Norwegian University of Life Sciences

Master's Thesis 2019 30 ECTS

Faculty of Science and Technology

Cecilila Marie Futsæther

Deep learning for automatic tumor delineation of anal cancer based on MRI, PET and CT images.

Christine Kiran Kaushal

MSc. Environmental Physics and Renewable Energy

Faculty of Science and Technology

Acknowledgments

First of all I would like to thank my supervisor Prof. Cecilia Marie Futsæther for admirable guidance while writing this thesis. She has shown excellent support, been a motivational factor during the whole period and provided me with thoroughly feedbacks.

In addition, Yngve Mardal Moe gave me orientation about the work of his MSc and gave constructive recommendations throughout this semester. I appreciate the sharing of his knowledge with me. I would also like to thank PhD student Aurora Rosvoll Grøndahl and, certainly, Prof. Oliver Tomic for being available for discussions, the sharing of their ideas and guidance.

Without the availability of the ANCARAD dataset this thesis would not be possible, and therefore a big thanks to both Prof. Eirik Malinen and Marianne Grønlien Guren, M.D, PhD. is in order. In addition, PhD student Espen Rusten was of great aid while co-registering the medical images and has been available for additional questions this semester.

Hallgeir Maage also deserves a thank you for repairing the computer used for running my experiments and making sure that I could use the GPU available.

Finally, I would like to thank my closest friends and family who have given me the strength to complete my Master's thesis. Without your love and support I might not have come this far by now.

Christine Kiran Kaushal
Ås, 14th May 2019

Abstract

Purpose
Precise delineation of tumors is considered the weakest link and the largest source of uncertainty in radiotherapy planning. The purpose of this thesis is to explore some of the possibilities for automatic delineation of cancerous tumors in medical image data of anal cancer provided by Oslo University Hospital. The use of an autodelineation computer program could potentially save time, provide consistency and give the physicians the possibility to focus more on other challenges.

Method
The dataset consisted of MRI, PET and CT images from 85 patients with anal cancer, who were scheduled for radiotherapy or chemo-radiotherapy in the period 2013 to 2016. Three experienced radiation oncologists provided the dataset with target volume delineations of the primary tumor, which was considered the ground truth delineations. The dataset was split into a training, a validation and a test set, stratified based on the volume of the ground truth delineations.

The autodelineation of the primary tumors in the medical images was performed using a deep learning approach by sentiment image segmentation, with a U-Net architecture. Ten experiments based on different imaging modalities, and combinations of them, were conducted. To increase the training data, image augmentation was used when preprocessing the data. Furthermore, data cleaning was performed in order to exclude image slices with defects. Finally, the Dice performance of the experiments using different imaging modalities as input was compared and the effects of regularization and data cleaning were explored. The implemented framework along with the codes used for the preprocessing are available at: https://github.com/christinekaush/ANCARAD_autodel.

Results
Using PET and CT images together as input to the deep learning segmentation program seems the most promising for the purpose of autodelineation of cancerous tumors of anal cancer patients, with a Dice performance of 0.885 on the validation set. Furthermore, data cleaning and the removal of image slices with no delineation provided by an oncologist seemed to have the largest impact on the Dice performance of the model. In addition, the experiments using CT and T2W individually as input to the deep learning model also showed promising results with Dice coefficients of 0.877 and 0.861 respectively.

When inspecting the autodelineations on the validation and test set, the delineations made by the deep learning model matched the provided target volume well, resulting in high Dice performances per patient ($>$ 0.85). The model does not seemed to recognize image slices that did not contain any tumor tissue delineation made by an oncologist.

Conclusions
Deep learning autodelineation of primary tumor in medical images of anal cancer patient shows excellent potential, providing comparable performance to the overlap expected between oncologists. The tumors in this dataset are located in more or less the same region, which makes it easier for the model to learn how to find tissue that potentially are cancerous for anal cancer patients. Further exploration of autodelineation including more image slices representing regions without anal cancer tumors should be conducted.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	Anal cancer
ADC	Apparent diffusion coefficient
AI	Artificial Intelligence
ANCARAD	Anal Cancer Radiotherapy
CNN	Convolutional Neural Network
CT	Computed tomography
CTV	Clinical target volume
DICOM	Digital Imaging and Communications in Medicine
DPCT	Dose Planning Computed Tomography
DWI	Diffusion weighted images
FCN	Fully Connected Networks
FDG	Fluorodeoxyglucose
FN	False negative
FP	False positive
GTV	Gross tumor volume
HDF	Hierarchical data format
HU	Hounsfield unit
ICRU	International Commission on Radiation Units and Measurements

MICE	Medical Interactive Creative Environment
MRI	Magnetic resonance imaging
OUH	Oslo University Hospital
PET	Positron emission tomography
PTV	Planning target volume
REC	Regional Committees for Medical and Health Research Ethics
ReLU	Rectified Linear Unit
RF	Radio frequency
SSE	Sum of Squared Errors
SUV	Standardised uptake value
TE	Time of echo
TN	True negative
TP	True positive
TR	Repetition time
TV	Target volume

Chapter 1

Introduction

1.1 Motivation

1.1.1 Anal Cancer

Anal cancer is the development of cancerous tumors in the anus or in the rectal canal within 4-5 cm from the anal opening [1], [2]. This type of cancer is rare in Norway with about 40 to 50 incidents annually per 2008 [2]. In 2018 the occurrence increased to about 75 patients annually [1].Certain types of Human papilloma viruses (HPV) have been detected in the majority of the patients [1], [2].

Patients with anal cancer receive either radiotherapy, chemotherapy, both or surgery [1]–[3], but a combination of chemo- and radiotherapy has been shown to give the best tumor control [2]. Patients diagnosed with anal cancer have a high chance of survival. In 2008 the five-year survival was estimated to be between 80 and 95 % for two-thirds of the patients with tumors under 5 cm [1], [2]. However, many patients are left with discomfort post cancer treatment [1].

1.1.2 Some challenges with cancer treatment

A common challenge dealing with cancer treatment is the waiting time between the diagnosis and treatment. A trained radiologist can spend

more than 4 hours to evaluate and delineate a single case [4]. Although Loureiro et al. [5] concluded that the waiting time to radiotherapy shows no significant prognostic impact, the time spent could be costly for the hospitals and intolerable for the patients. In addition, resources are known to be scarce in the healthcare sector [6] and any time saved for the physicians is valuable.

Another challenge is the accuracy of the delineations of tumor volumes by the radiologists [7], [8]. Due to interobserver variability, the radii of the tumor delineations from radiologists might deviate with 0.3 cm [7] and there are often inconsistencies even if guidelines are provided. In the study conducted by Weiss and Hess [7], they could report that the uncertainties from organ motion and patient positioning was smaller than the uncertainties from tumor delineations. This was also demonstrated by Rusten et al. [9], who explored tumor delineations based on PET and MRI made by three experienced radio oncologists. Precise delineation of target volumes is considered the weakest link and the largest source of uncertainty in radiotherapy planning [8], [10]. This will, certainly, depend on the region in which the tumor delineation is performed. Nonetheless, in most cases, the precision of the delineated area is important for the further cancer treatment and might be a crucial factor for both the outcome of the treatment, as well as for recurrence and life quality of the patient post cancer treatment. An inaccurate delineation could lead to irradiation, and thereby damage, of healthy tissue which may cause discomfort for the patient.

1.1.3 Artificial intelligence in the healthcare

The interest in artificial intelligence (AI) has been growing during the last few years, especially with the increased availability of both computational power and data [11]. Today, companies such as Google, Apple and Huawei use AI for semantic image segmentation in computer vision tasks to, for instance, extract the foreground in images [12], [13]. The use of AI for the purpose of segmenting biomedical images has been a popular and interesting approach for the healthcare industry [6], [11], [14]–[16].

However, there is also skepticism regarding the use of AI in healthcare. Physicians, such as radiologists and pathologist, might be worried about losing their job [6]. Other concerns regarding the use of AI are whether it

can provide trustworthy and accurate medical information, and, certainly, the question of privacy and security of medical data [6].

Nonetheless, using machine learning to automate some of the routine tasks of a physician or providing a radiologist with suggestions for delineations could reduce some of the workload in healthcare [6]. This could save time for the physicians, decrease the chance of burnout and give them more time for other challenges that require their attention. The blend of AI and human experience is believed to be a natural settling point which may improve the delivery of care [6]. In addition, the 'Ethics guidelines for trustworthy AI" [17], requires that proper oversight is ensured while developing an AI system, by for instance, utilizing a 'human-in-command' approach.

1.2 Aim of this Master's thesis

This project is part of the observational study Anal Cancer Radiotherapy (ANCARAD, NCT01937780) [3], led by Marianne G. Guren, MD, PhD from Oslo University Hospital (OUH). All patients in this study were scheduled for chemo-radiotherapy in the period 2013 to 2016. This is a prospective study of treatment outcome, where the effect of the treatment, in terms of survival, recurrence and life quality, are followed up for 5 years. The delineations of the tumor tissue volume often carry a high degree of uncertainty [9], [10]. As a sub-study, the project explores potential aids for identifying and delineating tumor tissue. The author of this thesis has worked with autodelineation of medical images for anal cancer patients provided by Oslo University Hospital.

The aim of this thesis is to increase the knowledge about automatic tumor delineation for patients included in the ANCARAD study, but also for automatic delineation of cancerous tumors in general. The results of this project could give indications of how the overall research for using AI, and especially convolutional neural networks (CNN), for the purpose of semantic segmentation of medical images may be conducted. Such a tool could potentially save time for the radiologists and increase their efficiency and performance in their work. The CNN architecture used in this thesis are based on the framework provided by Yngve Mardal Moe in his MSc for the Norwegian Univeristy of Life Sciences, February 2019 [18].

Moe's MSc thesis [18] inspects semantic image segmentation using a CNN on PET and CT images from 197 patients with head and neck cancer, also in cooperation with OUH. The dataset for head and neck cancer consisted of PET and CT images, however, the ANCARAD dataset also contains images from MRI scans. This thesis will therefore compare the Dice performances of tumor autodelineations for images of the anal cancer patients based on PET, CT and MRI, and propose which of the imaging modalities, or a combination of the imaging modalities, seem most promising for the purpose of autodelineation. Furthermore, this thesis will explore the effect of some additional techniques added to the proposed framework of Moe [18] to increase the model performance of the delineations.

1.3 Organization

This thesis will in chapter 2 and 3 give an introduction to the theory behind the methods used in this project, in chapter 2 and 3. Basic knowledge and concepts of the imaging modalities provided in the ANCARAD dataset will be presented. Chapter 3, covers principles of artificial intelligence and the basic theory behind the code used for the tumor autodelineation in medical images. The next chapter presents the dataset and actions taken for data quality assurance. In chapter 5 assumptions, preprocessing of the dataset and the experiments conducted are described. The results from these experiments are presented in chapter 6. Evaluation of the choices made for the experiments and the results, along with possible improvements of the methods used are discussed in chapter 7. Finally, chapter 8 provides conclusions of the results and the experiments.

Chapter 2

Medical imaging

2.1 Principles of medical imaging

Medical imaging gives the opportunity for physicians to examine and make a clinical assessment of the interior of the human body without performing an invasive surgical procedure on the patient. Imaging is an extremely useful tool in diagnostic medicine [19]. Today, different medical imaging techniques are crucial for clinical diagnosis, treatments and monitoring of medical conditions. There are different imaging modalities that can be utilized depending on the type and site of the lesion. Some of these are Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) and ultrasound [19]. Effective, safe and high quality imaging is pivotal for the outcome of these medical examinations.

2.1.1 Resolution

The resolution of the image provides a measure of the imaging quality. If, for instance, lesions or other medical conditions are not apparent in the medical image, they might not be detected and the diagnosis of the examiner could be incomplete. The resolution of medical images depends on the imaging modality and the corresponding physical limitations, such as, the imaging machine, imaging environments and noise or blur [20].

There are mainly three different categories for describing the resolution of medical images: spatial resolution, temporal resolution and contrast resolution. The number of elements, or pixels, that an image consists of gives its spatial resolution. An increased number of elements corresponding to an image can potentially capture more details in the imaged object, but might also be more prone to noise. Temporal resolution is the precision of a measurement from an imaging modality, based on the time the scanner takes for each measurement [20]. The temporal resolution is of little importance if the imaged objects have no or minimal motion [20]. Contrast resolution is how well the image can distinguish between intensities. Figure 2.1 provides an overview of common resolution values of cardiac imaging:

Table 2.1: Spatial, contrast and temporal resolution presented in [20] of cardiac imaging methods. Spatial and temporal resolutions for PET may vary depending on the trade-off between resolution and noise when reconstructing the images.

	Spatial resolution	**Contrast resolution**	**Temporal resolution**
CT	0.5-0.625 mm	Low to moderate	83-135 ms
MRI	1-2 mm	High	20-50 ms
PET	4-10 mm	Very high*	5 s to 5 min

* May vary depending on the radio tracer

2.2 Computed Tomography

Computerized tomography (CT) can visualize soft tissues, blood vessels and bone structures quite well and is especially known for its excellent spatial resolution (see section 2.1.1) compared to other modalities in radiology [21], [22]. CT images typically have spatial resolution between 0.5 and 0.625 mm [20]. This modality utilizes several X-ray scans to generate two-dimensional, cross sectional images in very fine slices [19]. The CT scanner consists of a ring of several hundred detectors and an X-ray source rotating along the same ring (see Figure 2.1) [23]. The patient is placed on a bed that can slide in and out of the center of this ring [21],

[22], generating several slice-images, resulting in a CT image sequence of the region of interest.

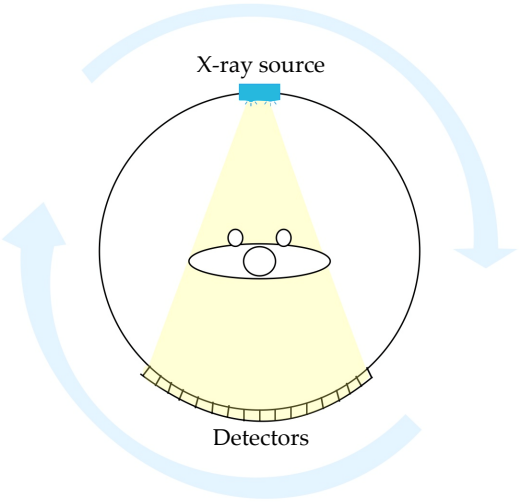


Figure 2.1: Illustration of the mechanics of a CT scanner. With permission from Kari Kvandal [24].

Principles of CT

The X-ray source circles the patient and X-rays are beamed many times along the ring. As the X-ray beam passes the tissue one can calculate the attenuation coefficient in the volume of the imaged object [22]. The reconstruction of the density of the traversed tissue, can be explained by the simplest form of Beer's law:

$$I = I_0 e^{-\mu \Delta x} \tag{2.1}$$

where I_0 is the initial intensity from the X-ray source, μ is the effective linear attenuation coefficient of the tissue and Δx is the length of the X-ray path [19]. I is then the intensity of unscattered rays that reaches the receiver/detector [19]. Eventually, all the calculated intensities collected by the detector can form a two-dimensional matrix representing the densities of the tissues in the imaged body.

Since each matrix element in the reconstructed image represents a volume of the tissue of the patient, a voxel, the resulting attenuation coefficient,

for that particular element, is the sum of the attenuation coefficients through the volume [22]. Moreover, each voxel has a degrading contrast, resulting in blurred boundaries for the objects in the image. A filtered back projection algorithm is used when reconstructing the image to avoid the blurriness [22]. The filtered back projection works like a filter and leaves the resulting image object with sharper edges [22].

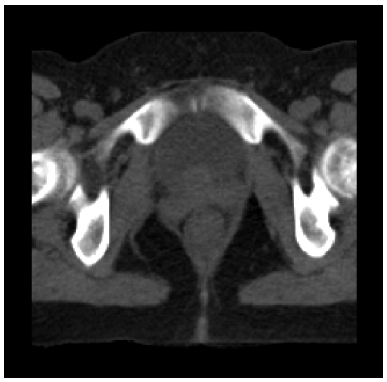


Figure 2.2: CT image of the anorectal region in an AC patient ('M033', slice 27). The white pixels represent bones, the gray areas are soft tissue (including muscles) while fat is shown in dark gray. In about the center of the image, right above the intergluteal cleft, one might dicern an oval object, which is a cancerous tumor.

Voxel values

A CT number that determines the voxel value can be detected from the reconstructed image. CT numbers are generically the gray-level data values in CT images, but the values may vary between the different scanner vendors and even between each scan [22]. The values are expressed in Hounsfield units (HU), where air has a value of -1000 HU, fat typically varies between -60 and -120 HU, water is 0 HU and compact bone has a CT number higher than 1000 HU [22]. The CT number can be calculated by:

$$CT_{number} = 1000\frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \tag{2.2}$$

where μ_{tissue} and μ_{water} are the linear attenuation coefficients of the tissue and water, respectively [25]. The beam attenuates differently according

to the tissue type and the corresponding density, and tissues with similar densities will have similar gray levels in the image. In Figure 2.2 the soft tissues have a dark gray nuances while bone structures are bright. Muscles and air have low attenuation coefficients, resulting in very dark voxels in the two dimensional image.

