 and 40 patient regarding the oral cavity.

The DSC was improved in a similar way by *partial* and *fine tuning*, but here again, the former was outperformed by the latter with an average increase of 2.5% and 6.3% for the higher doses, which is 31% better than *partial*. *Fine tuning* also has an earlier and stiffer increase of the Dice similarity coefficient than every other model, showing that it brings better help to predict lower doses. Precisely, for doses up to 25% of the prescription, *partial* did not bring changes in average where *fine tuning* improved the DSC by 4.8%.

An interesting result is that the mean DSC of the prediction from *fine tuning 10* is worst than that of *baseline 40* by 0.54% while using 4 times less patients. Moreover, *fine tuning 40* significantly increased the DSC in comparison with *baseline 40*, reaching a score of 82.53% versus 80.65%.

Looking at the distribution of the DVH metrics in general, it can be observed that the left parotid has a larger error distribution than the right one with all four models. This probably highlight a lack of bilateral symmetry during the training.

Finally, Figure 5.11 raises an alarm regarding the spinal cord and brain stem. They are indeed the two organs with the lowest dose and yet there are errors up to 30% of the prescription on their D2 metrics, whether with the *baseline 10* model or using transfer learning techniques.

It is quite hard to evaluate the *partial* transfer method without a medical expertise. It improved the precision of the prediction on the series organs, that are the brain stem and the spinal cord, both believed to be of major importance and on which the error was the biggest in the *baseline* model, as well as on the oral cavity. The downside is that it clearly deteriorated the reliability on the parotid glands.
On the other side, the *fine tuning* transfer method improved every metric, and better than *partial* each time, except on the spinal cord where the deviation of the error on D2 was slightly increased in comparison with the *baseline* model.

The DSC shows an average increase of 0.011 and 0.025 with the first and second strategies, respectively. Moreover, the *fine tuning* method improved the DSC by 0.048 for the lower doses and by 0.063 for the higher ones. The transfer learning methods have shown to reduce the deleterious impacts of a lack of patients with *fine tuning 10* reaching performances of a *baseline* model trained with 40 patient

6.2 Limitations

The present study has some limitations that are detailed in this section.

First, the need for similar inputs for the transfer learning between radiotherapy and protontherapy led to the disuse of some channels originally available in the protontherapy plans dataset. The baseline models used only the 9 input channels selected but no comparison was done with models that use all the channels available. Having more information at hand for each prediction might indeed lead to better improvements than using transfer learning from another modality. For the sake of readability of the results, the choice was made to compare models that took the same inputs, but still, answering this question would have been interesting.

Second, the baseline models did not benefit from the same total training time. Part of the improvement brought by the transfer learning might be due to a longer training with the 150 epochs on the radiotherapy dataset followed by the 150 on the protontherapy dataset against only the last 150 epochs for the baseline models.

Third, the model was trained with similar treatments for every patient, which is probably a reason of the differences in precision between the parotid glands. In real life applications, a model should be able to select between different styles the one that is the best for a particular patient. Filling the dataset with more diverse treatment plans would give results closer to those achievable in a clinical context.

Fourth and lastly, the tensors used had voxels with 3.5mm sides. This resolution is lower than the clinical standards but predicting with higher precision should still be possible at the cost of longer training and prediction times.

6.3 Perspectives

At the light of the results of this study, a few perspectives of investigations are proposed in this section.

As stated in the literature review, some state-of-the-art methods for dose prediction in protontherapy use a special type of U-Net called HD-UNet, they are simpler in terms of trainable parameters but still yield better predictions than the regular U-Net used in this study. If it happens that this architecture also benefits from transfer learning, it could bring improvements to the performances of

the currently best models, especially when low amount of good quality plans are available.

Finally, one way to palliate the loss of information due to the channel selection, could be to match contours of organs that are different but located in similar areas or for which the clinical objectives might be alike. It could also be tried to take all the channels available in the proton therapy dataset but, if such channel is not available in the radiotherapy dataset, to put their corresponding inputs to either 0 or random values during the preliminary training. That way, it would still be possible to use them as input for the protontherapy dose prediction model. Particular attention should be paid to the freezing with this strategy because the network trained on the radiotherapy should have learnt disregarding the empty channels and thus would need to be able to unlearn that, which could be harder if the first layers are frozen.

Chapter 7

Conclusion

A reliable dose prediction is essential to automate the treatment planning in head and neck cancer. The goal of this thesis was to improve dose prediction of a U-Net through transfer learning from radiotherapy, where data is abundant, to protontherapy, where it is rare due to the youth of the modality. Transfer learning was able to help models to produce accurate, fast and achievable doses through dose mimicking.

Two transfer methods were tested and it appears that they both improved the model accuracy, especially when confronted to a lack of training examples. The fine tuning strategy, which consists in transferring all the weights without freezing, has shown to be a specifically good way to improve the performances of a model trained with 10 protontherapy patients.

It increased the mean validation accuracy by nearly 18%. It also reduced the uncertainties on most the clinical metrics, with either a shortening of the error distribution, up to 60% lower in the CTV. The only volume that did not benefit from the transfer is the spinal cord were the range of errors increased by 15%. Finally, it increased the average DSC of the prediction by 2.5% with greater improvements for the dose values were the model struggled the most. The DSC was increased by 4.8% for the lower doses and by 6.3% for the higher ones.

As a take-home message, using transfer learning with a fine tuning method from radiotherapy to protontherapy enabled a model trained with only 10 protontherapy patients to reach performances similar to a model trained only on protontherapy with 40 patients. This is valuable to speed up the deployment of automatic planning in clinical practice where building a dataset with 40 patient takes around 1 month.

Future experiments could be led to explore the following perspectives:

- HD-UNets have shown the best performances on the dose prediction task yet.

It could be interesting to investigate how transfer learning would affect their performance.

- The contours missing in the radiotherapy dataset could be replaced with fake inputs or matched with remotely similar organs of the protontherapy dataset during the preliminary training, instead of removing them from the inputs of the protontherapy model. This would let the latter use all the contours on which the plans were prepared and could improve its performances even more.

Bibliography

[1] Avo. *The potential of protontherapy*. URL: https://www.avoplc.com/en-gb/Technology/The-Potential-of-Proton-Therapy.

[2] Ana Barragán-Montero et al. "Artificial intelligence and machine learning for medical imaging: A technology review". In: *Physica Medica* 83 (2021), pp. 242–256.

[3] Ana María Barragán-Montero et al. "Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations". In: *Medical physics* 46.8 (2019), pp. 3679–3691.

[4] Anders Bertelsen et al. "Single Arc Volumetric Modulated Arc Therapy of head and neck cancer". In: *Radiotherapy and Oncology* 95.2 (2010), pp. 142–148. ISSN: 0167-8140. DOI: https://doi.org/10.1016/j.radonc.2010.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0167814010000629.

[5] Alejandro Carabe-Fernandez et al. "Is there a role for arcing techniques in proton therapy?" In: *The British Journal of Radiology* 93.1107 (2020). PMID: 31860338, p. 20190469. DOI: 10.1259/bjr.20190469. eprint: https://doi.org/10.1259/bjr.20190469. URL: https://doi.org/10.1259/bjr.20190469.

[6] Pierre Castadot et al. "Adaptive radiotherapy of head and neck cancer". In: *Seminars in radiation oncology*. Vol. 20. 2. Elsevier. 2020, pp. 84–93.

[7] Seattle CancerCare Alliance Proton Therapy Center. *Prostate Cancer Treatment*. URL: https://www.sccaprotontherapy.com/cancers-treated/prostate-cancer-treatment.

[8] David S. Chang et al. "Interactions of Electromagnetic Radiation with Matter". In: *Basic Radiotherapy Physics and Biology*. Cham: Springer International Publishing, 2021, pp. 35–40. ISBN: 978-3-030-61899-5. DOI: 10.1007/978-3-030-61899-5_4. URL: https://doi.org/10.1007/978-3-030-61899-5_4.

[9] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26.6 (2013), pp. 1045–1057.

[10] Ximin Cui et al. “Multiscale Spatial-Spectral Convolutional Network with Image-Based Framework for Hyperspectral Imagery Classification”. In: *Remote Sensing* 11 (Sept. 2019), p. 2220. DOI: 10.3390/rs11192220.

[11] Albin Fredriksson, Anders Forsgren, and Björn Hårdemark. “Minimax optimization for handling range and setup uncertainties in proton therapy”. In: *Medical physics* 38.3 (2011), pp. 1672–1684.

[12] J. Lee G. Janssens E. Sterpin. *Engineering challenges in protontherapy, gbio2070 course*. 2017.

[13] Mohsen Ghafoorian et al. “Transfer learning for domain adaptation in mri: Application in brain lesion segmentation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 516–524.

[14] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, May 2010, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html.

[15] DT Goodhead. “Spatial and temporal distribution of energy”. In: *Health physics* 55.2 (Aug. 1988), pp. 231–240. ISSN: 0017-9078. DOI: 10.1097/00004032-198808000-00015. URL: https://doi.org/10.1097/00004032-198808000-00015.