The human eye can in the most optimal conditions differentiate between about 720 different shades of gray. A CT image can, however, potentially contain more than 65 000 shades of gray [26]. In addition, the attenuation coefficient of a voxel of about 1 cm in diameter, must differ from its surroundings by at least 10 % in order to be distinguished from the surroundings [23]. As a result, examining areas where the tissues have similar densities can be challenging.

CT Windowing
Radiologists use CT windowing to adjust the interval of gray levels by manipulating the CT numbers [27]. The main purpose of CT windowing is to better differentiate the organs and tissues in the region of interest, or to highlight structures. The brightness of the image is adjusted by the window level (L), while the contrast can be adjusted by the window width (*W*) [27]. Typical window values for soft tissue in the abdomen are, for instance, $W = 400\ HU$ and $L = 50\ HU$, but may vary depending on the vendor and institution [27].

Contrast medium
Each CT scan is customized specifically according to the body, the region and the condition that is to be examined [21]. In most cases, the patient will be given a contrast medium injection to show, for instance, blood vessels more clearly when reconstructing the medical images [21], [22]. Contrast medium might also make cancerous tissue more apparent as opposed to surrounding healthy tissue. The contrast medium has a higher attenuation coefficient than, for instance, blood. CT images of blood vessels injected with contrast medium will therefore obtain a higher CT number in the reconstructed image.

2.3 Positron Emission Tomography

In Positron Emission Tomography (PET) the patient is injected with a positron-emitting radioactive tracer [23]. When the positron comes to rest it annihilates with an electron resulting in two 511-keV γ photons [19], [23], [28]. The two photons leave the annihilation with 180° relative to each other as the energy and momentum are conserved [28] (as shown in Figure 2.3).

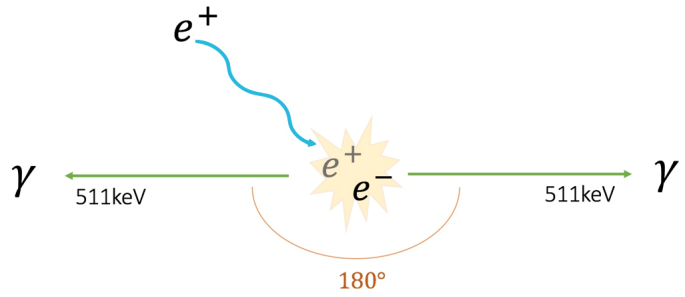


Figure 2.3: Illustration of annihilation of a positron e^+, with an electron e^-. Two γ photons with 511 keV energy are emitted. Presented with permission from Kari Kvandal [24]

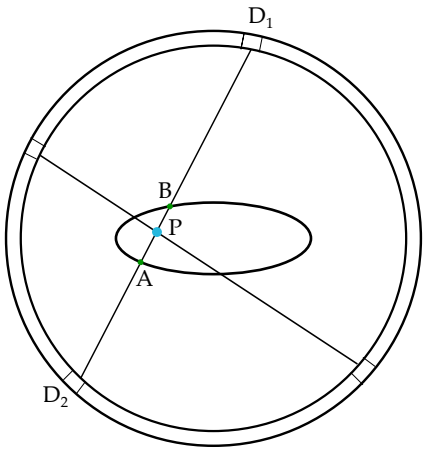


Figure 2.4: A positron annihilation from point P emits γ rays hitting detector D_1 and D_2, which will record coincident γ photons distributed along the line segment AB. The oval object in the center represents a patient. Illustration inspired by an illustration of a PET detector in *Nuclear Physics: Principles and Applications* by John Lilley (2001) [23].

The patient is placed in the center of a ring of detectors (shown in Figure 2.4). If two γ photons are detected in coincidence by detectors D_1 and D_2, they must have been emitted from the same point, P (see Figure 2.4) [23].

Eventually, the information gathered from all the detector pairs in the ring generates a PET image slice [23] of the scanned region, in vivo. The generated image does not capture all of the photon pairs emitted from the scanned region. Far more photon pairs will leave the body undetected because they are not in the plane of the detector ring [28]. However, the distribution of the count rate detected in one direction, will be a projection of the real distribution of radioactivity [28].

Possible false detection
It is assumed that the two detectors have zero lag and that the γ photons hit the detectors within some small interval of time (typically 2 to 5 nanoseconds) [28]. In reality, there are several γ photons from different positrons in the imaged object that reach the detector simultaneously. As a consequence, the two observations that appear to be detected at the same time are paired up. Lag in detector response could result in false γ pairs being selected by the detectors. Such random coincidences can cause false signals in the PET image [28]. In addition, positron emission also occurs due to scattering or absorption of one or both of the annihilation photons [19]. During the reconstruction of the PET data, an attenuation correction process is applied to restore the quantitative accuracy and qualitative integrity of PET [28].

FDG
The most widely used radionuclide for tracer in PET is Fluorine-18, ^{18}F, which decays 97% by positron emission and has a nearly 2-hour half-life [28]. The radionuclide is often combined with glucose to highlight areas of increased metabolic activity [23]. Consequently ^{18}F becomes the radioactive tracer ^{18}F-FDG. Due to the high metabolism in tumors relative to healthy tissue, the absorption of ^{18}F-FDG is high and tumors light up in PET images [29]. For untreated tumors, the FDG uptake in a wide range of tumor types has often shown to be well and positively correlated with the cell number in that tumor [28].

SUV

The uptake of the radioactive tracer may vary between each PET scan. The two most significant sources of variation are the patient size and the concentration of radioactive tracers injected [30]. Therefore, the relative tissue uptake of the radioactive tracer is often used [30]. As a standardized measurement of the uptake, the standardized uptake value (SUV) is used [28], [30]. SUV is the ratio between the image derived radioactive tracer concentration C_r and the concentration of radioactivity in the whole body C_b, which can be calculated by the injected dose d and the body mass m:

$$SUV = \frac{C_r}{C_b} = \frac{C_r}{\frac{d}{m}} = \frac{C_r m}{d} \tag{2.3}$$

False positives

High SUV should result in higher coincidence count rate and thus brighter voxels in the reconstructed image [29]. However, tumors are not the only tissue that absorb the radioactive tracers. Lymph nodes and tissues with lesions such as inflammation, auto immune processes or infection also have high metabolism, resulting in high uptake of, for instance, FDG [29]. Moreover, variable gas in the bowel can lead to false areas of increased uptake. The presence of ^{18}F in urine, when dealing with ^{18}F-FDG as tracer [29], will result in a bladder that lights up in the generated PET image.

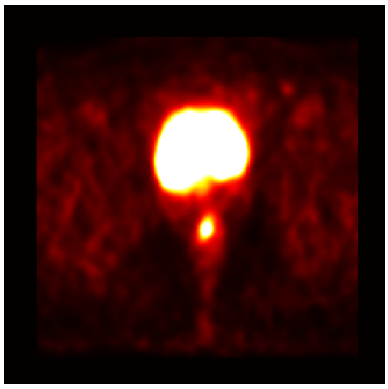


Figure 2.5: PET image of the anorectal region in an AC patient ('M033', slice 27). Tissues with uptake of the ^{18}F-FDG tracer light up. The large, bright area represents the bladder, while the smaller, bright area is a cancerous tumor.

This carries the risk of false positives and misdiagnosis of the patient's condition. The surgeon and oncologist must therefore not base their

diagnosis of the lesion solely on the PET image [29]. In Figure 2.5 there are two areas that are especially bright. The upper, larger area represents the bladder while the lower, smaller area is tissue of a cancerous tumor. The bladder is neither cancerous nor a lesion, but will light up in the same manner as the tumor.

PET/CT scanners

The PET image is dependent on the tracer uptake. However, localization of the tracer activity is difficult or sometimes even impossible [31] since the generated images provide relatively little anatomic information. In addition, images generated from PET scans have spatial resolution typically between 4 and 10 mm [20], [28] which is poor compared to CT or MRI.

Image fusion is a technique used to form an ”anatometabolic” image of PET and either MRI or CT [28]. At first, this was just a software approach, where the images from the different modalities were fused subsequently [28]. The combination of both anatomic and metabolic data makes it much easier to localize the tracer activity [31]. Today, a PET/CT scanner can take both images during a single examination. By doing so, the anatomic structures in the images are more likely to match, and localization of the PET signals are more likely to be correct. In addition, the CT images can be used for more precise attenuation correction of the PET data [31]. Figure 2.6 is an example of PET and CT images acquired from a single examination can be fused.

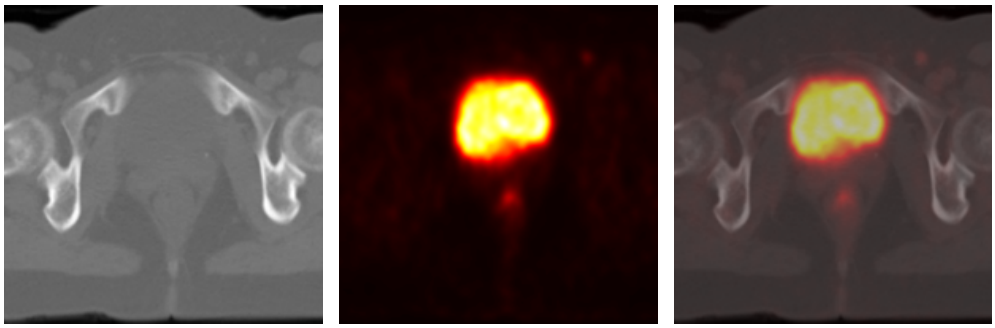


Figure 2.6: An example of how a CT image (left) and PET image (middle) of the anorectal region of an AC patient (’M033’, slice 27) can be fused (right).

2.4 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is known for its high contrast resolution, providing detailed images [32]. As a result, it is very good at differentiating between soft tissues of different densities. MRI, as opposed to PET or CT, uses properties of stable atomic nuclei to obtain images of the interior of the imaged object [23]. As a result, the patient is not exposed to any risk of ionizing radiation.

Principles of MRI

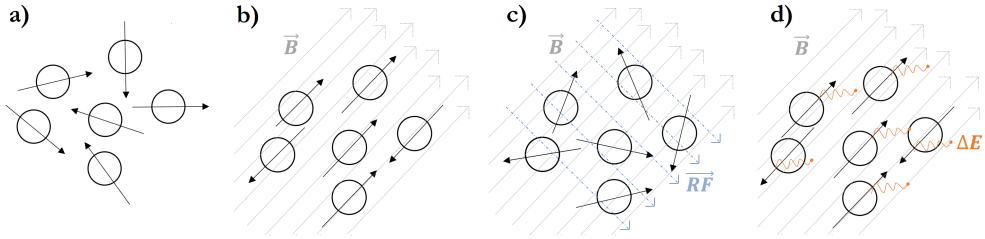


Figure 2.7: Simplified illustration of the steps in MRI. Starting off with a) randomly oriented nuclei, followed by b) aligned nuclei with an static, external magnetic field with $\vec{B}$, then c) a radio frequency pulse $\vec{RF}$ is added, tipping the nuclei spins and after a while d) the nuclei flip back in alignment emitting radio frequency energy ΔE.

In MR imaging one utilizes the abundance of hydrogen nucleus in water and fat. The hydrogen nuclei in the human body is normally randomly oriented with an angular momentum (spin) as in Figure 2.7 a). The patient is exposed to a powerful, static magnetic field that aligns the orientation of the nuclei either in parallel or anti-parallel to the magnetic field [19], [32], as shown in Figure 2.7 b). The nuclei are disrupted by an external radio frequency (RF) energy pulse, causing the protons to flip to a higher energy state [19], [32]. The spins of the protons are now tipped away from the direction of the static magnetic field as illustrated in two-dimensions by Figure 2.7 c). A certain period after the initial radio frequency, the excited nuclei spontaneously return to their lower energy (relaxing) state, emitting RF photons in the process [19]. In MRI the emitted RF energy, ΔE, is a measure given by:

$$\Delta E = 2\mu_p B = h\gamma \tag{2.4}$$

This energy is dependent on the magnetic moment of the proton, μ_p, and the magnetic field B, but can also be described in terms of Planck's constant h and a frequency γ (see Equation 2.4). It is important that the RF pulse matches the frequency γ (which is also called the Larmor frequency) in order to excite the protons [23].

Images acquired from MRI

There are many methods by which MR images can be acquired and reconstructed [19]. One can, for instance, vary the sequence of RF pulses applied and collected [32]. The time between each successive pulse sequence is called the repetition time (TR) and the time between the applied pulse and the collection of the emitted, echo signal is called the time of echo (TE) [32].

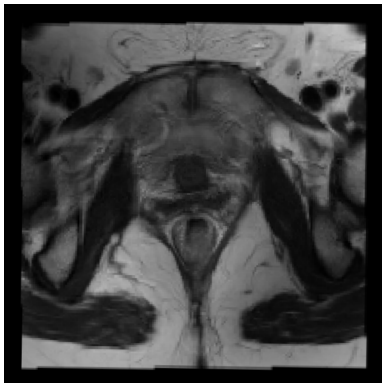
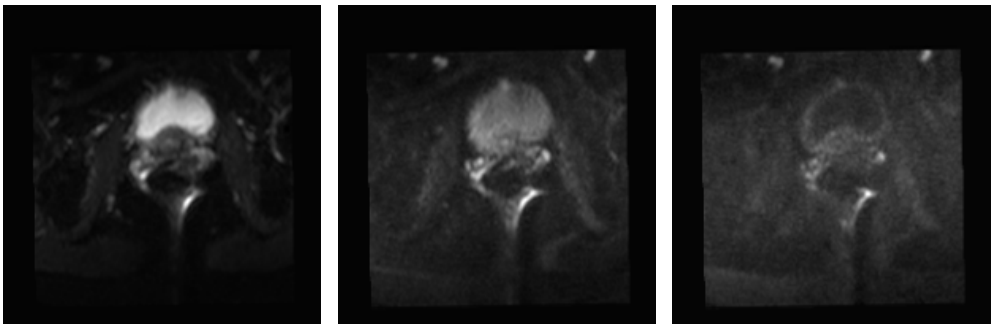


Figure 2.8: T2-weighted image of the anorectal region of an AC patient ('M033', slice 27). Here, fat appears in a white or light gray tone and muscles become dark gray and almost black. The areas with gray levels in between these, are soft tissue. In about the center of the image, right above the intergluteal cleft, an oval object, which is a cancerous tumor can be seen.

T1-weighted images are a result of using short TR and TE. This causes soft tissues and fat to appear in lighter shades of gray while tumors, inflammation or cysts become darker gray. By increasing the TR and TE one can generate a T2-weighted (T2W) image, which is more commonly used. In a T2W image, the soft tissues are darker in comparison with the T1-weighted images, while tumors, inflammation and cysts appear in a lighter shade of gray. Figure 2.8 shows an example of a T2-weighted image.

Another method of collecting the data in MRI is by generating diffusion weighted images (DWIs). DWIs exploits the random motion of water nuclei [33] and is especially sensitive for detecting restricted water movements, such as the flow in blood vessels. Detected diffusion will light up in the reconstructed images [32]. DWI is also widely used to assess stroke, which is often visible by DWI before any T1-weighted or T2-weighted image [34].

By adjusting the timing and strengths of the gradients for constructing a DWI, one determines the degree of diffusion weighting and can capture different diffusion processes in the imaged object [34]. The degree of diffusion weighting is also referred to as the 'diffusion sensitivity', 'b-factor' or 'b-value' [34], [35], and has unit s/mm^2. Higher b-values capture slow moving water nuclei, while lower b-values capture the more fast moving water nuclei [35]. In Figure 2.9 one can observe that the water nuclei in the bladder are slow moving, since the signal from the bladder is much higher for lower b-values.



(a) $b = 0\ s/mm^2$ **(b)** $b = 800\ s/mm^2$ **(c)** $b = 1500\ s/mm^2$

Figure 2.9: DWIs of the anorectal region for three different b-values in an AC patient ('M033', slice 27). The brightest area in (a) is the bladder and slightly above is a cancerous tumor. Note that the water nuclei flow, in this case, seem to surround the tumor but is not present in the center of the tumor, located about in the center of the images.

By combining two or more DWIs, of different b-values, one can generate an Apparent Diffusion Coefficient (ADC) map [33]. The aim of ADC maps is to obtain a less noisy image containing more information than just one single DWI. The gray levels in an ADC map reflect the degree of diffusion

of water molecules through different tissues [33]. For instance, blood vessels can more easily be differentiated from muscles due to the stream of water molecules. Regions with no or very few water molecules in motion, such as air or bones, will appear much darker in the ADC map.

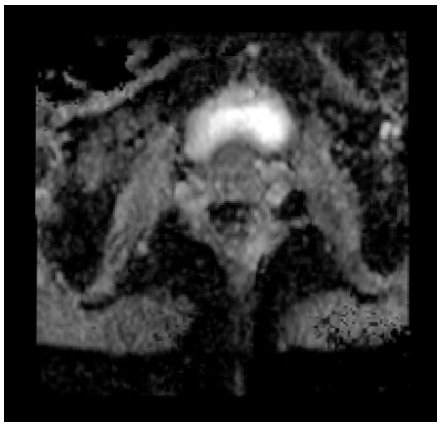


Figure 2.10: ADC of the anorectal region, created from MRI of an AC patient ('M033', slice 27). This ADC map is made based on the b-values 0 s/mm^2, 10 s/mm^2 and 20 s/mm^2.

2.5 Volume delineation

As of today there is a high degree of uncertainty associated with the target volume [10], that is the volume of a lesion which is of interest. Errors in the target volume might be caused by motion of the target, errors in the positioning of the patient or the delineation of the target volume. Radiologists are encouraged to use international guidelines for the definition of target volume, such as those provided by the International Commission on Radiation Units and Measurements (ICRU) [36]. However, this does not ensure that the inter and intrapractitioner variability of the delineations is sufficiently small [7], [37]. A study of interobserver variability [7] found that the ratio of the contoured volume for tumors in the prostate ranged between 1 and 1.6 [7]. Another study conducted by Guda et al. [8] on data of head and neck cancer patients, found that the overlap of GTV delineation, made by three radiation oncologists with 10 years of individual experience, was moderate to good (Dice similarity coefficient of 0.57 without PET and 0.69 with the use of information from PET).

Different delineations are used to describe the target volume. The guidelines [36] describe, among others, the following volumes:

- Gross Tumor Volume (GTV)
- Clinical Target Volume (CTV)
- Planning Target volume (PTV)

GTV is defined as the most probable position and extent of the tumor, which is visible [10]. The GTV may include the primary tumor, metastatic regional nodes (such as lymph nodes) or distant metastasis (spread of cancer) [36]. A complete and accurate description and report of the GTV is required for staging of the cancer, dose planning and evaluation of the CTV and the PTV [36]. Moreover, this report should preferably specify the diagnostic modality used since this can vary, as may the methods used to delineate the GTV [36]. By combining clinical examinations and the use of various imaging modalities, the radiologist has more information about the extent of the target volume. Several methods may have been used when evaluating the generated image and determining the size of the GTV. Therefore, the radiologist should specify on what basis the decisions for the delineation have been made [36]. The GTV may also be used for non-malignant lesions that are to be treated with radiation [36].

The GTV is often surrounded by tissue that is subclinical. This tissue might contain cancerous tumor cells which cannot be detected through clinical examination. The CTV includes the GTV in addition to the assumed subclinical microscopic tumor spread [10] (see Figure 2.11). The subclinical patterns might often be hidden because of the resolution limits in imaging techniques [10]. Based on clinical experience, this is accounted for by adding a margin of, for instance, 2 cm around the GTV to generate the corresponding CTV [10]. A CTV of a benign tumor (a non-cancerous tumor) may not be generated since there is no risk of microscopic tumor infiltration [36]. The CTV will in this case coincide with the GTV.

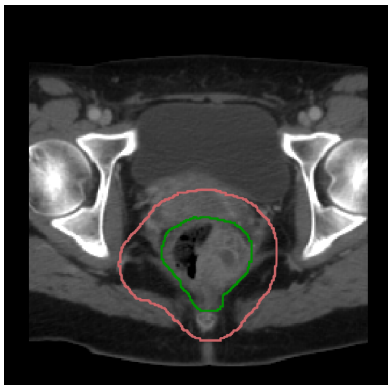


Figure 2.11: The delineation of GTV (in green) and CTV (in pink) for an AC patient ('M007', slice 21), on a CT image of the anorectal region.

The PTV was introduced for treatment planning and evaluation [36]. It is mainly used to ensure that the treatment dose will be delivered to the CTV with an acceptable probability [36].

2.5.1 Approaches for tumor delineation/diagnosis

If the tumor is accessible, a physician starts by performing a physical examination of the patient and looks for circular lumps in the affected area [1]. Furthermore, a radiologist may take one or more scans of the patient, depending on the region affected, the condition of the patient and the assumed stage. At last the clinical data from blood tests, the physical examination and imaging of the interior of the patient, is used to finally make a delineation of the target volume on one of the imaging modalities [8].

The process of delineating may be based on a combination of factors and needs to be carefully evaluated [36]. As a result, this can be a very time consuming process, taking between 18 minutes and 2.7 hours on an average [4]. For anal cancer, the oncologist would start by localizing the primary tumor using PET images. Next, he/she would consider an extension of the delineation depending on how the surrounding tissue seems in the MRI or CT image used [9].

In recent years, the exploration of artificial intelligence for the purpose of segmenting biomedical images has become popular [11], [14]–[16]. This

is mainly due to the increased availability of computational power and the increased available medical image data [11]. Automatic delineation of medical images, using artificial intelligence, to detect and segment tumors could potentially save time and resources for the hospitals, but in addition potentially uncover new information about the properties of medical lesions.

Chapter 3

Artificial intelligence

3.1 Basic principles of artificial intelligence

The main idea behind artificial intelligence (AI) is to give computers the ability to learn, and potentially improve, the performance of their tasks. As a subfield of AI, machine learning focuses on self-learning algorithms that extract knowledge from a given dataset to make predictions in classification or regression problems from new data [38]. The learning algorithms for computers is inspired by how a biological neuron transmit signals in the brain [38].

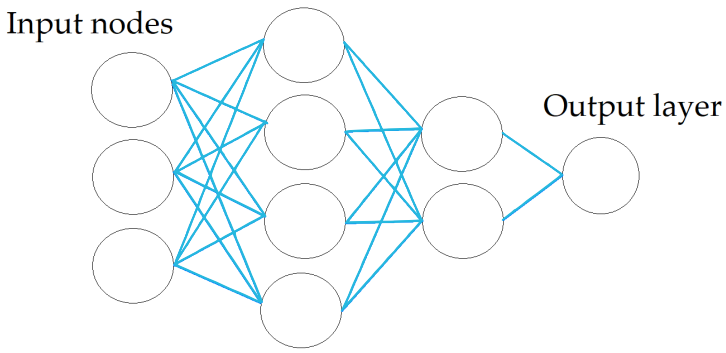


Figure 3.1: Illustration of how the composition of a Neural Network can be. The circles represent activation units. The number of activation units in the first layer depends on the number of variables in the input data. The final layer is the output signal from the network. In between are the hidden layers, where the information is processed. The blue lines represent connections, each with their own weight.

Artificial neural networks can consist of nmerous layers of neurons, that each evaluate its input signals and supply a processed signal to the next layer [38], as shown in Figure 3.1. Each connection is weighted, describing how important the connection is relative to the rest. Prior to training, it is common to either set all weights to zero or small random numbers [38]. When all neurons in a layer are fully connected with all of the neurons in another layer, the layer is called a Fully Connected Layer [38], [39]. Figure 3.1 is an example of a network consisting of Fully Connected Layers.

3.1.1 Neural Network architecture

Neural networks can have a architectures similar to the one presented in Figure 3.2. The input samples and the corresponding weights are processed through a net input function, an activation function and the weights are updated.

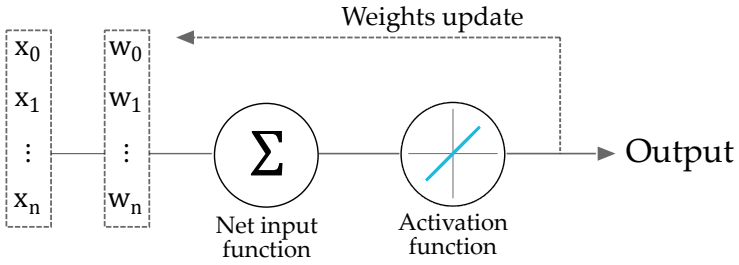


Figure 3.2: Illustration of how the architecture of a Neural Network can be. The two boxes to the left represents the input samples vector **x** and the weights vector **w**, respectively. The circles represent functions the input samples are processed through before the model obtains an output.

Activation

Based on the information from the network, the activation function is used to compute a prediction for a given sample. Each neuron can have, for instance, a linear activation function ϕ given by

$$\phi(z) = \mathbf{w}^T\mathbf{x} = a \tag{3.1}$$

where z is the net input (Σ in Figure 3.2) computed with the weights vector $\mathbf{w}^T$ (transposed) and the input samples vector **x**. The scalar a is the

resulting activation which will be forward propagated to the next layer [38]. For a binary classification task, a threshold for z is used in the last layer to decide which of the two classes the sample may belong to. When working with a regression task, an activation function that provides a more continuous range of outputs would be favoured. In this way, the choice of the activation function depends on the desired outputs.

All neuron in each layer of a neural network must be activated by a particular activation function, in order to provide an input value a for the next layer [38]. Another example of an activation function is Rectified Linear Unit (ReLU), defined as:

$$\phi(z) = max(0, z) \tag{3.2}$$

ReLU will only send an activation signal to the next neuron layer if the input value is above zero (see Figure 3.3). The advantage of ReLU is that it introduces non-linearity for the activation, as apposed to the linear activation function 3.1. One disadvantage with ReLU is in the case where the input values z are consistently negative, inhibiting that particular neuron to activate. This is referred to as the "Dying ReLU" problem.

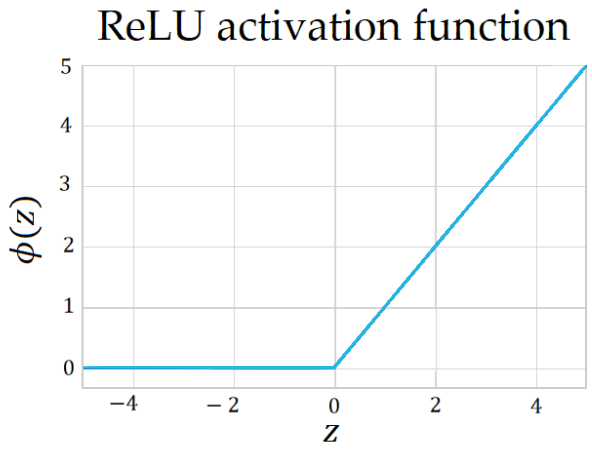


Figure 3.3: ReLu activation function where the x-axis represents the net input values z and the y-axis represent output of the activation function ϕ.

Deep learning is a machine learning technique developed to train such artificial neural networks [38], often used for classification tasks. When training a deep learning network, one iterate through the network several times and the weights of the connections are updated (as showvn in Figure 3.2).

3.1.2 Model optimization

Machine learning is essentially an optimization problem. By iterating the signals through the network the goal is to optimize the weights, and thereby improve the performance of the model. One can compare the connections between the neurons and the weights in a neural network to human brain connections: connections that are often used and are considered important are strengthened while connections that are not used will eventually become very weak.

Loss function
The update of the weights are made in order to the minimize the outcome of a loss function [38]. It is also sometimes referred to as the error function [39] or a cost function. A loss function J can be any wanted metric, defined as a function of the weights, $\mathbf{w}$. One example is the Sum of Squared Errors (SSE):

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i}(y^{(i)} - a^{(i)})^2 \tag{3.3}$$

where i represents the current sample, $y^{(i)}$ is the true class label and $a^{(i)}$ is the predicted class label, for sample i, from an activation function [38].

Gradient descent for cost minimization
The optimizers task is to update the weights in a way that will lead to a lower loss [38], [39]. For each sample, the weights are updated according to the output of the loss function and the optimization function, as given in Equation 3.4. Here the new weights $\mathbf{w}^{i+1}$ are the sum of the weights in the current layer, $\mathbf{w}^{(i)}$, and an update for the weights, $\Delta\mathbf{w}^{(i)}$.

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} + \Delta\mathbf{w}^{(i)} \tag{3.4}$$

The simplest approach for updating the weights is using the gradient of the loss function [39]:

$$\Delta\mathbf{w}^{(i)} = -\eta\nabla J(\mathbf{w}^{(i)}) \tag{3.5}$$

where η is the learning rate and $\nabla J(\mathbf{w}^{(i)})$ is the gradient of the loss function [38], [39]. The update of the weight will be in the opposite direction of $\nabla J(\mathbf{w}^{(i)})$, which should be where the loss function has the greatest decrease. This approach is known as the gradient decent [38]–[40].

Perhaps the more commonly go-to optimizer in deep learning today is the one called 'Adam'. Adaptive moment estimation was proposed by Kingma and Ba [40] for efficient stochastic optimization. Such an optimizer, is computationally less expensive relative to simply using the gradient descent as presented above. Adam uses less iterations through the network before the loss value converges, and is known for its robustness and that it is suited for a wide range of optimization problems [40]. The algorithm behind Adam uses the estimated mean of the gradient for the next layer $\hat{m}^{(i+1)}$, the uncentered variance of the gradient $\hat{v}^{(i+1)}$ and an error or noise parameter ϵ, in addition to the learning rate η [40]:

$$\Delta \mathbf{w}^{(i)} = -\eta \frac{\hat{m}^{(i+1)}}{\sqrt{\hat{v}^{(i+1)}} + \epsilon} \tag{3.6}$$

The gradient of the loss function is used to update the estimated mean and uncentered variance of the next gradient.

3.1.3 Training, validation and test set

The process of optimizing a neural net is called training [38]. It is here the model learns relevant patterns of the input samples. How much a network can learn depends on the number of weights or parameters, and are often referred to as the capacity of the network [38]. The samples used for optimizing are called the training set or the training samples. In addition to a training set, it is also necessary to have a validation and a test set.

The purpose of a validation set is to validate the proposed weights after the training, and observe how the model performs on new, unseen data [38]. Depending on how poor or well the performance is on the validation set, one can then go back to training the network. The test set is used as a final evaluation of the network, and contains unseen samples.

Splitting the data into training, validation and test set should be as stratified as possible. Consequently, each sub-dataset should be as representative of the true diversity in the data as possible. In addition, a well-represented training, validation and test split should reduce the chance for getting a biased autodelineation and thereby decrease the chance of overfitting [38].

3.1.4 Overfitting and Regularization

If the models provide excellent results during the training, but perform much worse on new, unseen data, the model is overfitted. An overfitted model has managed to capture pattern in the training data well, but performs poorly on unseen data [38]. Neural networks are prone to overfitting the data, and the main reason for this is the lack of invariance in the training set [38], [39].

To limit the chance of overfitting one can increase the invariance in the data. In data augmentation new variations of the existing data, or a subset of the existing data, are created by transforming the data. If the augmented data changes the expected target, the target should also be transformed in the same manner. Another approach for regularizing the network is to add penalization on the weights when training the network [38], [39].

Batch size
The weights are updated after each sample or each batch of samples [38]. The number of updates may therefore be dependent on the batch size. The batch size is how many samples the network should use for each weight update, and the bigger the batch size, the more generalized and less overfitted will each weight update, and eventually the model, become. But, by decreasing the batch size, one may capture structures of noise in the data, which may not be relevant for the prediction.

Dropout
When using a Dropout-activation, a chosen fraction of neurons in the layer are randomly dropped for each iteration [41]. This is often referred to as the keep probability. Dropout can be viewed as the averaging of an ensemble of models [38]. Srivastava et al. [41] explain that Dropout can prevent overfitting as well as ”approximately combining exponentially many different neural network architectures efficiently”. By not activating a random set of neurons in a layer, the network is forced to learn a robust and redundant representation of the data. Dropout is often applied to one of the higher layers [38], [41].

3.2 Performance of a classification model

How the performance of a model should be defined depends on the requirements of the specific problem. There are numerous ways to quantify the performance of a model and measure a model's relevance. Perhaps the most commonly used performance metric is the classification accuracy, which is defined as:

$$accuracy = \frac{\text{Successfully classified samples}}{\text{Total number of samples}} \tag{3.7}$$

This is generally a useful metric [38] but does not cover any information about the degree of error for each classification instance. For example, in the presence of a major imbalance between two classes, a model could gain a high accuracy score by simply assigning all instances to the class with highest frequency.

3.2.1 Confusion matrix

Table 3.1: The typical setup for a confusion matrix, where the x-axis represents the predicted class and the y-axis are the actual class. The values within each box is typically presented as a frequency.

		Predicted class P	*Predicted class* N
Actual class	P	True positive (TP)	False negative (FN)
	N	False positive (FP)	True negative (TN)

Several performance metrics are based on the confusion matrix. A confusion matrix reports the occurrences of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) predictions of a binary classification model [38], as displayed in Table 3.1. For a binary

classification problem with classes 'positive class' and 'negative class', True Positives is the count of instances that were correctly classified as 'positive class'. False Positives are the instances that were classified as 'positive class' but in reality belong to 'negative class'. Furthermore, True Negatives are the instances that were correctly classified as 'negative class' whereas False Negatives are instances misclassified as 'negative class'.