[16] Wenbo Gu et al. “Integrated beam orientation and scanning-spot optimization in intensity-modulated proton therapy for brain and unilateral head and neck tumors”. In: *Medical Physics* 45.4 (2018), pp. 1338–1350. DOI: https://doi.org/10.1002/mp.12788. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12788. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12788.

[17] Andrea Holt et al. “Multi-institutional comparison of volumetric modulated arc therapy vs. intensity-modulated radiation therapy for head-and-neck cancer: a planning study”. In: *Radiation Oncology* 8.1 (Jan. 2013), p. 26. ISSN: 1748-717X. DOI: 10.1186/1748-717X-8-26. URL: https://doi.org/10.1186/1748-717X-8-26.

[18] Roya Norouzi Kandalan et al. “Dose prediction with deep learning for prostate cancer radiation therapy: Model adaptation to different treatment planning practices”. In: *Radiotherapy and Oncology* 153 (2020), pp. 228–235.

[19] Huet - - Dastarac Margerie. “Dose prediction for protontherapy using neural networks”. Prom. : Lee, John Aldo ; Sterpin, Edmond. 2019. URL: http://hdl.handle.net/2078.1/thesis:19554.

[20] Patricia van Marlen et al. “Ultra-High Dose Rate Transmission Beam Proton Therapy for Conventionally Fractionated Head and Neck Cancer: Treatment Planning and Dose Rate Distributions”. In: *Cancers* 13.8 (2021). ISSN: 2072-6694. DOI: 10.3390/cancers13081859. URL: https://www.mdpi.com/2072-6694/13/8/1859.

[21] Philip Mayles, Alan Nahum, and Jean-Claude Rosenwald. *Handbook of radiotherapy physics: theory and practice*. CRC Press, 2007.

[22] SE McGowan, NG Burnet, and AJ Lomax. “Treatment planning optimisation in proton therapy”. In: *The British journal of radiology* 86.1021 (2013), pp. 20120288–20120288.

[23] Wayne D Newhauser and Rui Zhang. “The physics of proton therapy”. In: *Physics in Medicine and Biology* 60.8 (Mar. 2015), R155–R209. DOI: 10.1088/0031-9155/60/8/r155. URL: https://doi.org/10.1088/0031-9155/60/8/r155.

[24] Dan Nguyen et al. “A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning”. In: *Scientific reports* 9.1 (2019), pp. 1–10.

[25] Dan Nguyen et al. “Three-dimensional radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. In: *arXiv preprint arXiv:1805.10397* (2018).

[26] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[27] American Association of Physicists in Medicine. *The Open Knowledge-Based Planning Challenge (OpenKBP)*. URL: https://www.aapm.org/GrandChallenge/OpenKBP/.

[28] Ponor. URL: https://commons.wikimedia.org/w/index.php?curid=91353714.

[29] Protom. *The Bragg Peak*. June 2018. URL: https://www.protominternational.com/2018/06/bragg-peak/.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[31] Satomi Shiraishi and Kevin L Moore. "Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy". In: *Medical physics* 43.1 (2016), pp. 378–387.

[32] Kevin Souris et al. "Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy". In: *Medical physics* 46.10 (2019), pp. 4676–4684.

[33] David I Thwaites and John B Tuohy. "Back to the future: the history and development of the clinical linear accelerator". In: *Physics in Medicine and Biology* 51.13 (June 2006), R343–R362. DOI: 10.1088/0031-9155/51/13/r20. URL: https://doi.org/10.1088/0031-9155/51/13/r20.

[34] Ovidiu Veresezan et al. "Adaptive radiation therapy in head and neck cancer for clinical practice: state of the art and practical challenges". In: *Japanese journal of radiology* 35.2 (2017), pp. 43–52.

[35] Frank Verhaegen et al. "Considerations for shoot-through FLASH proton therapy". In: *Physics in Medicine & Biology* 66.6 (Mar. 2021), 06NT01. DOI: 10.1088/1361-6560/abe55a. URL: https://doi.org/10.1088/1361-6560/abe55a.

[36] Ilma Xhaferllari et al. "Automated IMRT planning with regional optimization using planning scripts". In: *Journal of applied clinical medical physics* 14.1 (2013), pp. 176–191.

[37] Yidong Yang et al. "An overlap-volume-histogram based method for rectal dose prediction and automated treatment planning in the external beam prostate radiotherapy following hydrogel injection". In: *Medical physics* 40.1 (2013), p. 011709.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**