3.2.2 Overlap based metrics

Overlap based metrics are performance metrics that focuses on the overlap of the predicted classification and the true classification. Two of these are Specificity and Recall. Specificity is essentially the true negative rate (TNR), while Recall is the true positive rate (TPR) [38], [42]:

$$Specificity = TNR = \frac{TN}{TN + FP} \tag{3.8}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{3.9}$$

Specificity is the fraction of instances correctly classified as negative (the true negatives) compared to the total number of instances that should be negative [42]. Recall, on the other hand, is the fraction of instances correctly classified as positive, given the total number of instances that really are positive. These measures are much more sensitive for small segmentations compared to bigger segmentations, and are therefore not common to use as evaluation of medical image segmentations [42]. Precision, on the other hand, is commonly used for validating medical images [42]. This metric measures the rate of true positives based on the total number of instances that were classified as positive. Precision is therefore also called the positive predictive value (PPV) [42] and is mathematically defined as:

$$Precision = PPV = \frac{TP}{TP + FP} \tag{3.10}$$

A combination of PPV and TPR provides the F-score, also named 'the Dice coefficient' or 'the overlap index':

$$Dice = F1 = 2\frac{PPV * TPR}{PPV + TPR} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.11}$$

The Dice coefficient is the most used metric for medical volume segmentations [38], [42]. Figure 3.4 shows the inpact of Precision and Recall on Dice.

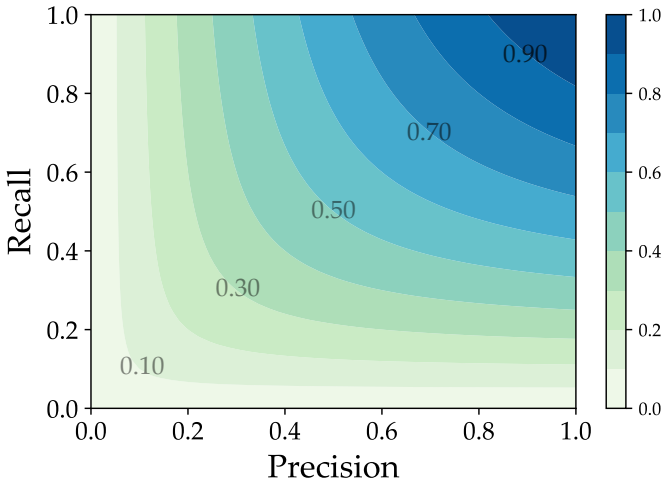


Figure 3.4: Illustration of the Dice coefficient as a function of Precision and Recall (Eq. 3.11). Precision is represented by the x-axis and the y-axis represents the Recall. The values within the square box shows the contouring for how the Dice coefficient changes.

By weighting the PPV and TPR differently, this metric can put more emphasis on one or the other, depending on what is important for the particular case. A more general definition of the F-score is therefore:

$$F_\beta = \frac{1+\beta^2}{\frac{\beta^2}{TPR} + \frac{1}{PPV}} = \frac{(1+\beta^2)PPV * TPR}{\beta^2 PPV + TPR} \tag{3.12}$$

where β is the weighting of the Precision. Figure 3.4 illustrates how the Dice coefficient evolves as the relationship between Precision and Recall changes. To obtain a Dice coefficient above 0.50, both the Precision and the Recall must be at least 0.50. If either one of them are under 0.50, the Dice coefficient cannot exceed 0.50.

Apart from the group of overlap based metrics there are also spatial distance based metrics, probabilistic metrics and pair-counting based metrics, among others [42]. A guideline for selecting evaluation metrics is described in [42] where several different performance metrics are compared for the purpose of segmenting medical images.

3.3 Image classification

Convolutional Neural Networks (CNN) can be described as 'models that were inspired by how the visual cortex of human brains works when recognizing objects' [38], [39]. There are mainly three different types of layers that make up CNN architectures: Fully Connected layers, Convolutional layers and Pooling layers [38].

3.3.1 Convolutions

Convolutional layers can be compared to applying filters to images. Each convolutional layer may extract or attenuate properties in the input image. By creating a feature map based on small patches, or subregions, of the input image [38], [39], the convolutional layer can exploit the fact that nearby pixels are more strongly correlated than distant pixels [39]. Another advantage is that the network can find patterns within each patch which can be a useful tool for image recognition. Furthermore, padding is often used on the input image.

Padding

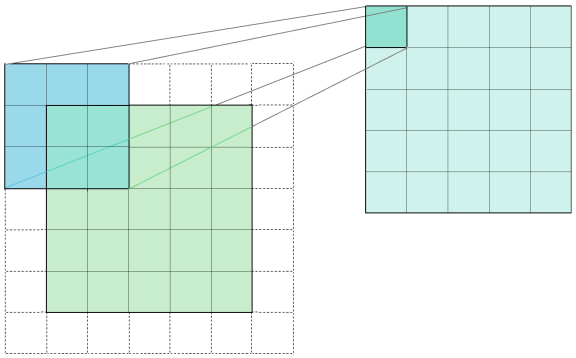


Figure 3.5: Illustration of same padding. The smaller square in blue represents the padding kernel, while the bigger green square is the input image. As the kernel moves through the input image, the output image (in seagreen) is generated.

Padding is adding pixels to a two-dimensional image in order to influence output dimension as well as how the convolution is applied to the input

image [38]. The choice of padding also affects the importance of the edge pixels in the input image [38]. 'Full padding' increases the size of the output image, 'same padding' maintains the input dimensions while 'valid padding' decreases the dimension of the image. Figure 3.5 is an illustration of same padding.

3.3.2 Pooling

Pooling layers (also called subsampling layers) are used to decrease the capacity of the network by reducing the amount of features, or pixels, in an image [38]. There are two common types of pooling: max pooling and mean pooling. For a given dimension of a patch of the input image, max pooling reduces each patch to the maximum value present. In the same manner, mean (or average) pooling extracts the mean value for each patch. The size of the patch is specified. Figure 3.6 is an example of max pooling with a patch size of 2 x 2 pixels. The resolution of the output image is considerably reduced. Larger patch size, results in more reduction in resolution and vise versa.

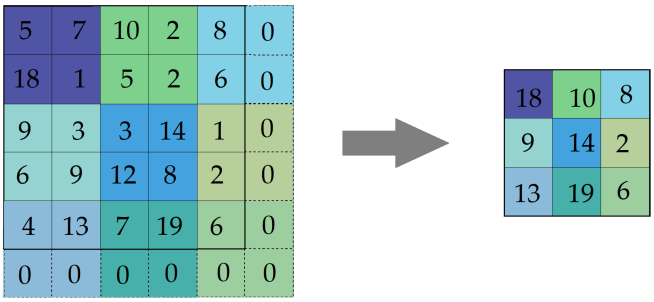


Figure 3.6: Illustration of max pooling with kernel size 2 x 2 pixels. Each shade of green and blue in the small box is the maximum value among the pixels with the corresponding color in the larger box. Note that padding is temporarily added to the input image, if needed to fulfill the kernel operations.

Including pooling layers to a network will result in a higher computational efficiency and reduces the chance of overfitting the network to the data. While pooling layers create more robust features, one also disregards the 'where' information of the sub-samples [16], [38] and loose resolution.

3.3.3 Upsampling

The closest to the opposite of pooling layers may be upsampling layers. Upsamling layers aims to reconstruct a dense map of the input data [43]. There are two types of upsampling: non-guided depth upsampling and guided upsampling. Non-guided upsampling methods often use techniques such as interpolation [43]. Guided upsampling methods upsample using guidance from a high resolution image. A third option for upsampling sparse data is to predict the depth value, such as Ma and Karaman explored in [44].

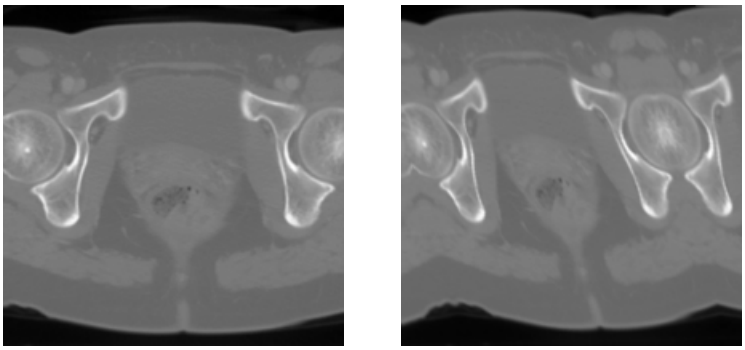
3.3.4 Image augmentation

As mentioned in Section 3.1.4, by augmenting the image data, one can present different versions of the same images, such as deformations, translations, rotations, croppings, flippings or shadings [45], among others. Consequently, one can increase the invariance in the dataset [39]. If the training data is adequate, it may already contain sufficiently different variations of the images, and a neural network can learn the invariance. However, when there is lack of training data, the network may not be presented with all the options during the training, and may perform poorly in the validation set. Consequently, image augmentation can be utilized as a technique to handle lack of data [39], [45] by increasing the training dataset.

Elastic deformation
Elastic deformation is changing the length, volume or shape of the image. Simard et al. [46] proposed an elastic deformation algorithm, which performed well on the MNIST dataset in 2003. Parameters for this algorithm are σ, α and α *affine*.

σ is explained as the Gaussian standard deviation of the voxels allowed for the deformation of the image. The larger the chosen σ, the more deformed the generated image will become. Furthermore, α is a scaling parameter that controls the intensity of the deformation [46]. At last, *affine* transformation is a transformation that preserves points, straight lines and planes in an image, although the angle between lines might not be preserved. A simple example is the transformation from $y = ax$ to

$y = ax + b$. The properties of the straight line is the same, but the line is now moved parallel to the original line, in the same plane.

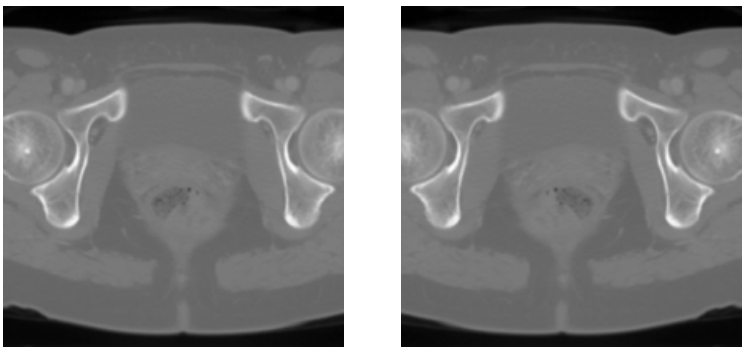


(a) Original image **(b)** Elastic deformated image

Figure 3.7: Example of elastic deformation, $\alpha = 90$, $\sigma = 15$ and α *affine*$= 25$ applied to a image of the anorectal region of an AC patient ('M007', slice 25) from the ANCARAD study.

Flipping

When the object that one aims to detect through an image classification task is independent of symmetry and the positioning in the image either horizontally, vertically or both, one can increase invariance by flipping the image [45]. An example is when classifying pedestrians in a dataset containing images of road junctions. Whether the pedestrians is located on the left or the right side of the image is of no importance, but flipping the image vertically would create an image that does not correspond to the other images in the dataset.



(a) Original image **(b)** Flipped image

Figure 3.8: Example of horizontal flip applied to a image of the anorectal region of an AC patient ('M007', slice 25) from the ANCARAD study.

3.4 Sentiment image segmentation

Sentiment image segmentation is the process of asssigning each pixel in an image to an object class [47]. Each object class needs to be delineated by boundaries [47], resulting in a partition of non-overlapping regions. Sentiment image segmentation may be applied in areas such as autonomous driving where the algorithm needs to differentiate between the road and a cyclist or a pedestrian [12]. For medical image analysis, sentiment image segmentation can be used to label organs, lesions or other regions of interest in a medical image.

Semantic image segmentation has proven to work well with convolutional neural networks [48]. Long et al. [48] showed recent results for semantic image segmentation using Fully Convolutional Networks (FCN), by achiveing a 20 % relative improvement (to 62.2 %) compared to contemporary classification networks of 2014.

A challenge by using convolutional networks for the purpose of semantic image segmentation are pooling layers [16]. Generic FCN models can only generate coarse global saliency maps, loosing detailed object structures [14]. Localization is crucial for medical images in order to give a diagnosis, thus making pooling a problem for the purpose of semantic segmentation. Qinghua Ren and Renjie Hu observed [14], that the most intrinsic challenges in deep learning methods today are to predict a saliency map with the same resolution as the input image, and increase the robustness and accuracy of the deep network.

3.4.1 Encoder-decoder architectures

An encoder-decoder architecture was introduced to maintain the resolution of the input image for a saliency map. In addition, these architectures capture context and enables precise localization [16]. This architecture class has shown to be a superior in performance in many computer vision tasks [14], [49], [50]. The root of a FCN with pooling layers is used to aggregate the features and consequently decrease the spatial resolution of the images [14], [16]. This encoding is often referred to as the contracting path. Next, the features are decoded in the expansion path with upsampling layers. While upsampling, the object details and

spatial resolution are gradually recovered using skip connections [14], [16].

Skip connections
The spatial information is recovered by merging the features skipped from various layers in the contracting path to layers in the expansion path [15], as shown in Figure 3.9. Hence, one re-uses the switch variables from the pooling layers in the contracting path, and can thereby reconstruct the detailed object structure more effectively [14]. These connections are called skip connections [15], [39]. Drozdal et al showed [15] that the choice of the combination of these skip connections can be of great importance when regarding the network performance for FCNs [15].

U-Net architecture
A popular architecture from this class is U-net, where the contraction and expansive paths are applied gradually. The illustration of this architecture often results in a 'U'-shape with a minima halfway. Figure 3.9 is an illustration of the U-net architecture created by Ronneberger et al. [16].

The need of a larger training dataset is often a challenge regarding biomedical tasks. The encoder-decoder architecture utilizes the available annotated samples available more efficiently, thus reducing the need of a larger training dataset [16].

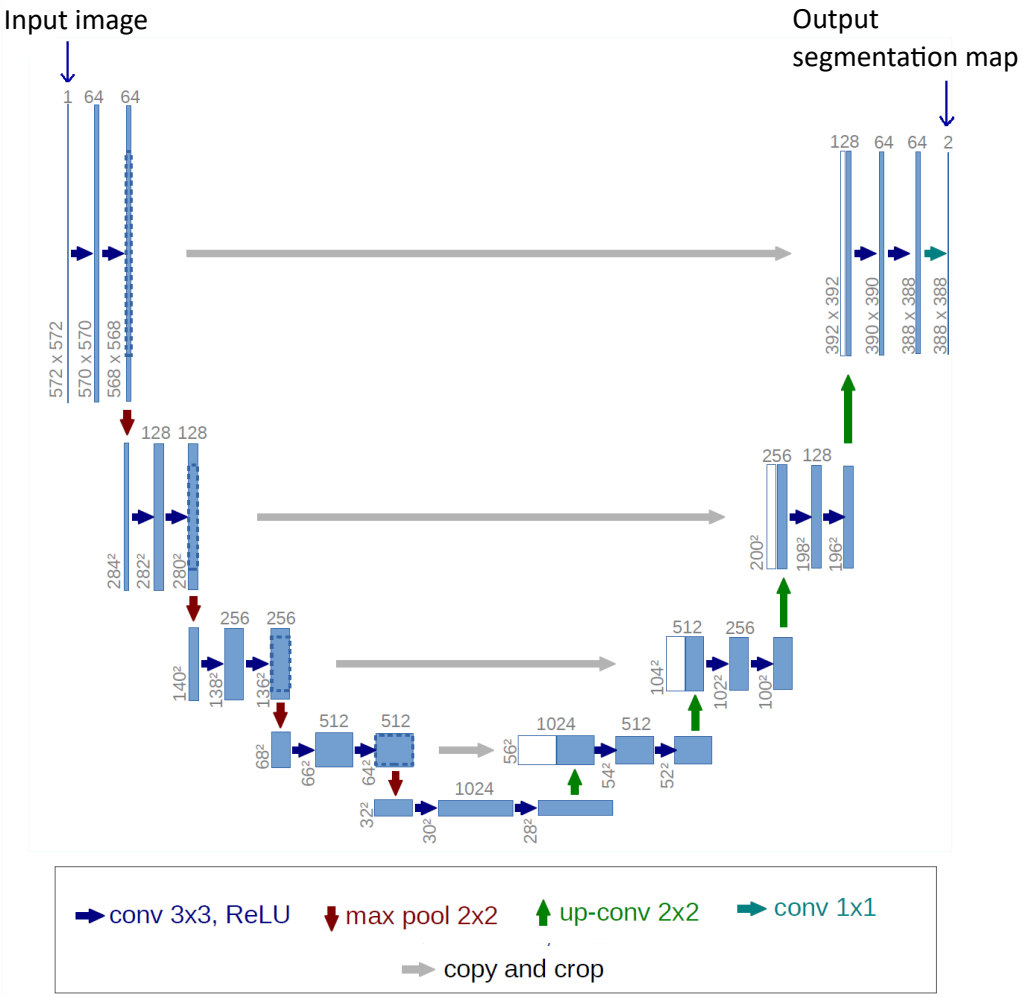


Figure 3.9: Illustration of U-net arcitechture, with permission from Olaf Ronneberger [16]. Each box represents a feature map, with the corresponding number of channels above. The resolution of the example image is denoted vertically at the bottom left corner of the boxes. The blue arrows are convolutional layers with a 3x3 pixels kernel. The gray arrows are skip connections, and the white boxes are the copied feature maps from previous layers. Furthermore, red arrows represent max pooling layers, while the green arrows are upsampling layers both with kernel sizes of 2x2 pixels. Finally, the seablue arrow is a 1x1 convolutional layer, creating the output segmentation map.

Chapter 4

Experimental setup

4.1 The data

4.1.1 Background

In order to use patient data for the purpose of research in Norway, it is a prerequisite that projects are pre-approved by the Regional Committees for Medical and Health Research Ethics (REC) [51]. One must also apply for data from the Norwegian Patient Registry to the Norwegian Directorate of Health [52]. Furthermore, each patient has to give consent, that their medical records for treatment and follow up can be used for this specific study. All patients in this study gave written informed consent. This process is time consuming, and it may take years before the dataset is large enough to be used for research purposes. When working with Deep Learning and image analysis, a substantial sample size is often crucial [38] but mostly beneficial [53]. The more data the algorithm can train on, the bigger are the chances for the algorithm to become robust and produce accurate results.

From 2013 to 2016, the Department of Oncology, OUH, collected data from 93 anal cancer (AC) patients treated with radiotherapy or chemo-radiotherapy [3]. All 93 patients were above the age of 18 years old. Even though both sexes were represented, the majority of the patients were female. The dataset consists of both clinical factors, such as age,

sex and mass, but also data from medical scans generated throughout the course of treatment of anal cancer. These medical scans include a Dose Planning Computed Tomography (DPCT) and possibly several PET, CT and MRI sequences. The patients were injected with CT contrast for the DPCT scan while the PET/CT scans were performed without CT contrast but with ^{18}F-FDG tracers for the PET scan. The MR sequences were generated without any contrast medium. The images provided by the hospital come in a format called DICOM ('Digital Imaging and Communications in Medicine').

The DICOM-image sequence of DPCT also contained a number of structures provided by three different, experienced radiation oncologists [9], including a GTV of the primary tumor which is considered the ground truth delineation. Since the target GTV was based on the DPCT, only the image sequences generated close to the time of the DPCT were of interest for this Master's thesis. These image sequences are referred to as baseline image sequences. However, if available, medical image sequences generated about two weeks after the DPCT were also provided as part of the dataset. Furthermore, the clinical factors were not used in this master thesis, and have therefore not been processed, evaluated or included to the dataset.

4.1.2 Processing and quality assurance

Of the 93 AC patients, 7 AC patients were not included since they did not have images from all modalities or did not pass the quality assurance for this study. During the autumn 2018, the author worked with co-registration of the patient data at the Norwegian Radium Hospital, along with her colleague Maria Cabrol. This resulted in co-registered image data from 36 AC patients. The co-registering and processing of the 86 AC patients were performed in MICE software [54], which is specifically developed for analysis of medical images. The process of image registration consisted of cropping the images according to chosen cropping values, co-registering the different modalities, interpolating the resolutions and finally saving the matrices as MATLAB-data. This is described in greater detail in a term paper, written autumn 2018 [55].

The resulting images have voxel size of 1 x 1 x 3 mm^3, and the generated image sequences from all the modalities are co-registered.

Co-registering the image sequences involves transforming the data of a moving image sequence and maximizing the image overlap [56] according to a reference image sequence. The DPCT was used as the reference for all the other image sequences. Moreover, the image sequences from each patient were carefully examined and evaluated prior to the conversion to MATLAB-data. If the co-registration was not successful, the chosen cropping values were reevaluated in order to perhaps get a more accurate overlap. This resulted in a final dataset of 85 AC patients, as one of the patients had an incomplete DPCT image sequence. All of the remaining patients had DPCT, baseline PET and CT scans, but only 36 AC patients had baseline MRI scans available (see Table 5.6).

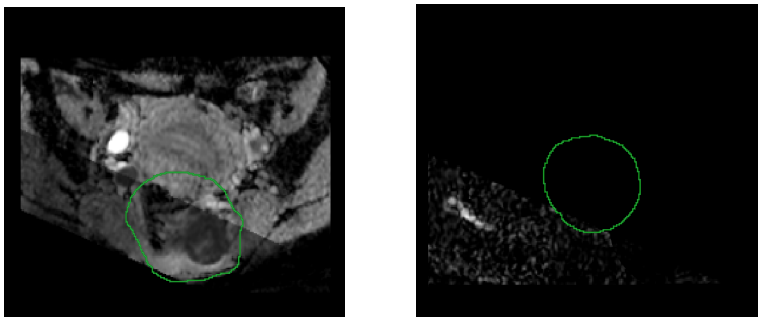
Table 4.1: Overview of the resulting dataset. The PET, CT and MRI images (columns three and four) refer to baseline images. Note that only 49.4 % of the slices in the dataset without MRI scans and 46.2 % of the dataset with MRI scans had target volumes (TV).

	Total	PET and CT	MR	Complete ADC maps
Patients	85	85	36	18
Slices	3492	3492	1501	764
Slices with TV	1726	1726	694	-

4.1.3 Content of the dataset

For the 36 AC patients that had MRI image sequences, two different ADC maps were created: 'ADC.mat' based on the three early b-values, *b0*, *b10* and *b20*, and 'Perf.mat' based on *b200*, *b400*, *b800* and *b1000* (as described in section 2.4). In addition, the MICE software also creates its own ADC map, named 'ADCsig.mat'. The latter was not used due to lack of information about how this ADC map was created. It was decided to be consistent when choosing b-values for making the ADC maps and not to customize this for each patient, which also was an option. This was to make the study and the results of this master thesis as reproducible as possible. The consequence of this is, of course, that the ADC maps might not always be as optimal as they might have been if one tailormade the ADC maps to each patient. In addition to the ADC maps, the DWIs for *b0*, *b10*, *b20*, *b40*, *b80*, *b160*, *b200*, *b400*, *b800*, *b1000*, *b1200* and *b1500* were included in the dataset.

Among the 36 AC patients with MRI image sequences yielding ADC maps, 14 patients had an incomplete ADC image sequence compared to the GTV image sequence. This means that the ADC image sequences lack image slices or that the generated image was not complete (see Figure 4.1). This occured either at the end or the beginning of the image sequence, where GTV delineations were present. This can be caused by how the DWIs were generated, the angle of the imaging (as a result of the positioning of the patient in the scanner) or a flaw in the co-registering [55].



(a) 'ADC.mat' slice number 48

(b) 'Perf.mat' slice number 20

Figure 4.1: Examples of discontinuities in the ADC maps of patient M007 where the GTV is present. The GTV is marked in green.

Included structures

The delineations of the affected lymph nodes and the Gross Tumor Volume (GTV) for the primary tumor, were included for all 85 AC patients. In most cases, these GTVs were distinct but for some patients the GTV of the primary tumor could include lymph nodes. In such cases, the masks of the lymph nodes were carefully evaluated and the delineations that were characterized as lymph nodes were either removed or saved as separate lymph node GTVs.

The reason for including the delineations of the lymph nodes was to possibly use these to reduce false positives. Lymph nodes share similar characteristics with malignant tumors, such as the metabolic activity and their spherical shape [29]. In a PET image sequence the lymph nodes will light up in the same manner as the primary tumor, and might mislead the algorithm into thinking that these are malignant tumors [29].

In addition, the Clinical Target Volume (CTV) for the primary tumor was included, as this could be useful for evaluating the accuracy of the autodelineation.

4.1.4 Sample size and class balance

The resulting dataset for this study consisted of 85 AC patients. Moreover, the number of voxels included in a delineation, representing the cancerous tissue was considerably smaller than the number of voxels representing healthy tissue in the image. The skewed label distribution is a typical challenge for real-world applications and is often referred to as a class imbalance [38]. Except for cropping of each image sequence for the purpose of co-registering, no actions were taken to reduces the class imbalance.

4.2 Finalized dataset

The finalized dataset is represented in 85 folders, one for each patient. Medical data contain sensitive, personal information about patients and it is therefore important to anonymize the data. The initials and corresponding date of birth have been replaced by a patient identification 'M***' in the folder names. Within a patient folder the image sequences in MATLAB-data formats are distributed among (i) a 'Base' folder, containing the baseline image sequences, (ii) a 'Mid' folder, containing the image sequences generated about two weeks post DPCT, and (iii) a 'ROI' folder containing the selected GTV and CTV structures. The specific MICE-code that was used to generate the MATLAB-data sequences for each patient is also available in the associated patient folder. In addition, a .txt file containing the chosen cropping regions from the original DICOM images (as explained in the term paper [55]) is also included.

4.2.1 Setup for the dataset

Hierarchical Data Format version 5 (HDF5) is a mechanism for storing and organizing large and complex amounts of data [57]. Some advantages

of this data format is that it can scale up to exabytes, and has sub-setting capabilities that consequently makes it very fast [57]. The HDF5-file is also more portable compared to storing data in a directory on the hardware. HDF5-files can consist of datasets, groups and/or attributes [57]. The dataset for this thesis was saved as a HDF5-file, where the first hierarchical level consists of a training group, a validation group and a testing group. Each group consists of a dataset (named 'dat'), information about patient ids and the corresponding target volume sequences (named 'mask'), as presented in Figure 4.2.

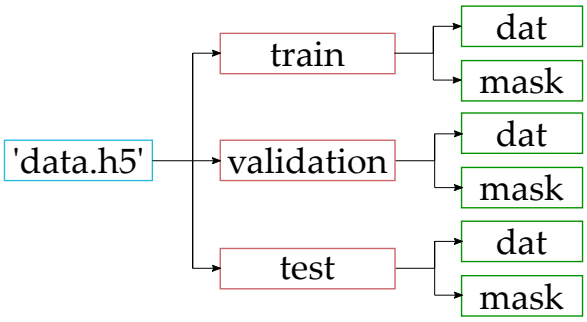


Figure 4.2: Illustration of the structure of the HDF5 file used in this thesis. The box with blue outline (far left) is the HDF5 file. Boxes with orange outline are groups, while boxes with green outline are datasets.

In order to collect all the patient image sequences in a HDF5-file, the patient image sequences must have the same resolution and number of channels. The resolutions in x- and y-direction across the patient image sequences were not the same, and neither was the number of channels. Therefore, padding (see section 3.3.1) was added to all images that had smaller resolutions than the largest resolution observed in the dataset. This resulted in resolution of 236 x 236 voxels for all images in the dataset.

Each slice in the image sequence can potentially consist of five channels, representing the corresponding slice from each modality, as shown in Figure 4.3. Furthermore, two HDF5 files were created: one for the 85 patients with DPCT, PET and CT, and another for the 36 patients whom in addition had the two ADC maps and T2W images. Table 4.2 provides explanations of the different channels.

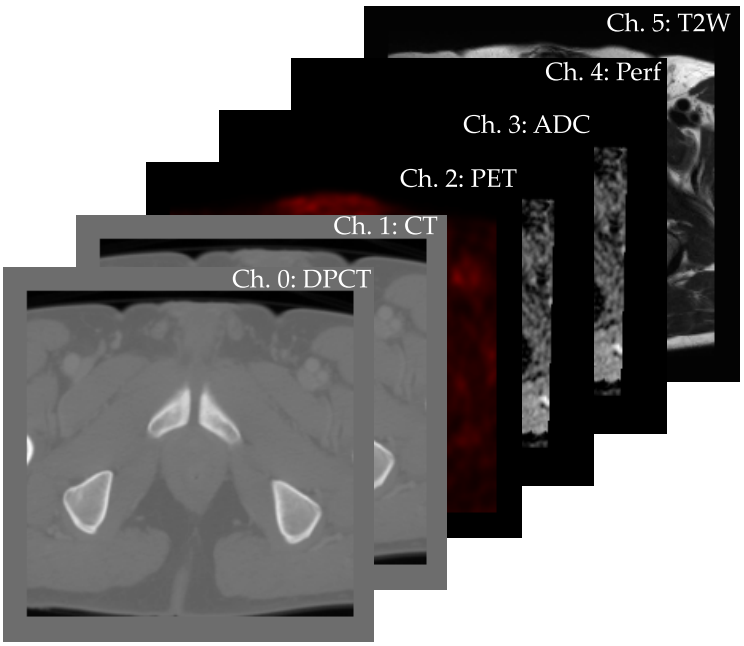


Figure 4.3: Illustration of the channels available in each image slice, according to the order in the finalized HDF5-file. Explanations of the channels are provided in Table 4.2. The images representing the different modalities are only for illustrative purpose. Note that 36 out of 85 patients did not have ADC, Perf or T2W channels.

Table 4.2: Explanation of the channels used for the experiments in this project.

Channel	Explanation
$DPCT$	CT image sequence generated for the purpose of dose planning
PET	PET image sequence generated by PET/CT examination
CT	CT image sequence generated by PET/CT examination
ADC	ADC map based on b-values: *b0*, *b10* and *b20*
$Perf$	ADC map based on b-values: *b200*, *b400*, *b800* and *b1000*
$T2W$	T2-weighted image sequence

4.3 Software and computer

For co-registering the images, the software MICE Toolkit 1.0.7 [54] was mainly used, in addition to MATLAB (version R2014b, The Mathworks Inc., Natick, MA) which was used for some of the code in the co-registering pipeline. Furthermore, the environment Spyder [58] version 3.3.2 under the terms of the MIT License was used for making the HDF5 files, preprocessing and inspecting the data. Python 3.6.4 (64-bit for Windows 10) was used for running the autodelineation program in this thesis. Moreover, TensorFlow r1.12 with GPU support was used for the neural network architecture. The computer used for the majority of the project had one GPU available, NVIDIA GeForce GTX 1080 Ti, which was used for running the experiments.

Chapter 5

Preparations and Experiments

5.1 Preprocessing

Before running any machine learning algorithm, it is crucial to preprocess the data in order to quality assure the model inputs. In most deep learning applications, preprocessing of the images is not necessary, as processing of the data usually is a part of the network pipeline. However, inspecting and evaluating the data thoroughly might be crucial to get a better understanding of the results of the CNN model. Moreover, medical data is highly inconsistent. Processing the data in order to increase the consistency and decrease unwanted noise, can aid the CNN model into learning the relevant information in the data.

5.1.1 Correction of T2W images

When inspecting the T2W image sequences of the AC patients it became apparent that some of the images were not in the same voxel value range as the others. Some of the T2W images appeared substantially darker than others. Furthermore, a shift in the voxel value ranges was discovered and the patients were divided into two groups: group A with the first seven patients in the dataset, and group B with the remaining 29 patients. The observed shift could be due to a change of MRI scanner or change in the routines of the oncologists. The voxel value ranges, in addition to the observed shift, are presented in Figure 5.1.

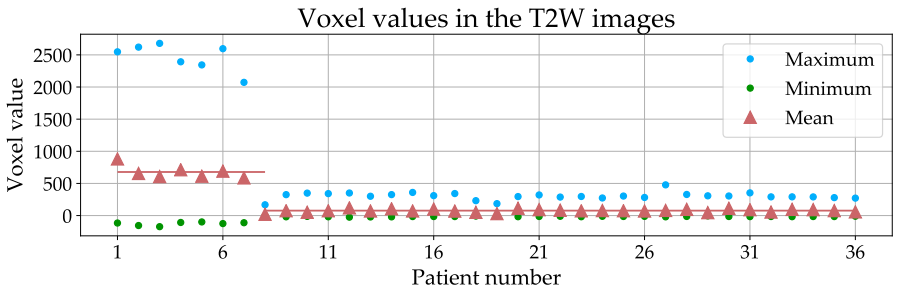


Figure 5.1: Maximum, minimum and mean voxel values for the T2W image sequences illustrated in blue, green and pink, respectively. The x-axis represents the AC patients, while the y-axis represents the voxel values. A conversion from patient number to the patient IDs is provided in Appendix A. Moreover, the pink lines are the mean of the voxel values within each of the two patient groups, A and B. Group A consists of the first seven patients, while group B is the remaining 29 patients.

The voxel value shift observed in Figure 5.1 was corrected for in order to obtain a more consistent dataset and the following equation was used:

$$X_{new} = X\frac{\mu_B}{\mu_A} \tag{5.1}$$

where X_{new} is the new, corrected image, X is the original image, μ_B is the mean of group B and μ_A is the mean of group A. As a result, the mean voxel value of group A will be equal to the mean of the voxel values in group B, as shown in Figure 5.2:

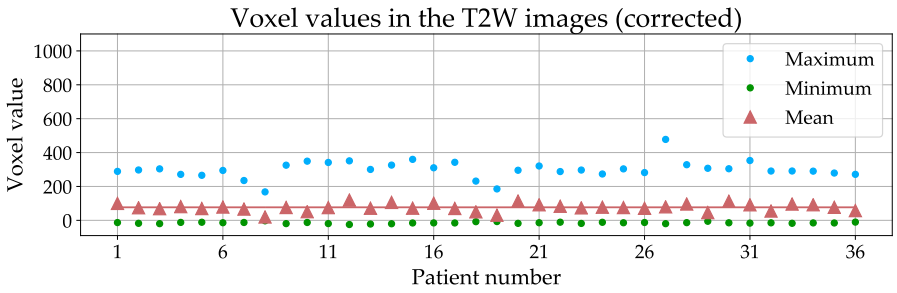
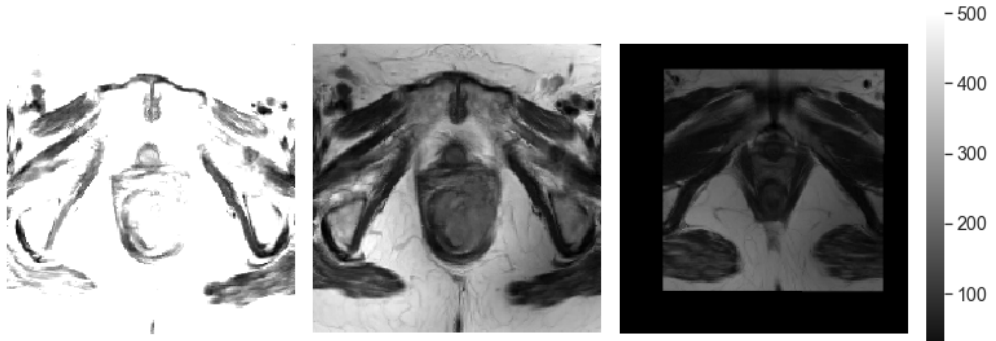


Figure 5.2: Corrected maximum, minimum and mean values for the T2W image sequences illustrated in blue, green and pink, respectively. The x-axis represents the AC patients, while the y-axis represents the voxel values. A conversion from patient number to the patient IDs is provided in Appendix A. Moreover, the pink line is the mean of all the voxel values

When displaying the images, one can also observe that the brightness of the corrected images from group A was more similar to the brightness of the T2W images from group B. An example of this is presented in Figure 5.3, where the uncorrected T2W image of AC patient 'M003' in group A was much brighter than the image of AC patient 'M110' in group B.



(a) 'M003', original **(b)** 'M003', corrected **(c)** 'M110', original

Figure 5.3: Correction of voxel value range in the T2W images of the anorectal region in two different AC patients ('M003' and 'M110', slice 21). The color bar provided to the right apply to all three images.

5.2 Data cleaning

The removal of data points which are considered outliers may be one of the most important process to clean the data for a machine learning task [38]. Such data points could have a negative impact on a classification task by misguiding the network model. Consequently, numerous image slices in this dataset were removed.

5.2.1 Discontinuities in MRI slices

As explained in section 4.1.3 in chapter 4, some of the T2W and ADC images were left with a diagonal discontinuity after co-registration in MICE. Table 5.1 provides an overview of the number of image slices which contained discontinuities, as shown in Figure 4.1, and image slices that were completely black (containing only zeroes).

Table 5.1: Number of image slices that were all-zeroes (first column), all-zeroes in the delineated region (second column), image slices evaluated with substantial discontinuity but not all-zeroes (third column) and the number of image slices with discontinuity in the target volume (TV) (last column). 'Dis.' stands for discontinuity.

	All-zeroes	All-zeroes in TV	Dis. in image	Dis. in TV
$T2W$	81	3	71	3
ADC	450	52	164	14

The evaluation of whether an image is substantially discontinued or not was made by looking for a patch of 70 x 70 voxels that only contains zeros. All image slices where the target volume only contained zeroes, either in the ADC map, the T2W image or both, were removed (second column in Table 5.1). In addition, all T2W image slices containing a patch of 70 x 70 voxels or bigger with only zeroes were removed. This includes slices in both the first and the third column of Table 5.1, for T2W.

Consequently, the MRI dataset was reduced by 204 slices or about 13.6 % of the total number of slices. Note that the sum of removed slices does not directly match the numbers provided in the columns in Table 5.1, since some of the slices can be represented in more than one column. An image slice could, for instance, both have only zero voxel values in the GTV area, while also having a discontinuity in the T2W image.

5.2.2 Slices without delineation

Originally, 49 % of the slices in the PET/CT dataset of 85 patients and 46 % of the slices in the MRI dataset of 35 patients did not contain any target volume delineations provided by the oncologists. The Pareto Principle [59] gives that, in general, 20 % of something always are responsible for 80 % of the results, and vise versa. In order to decrease the computational cost of the network and make each experiment less time consuming, it was decided to remove 80 % of the image slices not containing an oncologist's target volume delineation.

The 80 % removed should have been the upper 80 % of the image slices without target delineation which had the least effect on the model [59]. This would require a more complicated exploration and, consequently, the 80 % were chosen randomly, where each slice had a probability of 80 % of being removed. Thus, 645 image slices were removed, or about 43 % of the dataset.

5.3 Image augmentation

Ronneberger et al. [16] used shift, rotation and gray value variation of images for data augmentation to boost the network performance, during training [16]. Random, excessive elastic deformations were found to work well [16].

In this thesis, the elastic deformation used is based on the one presented by Simrad et al. [46]. By visual inspection of the outcome for some randomly chosen patient image sequences, the following parameters were chosen for elastic image augmentation:

Table 5.2: Parameter values chosen for elastic deformation of images. The first column is alpha, the second column is sigma and the last column is the alpha affine, as described in section 3.3.4, chapter 3.

α	σ	α affine
80	25	15

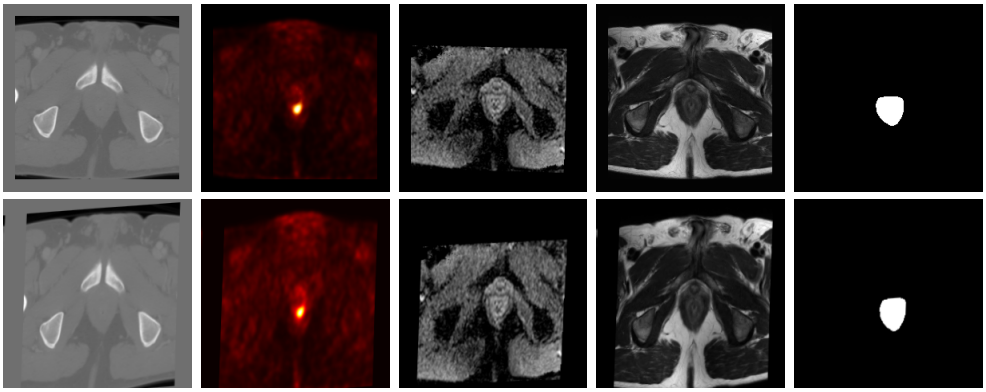


Figure 5.4: Performed elastic deformations on medical images from an AC patient ('M027', slice 27) with $alpha = 80$, $sigma = 25$ and $alpha\ affine = 15$. These parameters control the intensity of the deformation, as described in Section 3.3.4, Chapter 3. The medical images presented are, from the left: CT, PET, ADC, T2W and the provided target volume mask. The upper row gives the original images, whereas the second row show the augmented images.

When choosing the parameter values in Table 5.2, the aim was to generate augmented images resembling the existing images in the dataset, with only minor deformations. By doing so, the new, augmented images ought to represent fictional, new patient data, with the main purpose of increasing the training data for the network.

For each patient, about 35 % of the slices were randomly chosen for deformation. In addition, horizontal flip was applied on another subset of about 35 % of the slices for each patient. This subset was chosen independently of the slices chosen for elastic deformation. Both subsets were randomly chosen without replacement. Figure 5.4 shows the effect of applying elastic deformation to a given image slice.

Finally, the augmented data was only added to the training set, almost doubling the number of image slices. In order to retain information about which patient the slice stemmed from and to indicate that the image was a product of an augmentation process, these slices got a patient ID 'M***_aug'. The image augmentation was applied after both padding and removing of slices.

5.4 Train, validation and test split

Before running any experiments, the data was split into training, validation and test sets. The split was stratified, according to the tumor volume. The total target volume for each patient was determined based on the number of voxels included in the delineation provided. First, the total target volume for each patient was sorted in a list. Second, the patients were divided into two groups: those with the largest number of target volume voxels and another with the smallest number of target volume voxels. Thereafter, 70 % of each group were randomly placed in the training subset, 50 % of the remaining patients from each group were placed in the validation subset and the last patients from both groups were placed in the test subset. The 70-15-15 split of the dataset was evaluated based on the size of the dataset. To the author's knowledge, there are no other recommended guidelines for choosing the split as long as each subset of the total dataset are as representative as possible [38].

The motivation behind this strategy was to get a stratified split, based on the delineated volumes, while retaining some control of how the split was made. The subset should then contain both patients with large target volumes and small target volumes. Table 5.3 shows the resulting dataset after image augmentation.

Table 5.3: Number of patients and image slices in the datasets after image augmentation, data cleaning in the training, validation and test set. 'Patients' is the number of unique patient IDs, 'org' is the number of original instances and 'aug' is the additional instances as a result of the image augmentation. Note that all patients in the datasets are represented in the augmented images.

	PET/CT dataset				MRI dataset			
	Patients		*Image slices*		*Patients*		*Image slices*	
	org	aug	org	aug	org	aug	org	aug
Train	59	85	1434	1354	27	36	459	410
Validation	12	-	279	-	4	-	69	-
Test	14	-	316	-	5	-	99	-
Total	**170**		**3383**		**72**		**1037**	

5.5 Windowing

Based on the results of Moe [18], windowing of CT images was considered important for the performance of the delineation experiments. Therefore, an inspection of the most optimal windowing values was performed. Table 5.4 provides the maximum, minimum, mean, median and mode for the Hounsfield values of the voxels representing the delineated areas. DPCT and CT were inspected separately to detect any inter-channel variations. The metrics are based on all delineated voxels in the DPCT and CT image sequences for all 85 patients. The mode is the most common value represented in the set. If there by chance was more than one value with equal, highest frequency, the smallest HU value was chosen.

Table 5.4: Statistics of Hounsfield values in the delineated areas for DPCT and CT images in the dataset. STD stands for standard deviation.

Channel	Min	Max	STD	Mean	Median	Mode
$DPCT$	-993.0	3009	143.9	29.40	57.73	70.00
$PETCT$	-1034	834.7	108.8	0.2380	25.46	32.00

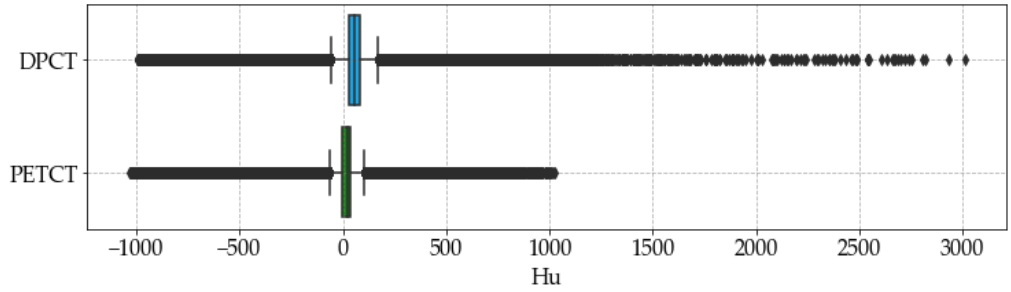


Figure 5.5: Boxplot illustrating the Hounsfield values for the delineated regions in the DPCT and CT images. The boxes in blue and green represent the area in which 50 % of the data, for the respective channel, is present. The whiskers show the highest and the lowest Hounsfield values, while the diamonds outside of the whiskers represent automatically detected outliers.

In Table 5.4 one can observe that since the minima and maxima are far away from the mean, median and mode, extreme values are present, most

probably due to artifacts in the images. This can also be observed in the boxplot in Figure 5.5. The center values for the windowing of each of the channels, was therefore chosen to be the mode values presented in Table 5.4. Furthermore, the widths were determined based on the double of the standard deviations, rounded up in order to include some extra Hounsfield values. This led to narrow windowing widths, which is recommended for areas of soft tissue [27].

Consequently, the windowing options provided in Table 5.5 were used. Typical windowing options for soft tissue in the abdomen is a center of 50 HU and a width of 400 HU [27], but should ideally be evaluated for each CT scan.

Table 5.5: The resulting windowing options chosen for the experiments, given in Hounsfield units (HU).

Channel	Center	Width
$DPCT$	70	300
$PETCT$	32	220

5.6 Baseline performance

It is possible to calculate the mean Dice performance expected if the deep learning model had simply used an average target volume mask based on all the 85 patients as the predicted tumour. This performance is the minimum Dice performance, the baseline performance, one should expect from the network in order for it to perform better than simply random guessing the mask.

Moreover, there are two options for generating an average target volume mask: one based on all image slices available, and another where the image slices without target volume delineations are excluded. These options where inspected separately and are presented in Figure 5.6.

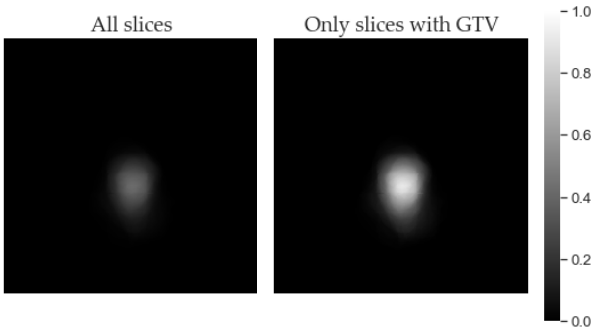


Figure 5.6: Resulting probability map for the average GTV mask based on all image slices (left) and only image slices with an oncologist' GTV (right). The color bar provided to the right apply to both images. Note that none of the pixel values in the left probability map are equal to or larger than 0.5.

The values in the probability map generated based on all image slices (to the left in Figure 5.6) did not exceed 0.5, and choosing a threshold value for creating an average GTV mask was not intuitive. Nonetheless, the probability map generated based on only the slices with an oncologist' GTV (to the right in Figure 5.6) had several voxel values equal to or larger than 0.5. This can be interpreted as if there is at least a 50 % chance for the voxel to be a part of the delineation.

Thus, an average GTV mask was generated by creating a binary image with a threshold value of 0.5, based on the probability map for image slices excluding slices without GTV delineations. This resulted in the average GTV mask shown in Figure 5.7.

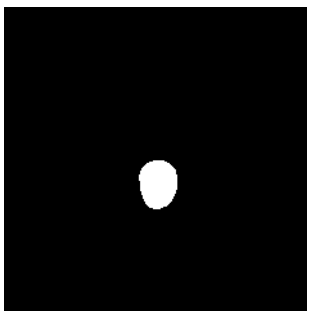


Figure 5.7: Average GTV mask based only on the slices containing GTV from all 85 patients, with a threshold value of 0.5.

This baseline GTV mask was used to evaluate the baseline performance by calculating the Dice coefficient between the average GTV mask and the

actual GTV for each image slice. This was calculated for both the PET/CT dataset and the MRI dataset (presented in Table 5.3). The calculated baseline Dice performances are provided in Table 5.6.

Table 5.6: Baseline performances using an average GTV mask for all slices.

Dataset	All slices	Only slices with GTV
PET/CT	0.347	0.701
MRI	0.324	0.701

5.7 The Code

The code used in this thesis is based on the code written by Yngve Mardal Moe for his MSc *Deep learning for the automatic delineation of tumours from PET/CT images.* (2019) [18]. Modifications have been made to fit the model according to the dataset and the experiments examined in this thesis. These modifications include new windowing options, a dropout activation, image augmentation and data cleaning outside of the pipeline proposed by Moe [18]. Note that, preprocessing was conducted prior to making the HDF5 file, used as input to the pipeline for autodelineation. Windowing, however, was performed as a part of the pipeline for autodelineation. For reproducibility, the programs used in this project are available on the Github repository: https://github.com/christinekaush/ANCARAD_autodel.

5.7.1 Network architecture

The model architecture used in this thesis is a basic U-Net architecture, shown in Table 5.7, with a total of 27 layers (disregarding the input image layer). As input, the model receives the medical images that we want to analyze, along with the target binary segmentation mask (the ground truth). The number of channels per input varied depending on the number of modalities used for each experiment. Each layer uses ReLU as activation function, except for the convolutional layer 8 where Dropout is used for activation to the next layer with a keep probability of 50 %. This is a modification relative to the original framework by Moe [18].

Table 5.7: The U-Net architecture used for the experiments in this project. All layers use the ReLU activation function except for *Conv 8*, which has a Dropout activation.

Layer	Type	Input	No. output channels
$Conv\ 1$	Convolutional	Input image	64
$Conv\ 2$	Convolutional	Conv 1	64
$MaxPool\ 1$	Max Pooling	Conv 2	64
$Conv\ 3$	Convolutional	MaxPool 1	128
$Conv\ 4$	Convolutional	Conv 2	128
$MaxPool\ 2$	Max Pooling	Conv 4	128
$Conv\ 5$	Convolutional	MaxPool 2	256
$Conv\ 6$	Convolutional	Conv 5	256
$MaxPool\ 3$	Max Pooling	Conv 6	256
$Conv\ 7$	Convolutional	MaxPool 3	512
$Conv\ 8$*	Convolutional	Conv 7	512
$MaxPool\ 4$	Max Pooling	Conv 8	512
$Conv\ 9$	Convolutional	MaxPool 4	1024
$Conv\ 10$	Convolutional	Conv 9	1024
$UpConv\ 1$	Upconvolutional	Conv 10	512
$Conv\ 11$	Convolutional	UpConv 1	512
$Conv\ 12$	Convolutional	Conv 11	512
$UpConv\ 2$	Upconvolutional	Conv 12	256
$Conv\ 13$	Convolutional	UpConv 2	256
$Conv\ 14$	Convolutional	Conv 13	256
$UpConv\ 3$	Upconvolutional	Conv 14	128
$Conv\ 15$	Convolutional	UpConv 3	128
$Conv\ 16$	Convolutional	Conv 15	128
$UpConv\ 4$	Upconvolutional	Conv 16	64
$Conv\ 17$	Convolutional	UpConv 4	64
$Conv\ 18$	Convolutional	Conv 17	64
$Conv\ 19$	Convolutional	Conv 18	1

* Dropout as activation function

5.8 Assumptions

For these experiments it is assumed that a physician has already performed a physical examination of the patient, and has a solid intuition of the region in which the cancerous tumor is located. Therefore, most of the image slices not containing delineations were removed.

The delineations provided by the oncologist are assumed to be correct, and are therefore used as a ground truth. Furthermore, it is assumed that the scans of the different imaging modalities has been conducted in more or less the same time period. Ergo, the cancerous tumor has not changed significantly in-between the scans and is located in more or less the same area. The co-registering of the imaging modalities is also assumed to be as optimal as possible.

It is assumed that the most optimal choices for the network parameters are independent of the imaging modalities used for the purpose of autodelineation. In the attempt to conduct a consistent and systematic investigation, all experiments comparing the modalities were run on the MRI dataset (see Table 5.3), since not all of the patient in the PET/CT dataset had image sequences from MR scans.

5.9 Experiments

To find the imaging modality, or the combination of imaging modalities, which is best suited for the purpose of sentiment image segmentation, experiments with the following sets given in Table 5.8 of imaging modalities were conducted.

Table 5.8: List of imaging modalities used for each experiment conducted in this project. Explanation of the different channels is provided in Table 4.2.

	Channel(s)
1.	PET
2.	$DPCT$
3.	CT
4.	ADC
5.	$T2W$
6.	$Perf$
7.	PET, CT
8.	$PET, DPCT$
9.	$T2W, ADC$
10.	$T2W, ADC, Perf$

5.9.1 Effect of regularization and data cleaning

In addition, the effects of image augmentation, Dropout activation and the removal of image slices were evaluated. This was conducted by running the highest performing experiment four times, removing each of the steps one at a time in addition to one experiment excluding all of the steps, as presented in Table 5.9. Consequently, the direct effects of the steps can be removed, and how much they impact the resulting Dice performance can be evaluated.

Table 5.9: List of additional experiments for inspecting the effect of regularization and data cleaning.

	Additional experiments
11.	*Excluding image augmentation*
12.	*Excluding Dropout activation*
13.	*Excluding data cleaning*
14.	*Excluding all regularization and Dropout*

5.10 Set-up

The data was prepared as described in Section 5.1. All experiments mentioned in Tables 5.8 and 5.9 were run with the setup provided in Table 5.10.

Table 5.10: Common setup for the experiments for evaluation of imaging modality for the purpose of autodelineation. The optimizer is presented along with the chosen learning rate and batch sizes are for the training, validation and test set respectively.

Optimizer	Adam, 0.0001
Loss	F1
Batch size	[16, 16, 16]
Iterations	5000
Dataset	MRI dataset

Chapter 6

Results

6.1 Model performance

The effects of the imaging modalities on the autodelineation are described in this chapter. In total, 14 experiments described in Tables 5.8 and 5.9) were run. All results provided are based on the validation set and all experiments were run exclusively on the MRI dataset and with the setup provided in Table 5.10.

6.2 Effect of input channels

The Dice performances and the standard deviations for each experiment in Table 6.1 is the mean over all image slices in the validation set. These mean Dice performances range from good (0.6-0.8) to excellent (> 0.8) as defined by Gudi et al. [8]. All except the poorest performing experiment, performed better than the calculated baseline performance, of 0.701 (Table 5.6), for this dataset (considering that about 80 % of the slices without target volume was excluded).

The combination of modalities giving the poorest performing model was the 'Perf' ADC map with a Dice performance of 0.676 (Table 6.1), which is lower than the baseline Dice performance. Second to that is the 'ADC' ADC map with a Dice performance of 0.748.

Table 6.1: Mean Dice performances with corresponding standard deviation for experiments run on the validation set, with the setup provided in Table 5.10. Explanation of the different channels is provided in Table 4.2.

Channel(s)	Windowing	Dice
PET	-	0.867 ± 0.166
$DPCT$	c70 w300	0.795 ± 0.186
CT	c32 w220	0.877 ± 0.168
ADC	-	0.748 ± 0.207
$T2W$	-	0.861 ± 0.177
$Perf$	-	0.676 ± 0.244
PET, CT	c32 w220	0.885 ± 0.164
$PET, DPCT$	c70 w300	0.885 ± 0.176
$T2W, ADC$	-	0.842 ± 0.195
$T2W, ADC, Perf$	-	0.780 ± 0.192

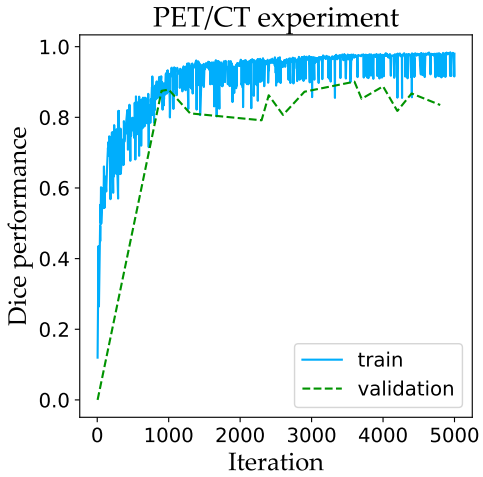


Figure 6.1: Training and validation curves for the PET/CT experiment for 5000 iterations. The blue line is the training curve, while the green, dashed line is the validation curve. The x-axis represents the iterations and the y-axis represents the Dice performance.

The experiment using PET and DPCT as imaging modalities performed equally well as the experiment with PET and CT when comparing the Dice

performances and considering three significant figures. Nevertheless, the standard deviation was somewhat lower when using PET and CT. The imaging modality combination which gave the best performance for the purpose of autodelineation was therefore concluded to be PET and CT (see Table 6.1). The training and validation curves for the PET and CT experiment is presented in Figure 6.1. Moreover, it should be noted that the experiments using the modalities PET, CT and T2W individually, also generated excellent performances (> 0.85 in Dice).

6.2.1 Effect of regularization and data cleaning

The proposed model network and the procedures included were investigated further. Four additional experiments were conducted to evaluate the effect of the regularization and data cleaning on the experiments: one excluding image augmentation, one excluding Dropout activation, another only excluding data cleaning and lastly, one experiment excluding all regularization or data cleaning. These experiments were based on the setup of the PET/CT experiment (experiment seven in Table 5.8), with modifications as explained in Section 5.9.1.

Table 6.2: Mean Dice performances and corresponding standard deviations per slice of experiments inspecting the effect of regularization and data cleaning. The change in percentage is the change in Dice performance relative to the original PET/CT experiment (first row). When removing the Dropout activation, a ReLU activation was used as a replacement.

	Dice	Percentage change
Original PET/CT experiment	0.885 ± 0.164	-
Excluding augmentation	0.747 ± 0.160	- 15.6 %
Excluding Dropout activation	0.874 ± 0.170	- 1.24 %
Excluding data cleaning	0.495 ± 0.440	- 50.3 %
Excluding all regularization and data cleaning	0.185 ± 0.328	- 79.1 %

Data cleaning refers to the removal of 80 % of the image slices without a oncologist' delineation, as described in section 5.2.2. The mean Dice performances per slice, with the corresponding standard deviations for these experiments are presented in Table 6.2. In addition, the change in Dice coefficient, relative to the original PET/CT experiment is also presented.

Table 6.2 shows that excluding data cleaning had the largest impact, causing a decrease in Dice performance of 50.3 %. Yet, the achieved Dice performance was above the baseline performance, of 0.324, for this dataset. Furthermore, excluding the Dropout activation decreased the Dice performance the least, by only a percentage change of 1.24 %.

Training and validation curves
By inspecting the training and validation curves in Figure 6.2, one can observe that the distance between the training and validation curves increased when the data augmentation was excluded. Hence, data augmentation decreased overfitting for this network.

However, the last experiment in Table 6.2, excluding data cleaning and any of the proposed regularization techniques, gave the largest decrease (79.1 %) in Dice performance. The Dice performance for this experiment is 0.139 below the baseline Dice performance of 0.324, for the MRI dataset with all slices included. The standard deviation of the experiment was larger than the Dice performance, which might indicate that the model is not learning any patterns sufficiently.

In addition, the training and validation curves for the experiment excluding all regularization and data cleaning in Figure 6.2, shows that the model performed poorly on the validation set and was more overfitted compared to the experiment where only data cleaning was excluded.

The training and validation curves for the experiment excluding data cleaning (bottom left in Figure 6.2) revealed that the model network appears to have difficulty in identifying relevant patterns in the images. Also here, the standard deviation for the experiment was large relative to the corresponding Dice performance.

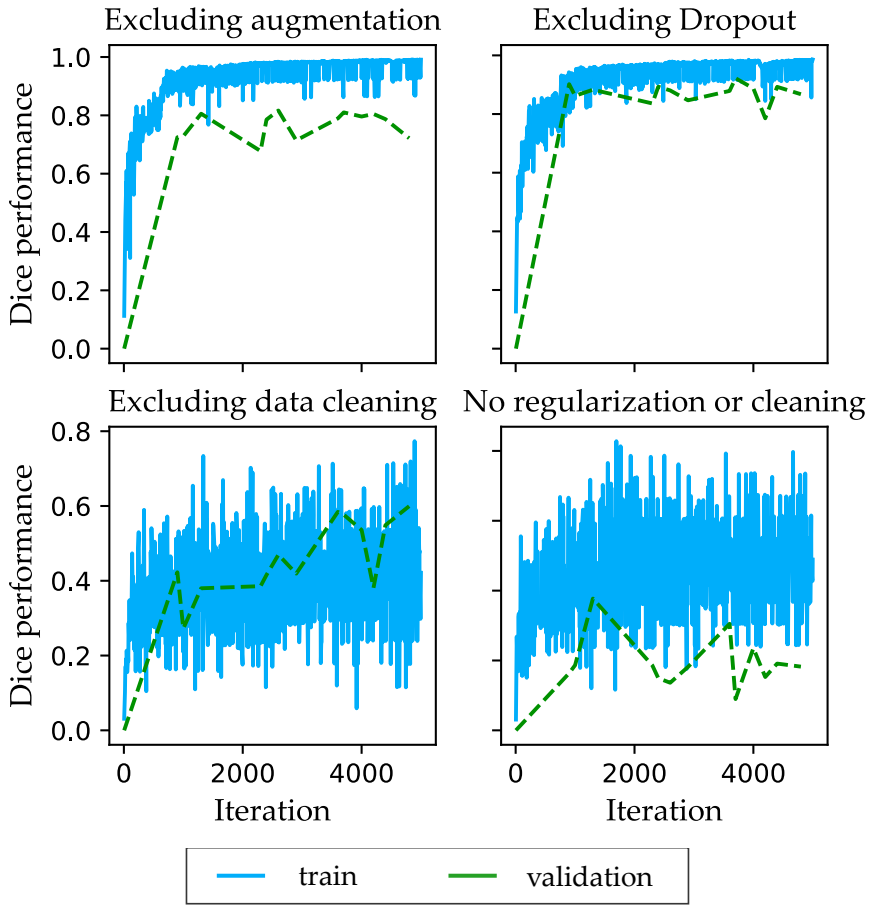


Figure 6.2: Training and validation curves for experiments excluding image augmentation, Dropout activation, data cleaning and all regularization and data cleaning, on the PET/CT runs (experiment seven in Table 5.8). The blue lines are the training curves, while the green, dashed lines are the validation curves. The x-axis represents the iterations and the y-axis represents the Dice performance.

6.2.2 Effect of an increased dataset

The PET/CT experiment (experiment seven in Table 5.8) could also be conducted with the PET/CT dataset (presented in Table 5.3). This opened up the possibility of inspecting the effect of an increased dataset as well. Table 6.3 provides the Dice performances for two versions of the PET/CT experiment: the original PET/CT experiment on the MRI dataset and the PET/CT experiment on the PET/CT dataset.

Table 6.3: Mean Dice performances and corresponding standard deviations per slice of experiments to examine the effect of an increased dataset. The change in percentage is the change in Dice performance relative to the original PET/CT experiment run on the MRI dataset (first row). The second row is the PET/CT experiment conducted on the dataset run on the PET/CT dataset. The datasets are presented in Table 5.3.

Dataset	Dice	Percentage change
MRI	0.885 ± 0.164	-
PET/CT	0.854 ± 0.169	- 3.50 %

In addition, the PET/CT experiment without any regularization or data cleaning was conducted on the PET/CT dataset. This resulted in a Dice performance of 0.664, which was 72.1 % larger than the corresponding model performance obtained using the MRI dataset without any regularization or data cleaning as presented in Table 6.4.

Table 6.4: Mean Dice performances and corresponding standard deviations per slice of experiments to examine the effect of an increased dataset. The change in percentage is the change in Dice performance relative to the PET/CT experiment excluding all regularization and data cleaning (run on the MRI dataset). The second row is the PET/CT experiment excluding all regularization and data cleaning on the dataset run on the PET/CT dataset. The datasets are presented in Table 5.3.

Dataset	Dice	Percentage change
MRI	0.185 ± 0.328	-
PET/CT	0.664 ± 0.378	+ 72.1 %

6.3 Inspection of the predicted delineations

The Dice performances provided in Table 6.1 were based on the mean Dice performance per slice across the validation set. It is, however, useful to inspect how the model performs on each image slice in order to obtain a better understanding of how the model delineates the tumor relative to the

oncologist' target volume. The following section will therefore present the resulting output images obtained using the PET/CT model (experiment seven in Table 5.8) on the validation and test set.

6.3.1 The validation set

Figure 6.3 present in sum 12 image slices from three different patients. The image slices presented are chosen from a set of 69 image slices, and are aimed to be as representative of the total validation set as possible. The base images are fused PET and CT images, where the PET signal is presented using colormap 'hot', ranging from black (no radiotracer signal) to yellow (maximum radiotracer signal). The target volume delineations provided by the oncologists are presented in green and the predicted delineations are marked in blue.

In most cases, the predicted delineation seem to match the provided target volume, resulting in a high Dice performance. Some examples are ('M007', slice 21) and ('M098', slice 27). However, the model fails on the image slice from patient 'M007', slice 3 (upper left in Figure 6.3), which has no oncologist' target volume delineation. Note that the model did not delineate the lymph node lightning up at the top in ('M007', slice 31).

It should be noted that all of the automatic delineations proposed for the validation set always include the region of the anal canal, and are more or less in the center of the image. However, the delineations do not resemble the baseline GTV, but rather seem to be customized according to the patient.

Table 6.5: Mean Dice performances and corresponding standard deviations per patient of the PET/CT experiment (run on the MRI dataset, Table 5.3) in the validation set.

Patients	Dice
'M007'	0.922 ± 0.236
'M064'	0.862 ± 0.111
'M066'	0.916 ± 0.044
'M098'	0.933 ± 0.065

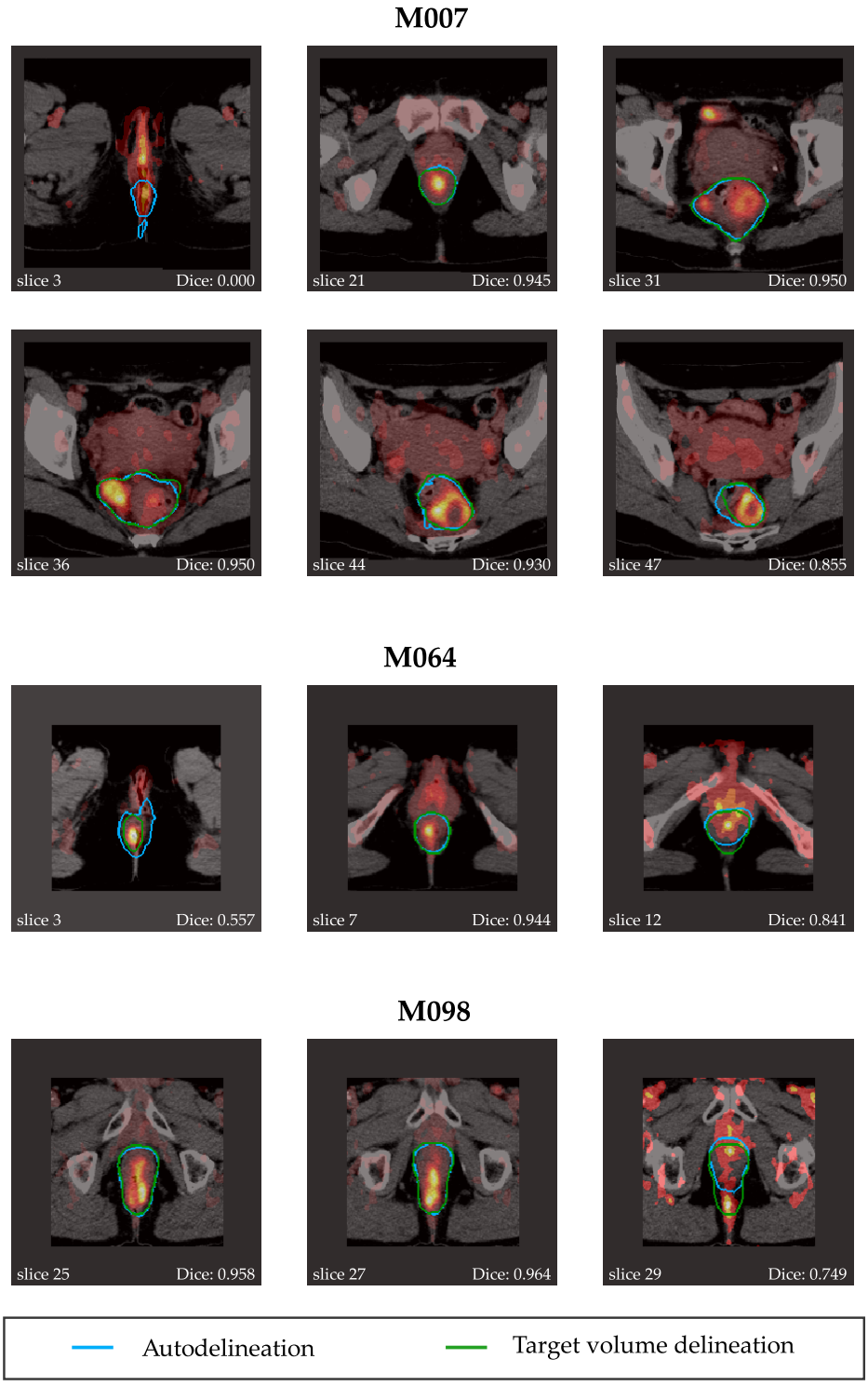


Figure 6.3: 12 fused PET/CT image slices as a result of the PET/CT experiment (run on the MRI dataset) on the validation set. The PET signal is presented using colormap 'hot', ranging from black (no radiotracer signal) to yellow (maximum radiotracer signal). The images are grouped based on the patient ID. The slice number is provided in the lower left corner of each image and the Dice performance is shown in the lower right corner. The blue lines represent the predicted autodelineations, while the green lines are the oncologists target volume delineations.

6.3.2 The test set

How the model performs on the validation set can be a good indicator of how well the model is trained. However, a more realistic evaluation of the model is the performance on an unseen, test set.

When running the PET/CT experiment on the test set, the Dice performance was 0.863 ± 0.133, which is slightly lower than the performance obtained on the validation set (presented in Table 6.1).

Figure 6.4 provide 12 image slices from three different patients. The image slices presented are chosen from a set of 99 image slices, and are aimed to be as representative of the total test set as possible. The base images are fused PET and CT images, where the target volume delineations provided by the oncologists are presented in green and the predicted delineations are marked in blue.

Also here, the predicted delineation seem to match the provided target volume in most cases, resulting in high Dice performances. However, the model fails on slice 37 for patient 'M055' (bottom right in Figure 6.4) which has no oncologist' target volume delineation, in similarity to image slice 3 from patient 'M007' (in Figure 6.3). Slice 14, for patient 'M026' is also an example where the model delineated a smaller area without any significant PET signal, or any sign of tumor tissue from the CT image (see in the upper, middle image of Figure 6.4).

Moreover, observing the autodelineations in Figure 6.4 (see 'M026', slice 29, 'M055', slice 28, 34, 37 and 'M068', slice 24,28, 30), it is apparent that the network model did not get confused by the strong PET signal from the bladder.

Table 6.6: Mean Dice performances and corresponding standard deviations per patient of the PET/CT experiment (run on the MRI dataset, Table 5.3) in the test set.

Patients	Dice
'M026'	0.832 ± 0.125
'M045'	0.896 ± 0.049
'M055'	0.887 ± 0.176
'M064'	0.856 ± 0.128
'M068'	0.886 ± 0.127

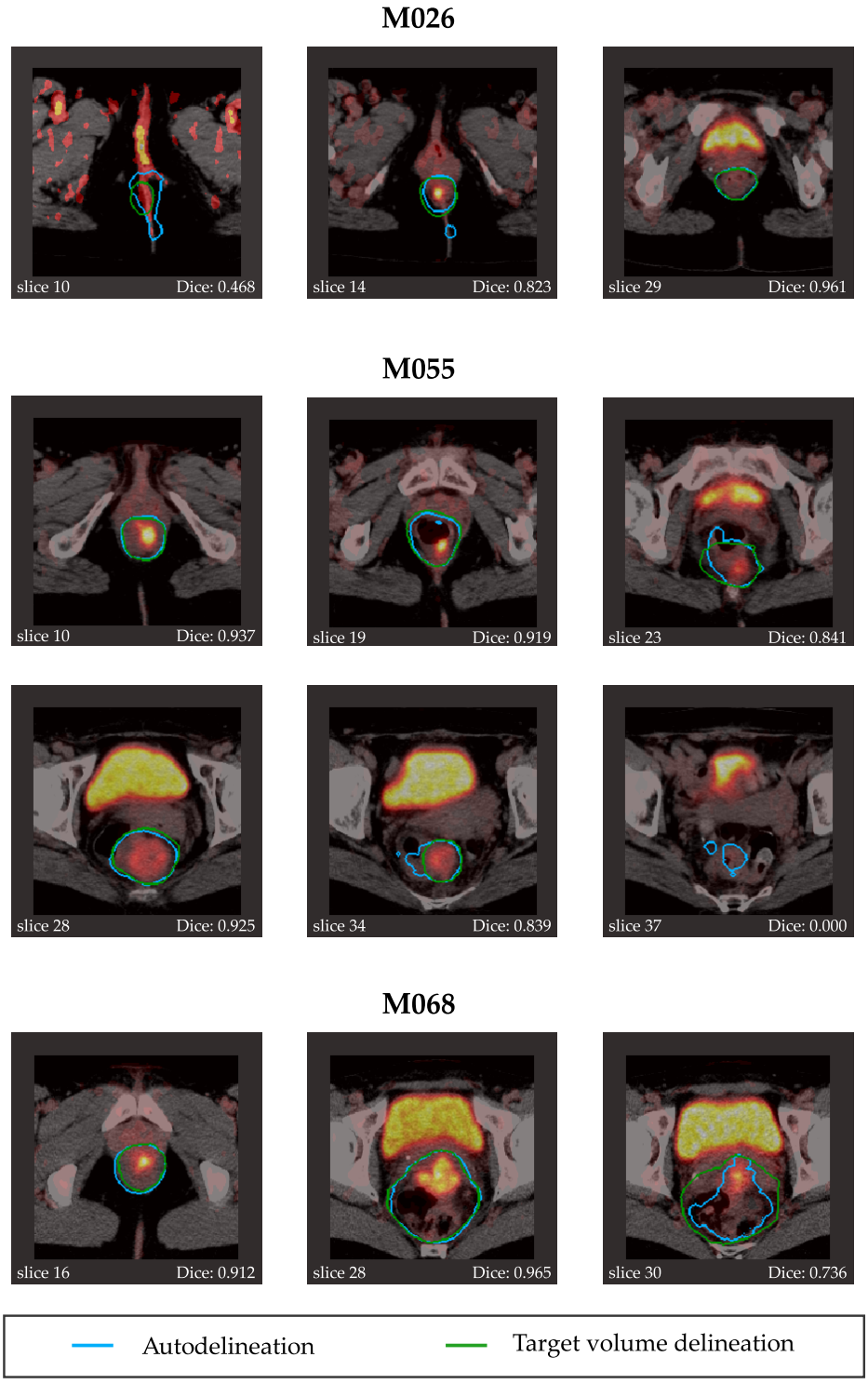


Figure 6.4: 12 fused PET/CT image slices as a result of the PET/CT experiment on the test set. The PET signal is presented using colormap 'hot', ranging from black (no radiotracer signal) to yellow (maximum radiotracer signal). The images are grouped based on the patient ID. The slice number is provided in the lower left corner of each image and the Dice performance is shown in the lower right corner. The blue lines represent the predicted autodelineations, while the green lines are the oncologists target volume delineations.

Chapter 7

Discussion

7.1 The aim of this Master's thesis

The MSc of Moe [18] explored the U-Net architecture for tumor segmentation and performed a vast parameter sweep of the proposed architecture on PET/CT images of head and neck cancers. Moe also used his framework to explore the benefit of combining PET and CT images for the purpose of autodelineation of tumors [18].

This thesis is a part of the ANCARAD study with a dataset consisting of PET, CT and MRI images of 85 AC patients. The aim of this thesis was to compare the Dice performance for sentiment image segmentation of PET, CT and MRI images. Furthermore, an inspection of which imaging modalities, or combination of imaging modalities, may be best suited for autodelineation of tumors in image sequences from AC patients was conducted. To evaluate the autodelineation performances, the overlap based metric Dice was used. The U-Net architecture proposed by Moe [18] was used with modifications and additional techniques, such as image augmentation and data cleaning, to increase the Dice performance of the network. Moreover, the effects of some of the additional techniques on the model performance was investigated.

7.2 Baseline performances

The baseline Dice performance for autodelineation of the ANCARAD dataset, excluding image slices without an oncologist' target volume

delineation, was 70.1 %. Hence, if one were to disregard the deep learning architecture provided in this thesis and only rely on the mean target volume mask based on the delineations made by the oncologists, one should expect a Dice performance of 70.1%. This is a high benchmark to beat, but indicates that the tumor tissues in this dataset are located in more or less the same region in the images, as can be seen in Figure 5.7. It is most likely that this is an effect of the co-registering of the image sequences.

7.3 Model predicted delineations

The Dice performances in Table 6.1 show excellent performances for most of the imaging modalities, but using PET and CT images as model input showed the highest Dice performance with the least standard deviation. However, all of the experiments run for comparing the imaging modalities showed good results. Thus, while concluding that the PET/CT experiment achieved the highest performance, further investigations should be conducted for the comparison of the imaging modalities.

Moreover, the resulting Dice coefficients are initial results based on only one run for each modality or a combination of imaging modalities. Ideally, each experiment should be run several times in order to get a better estimation of the expected Dice performance. This is especially made apparent by the equal Dice performances of PET/CT and PET/DPCT experiments. One might have expected the PET/DPCT to perform better than the PET/CT since the oncologist' delineation was made based on the DPCT. Nonetheless, the PET/CT are expected to be better aligned since these image sequences were generated during a single examination (see Section 2.3).

Validation and test performances
The Dice coefficient of the test set was slightly lower than the performance on the validation test. When the network is trained on a validation set, the model is not expected to perform better on a new, unseen dataset. As observed in Figure 6.2, the distance between the training an validation curves are relatively small which implies that the model is not overfitted either.

Tables 6.5 and 6.6 suggest that the model network performed better per patient on the validation set than on the test set. Using the

Dice performance per patient would give more importance to the total delineation per patient rather than the smaller tumor regions missed by the network per slice. This approach for evaluating the Dice performance of the model is similar to what is used in most medical studies exploring the interobserver variability [8], [9]. It is therefore recommended that any further research of the Dice performance of the network for the ANCARAD project should also inspect the experiment's Dice performance per patient.

Related work
In the study of Guda et al. [8], the Dice performance between the delineations of head and neck made by three different radio oncologists with 10 years of experience was 0.69. To localize cancerous tumors in the head and neck region is, to the author's knowledge, more challenging for the CNN model compared to AC tumors, since there are several options to where the tumor can be located [60]. Therefore, it would not be fair to compare the Dice performance between the radio oncologists in the study of Guda et al. [8] nor the Dice performances achieved by Moe in his MSc [18] to the performances obtained in this thesis.

However, Rusten et al. [9] conducted a study on the ANCARAD dataset, evaluating the interobserver variability of the delineations between three different oncologists, on PET and MRI. The obtained Dice coefficients of 0.80 and 0.74 for PET and MRI, respectively, showed a high degree of overlap between the observers [9]. Furthermore, the choice of modality seemed to have the largest variability.

The oncologists Dice performance of 0.80 for PET is comparable to the 0.867 ± 0.166 in Table 6.1 for autodelineation based on PET only. Moreover, the model surpassed the overlap coefficient when regarding MRI, where the model Dice coefficient for T2W was 0.861 ± 0.177 for T2W only as input, and 0.780 ± 0.192 (as shown in Table 6.1) when using T2W and the ADC maps 'ADC' and 'Perf' as input.

This demonstrates that the proposed model agreed well with the oncologists providing the target volume delineations. Thus, the autodelineations of the proposed model for AC patients, could be comparable to delineations provided by a radio oncologist. Nonetheless, it should be noted that the high Dice coefficient was obtained after excluding 80 % of the slices without anal cancer tissue. The PET/CT experiment on the PET/CT dataset without data cleaning gave a Dice

coefficient of 0.664 ± 0.378, which is inferior to the overlap between the radio oncologists in [9].

Image slices without an oncologist' target volume
There are at least two possibilities for correcting the error of making autodelineations on image slices without an oncologist' target volume: to penalize such mistakes during the training phase and increasing the number of image slices without an oncologist' target volume.

By increasing the weighting of the Precision for the F-score (as explained in Section 3.2.2), the rate of true positives to the total number of positives gains a higher importance. Using this as loss function rather than the Dice might make the model less prone to false positive mistakes, and thereby penalize the false autodelineations more.

In addition, the model needs to see a sufficiently amount of image slices without an oncologist' target volume in order to learn and recognize the patterns in the dataset. A substantially larger dataset would therefore be of great aid for avoiding false delineations on image slices without an oncologist' target volume.

7.3.1 Autodelineations

A pitfall in making an autodelineation program based on deep learning for this dataset is that it might learn the most "reasonable" location of the cancerous tumors for patients with AC. Inspecting the best performing model further yields an indication that the model has learned the location of the anal canal in addition to recogning cancerous tumor tissue to some extent. However, the model network fails at image slices without any oncologist' target volume, where it delineates a area not specified as cancerous.

Nevertheless, locating the anal canal might not be a poor assumption in this case. AC is cancerous tumors located within 4-5 cm from the anal opening [1]. Assuming that the physician has already conducted a physical examination of the patient, and is certain that the patient has anal cancer, the medical imaging can be utilized to localize the tumor for the radio and/or chemotherapy. In that case, localizing the anal canal itself and further adjusting the delineation for irregularities could be sufficient.

If, however, it is of interest to evaluate whether or not the patient has anal cancer, the proposed autodelineation program might not be optimal. Given the provided dataset, the network is not trained to distinguish between a healthy anal canal and an anal canal with cancerous tumor. In order to make sure that the network is not just learning the exact location of the anal canal, future research regarding autodelineation of AC tumors should include image data of the anorectal region from subjects without AC. In this way, one can be more certain that the network is learning to differentiate between a healthy anal canal and an anal canal with cancerous tumor. In that case, there should ideally be an equal amount of AC patients and patients without AC included in the dataset.

7.4 Effect of regularization and cleaning data

In order to regularize the model and increase the invariance in the dataset some procedures were added to the proposed model architecture provided by Moe [18]. These included data cleaning, a Dropout activation and image augmentation.

7.4.1 Cleaning data

The results show that removal of image slices not containing any GTV delineations had the biggest impact on the resulting Dice performance. Data cleaning was also demonstrated to be of importance in a study by Sun et al. [53] on the effectiveness of data in Deep learning.

In the training curves for the experiment excluding data cleaning in Figure 6.2, there was no increasing tendency in the Dice performance. Yet, the validation curve had a slightly increasing tendency. This may indicate that the network would need more than 5000 iterations to be able to recognize image slices that do not have any oncologist' delineation. This should be inspected further by running the experiment excluding data cleaning for an increased number of interations.

7.4.2 Dropout activation

Adding a Dropout activation in the eighth convolutional layer had the least impact of the added procedures, with only 1.24 % change relative to the original PET/CT model experiment (as shown in Table 6.2). Dropout is first and foremost a regularization technique [38]. Moreover, when inspecting the performance of the training and validation set, for experiments with and without Dropout (Figures 6.1 and 6.2), one may observe that there is no significant change in the performance curves. Consequently, one may state that Dropout activation is redundant for this network. However, a further inspection of the effect of Dropout should be conducted by tuning the keep probability (as explained in Section 3.1.4), using Dropout in additional layers or by relocating the Dropout activation to other layers towards the end of the contraction path in the U-Net as recommended by Ronneberger et al. [16].

7.4.3 Image augmentation

Image augmentation (see Section 3.3.4) was added in order to increase the amount of training data by including new variants of the already-existing data. Medical image data is known to already have large variation, since all patients are unique and consequently no two medical images of patients are alike. However, adding additional variants may increase the probability for the network to be familiar with new, unseen samples. Therefore, two different image augmentation techniques were implemented: elastic deformation and horizontal flip.

The elastic deformation applied in these experiments was not excessive, as proposed by Ronneberger et al. [16]. The deformed images only became slightly different versions of the original images. Yet, the effect of image augmentation on the Dice performance was 15.6 %. The reason for the very mild elastic deformation was to maintain the circular characteristics of the oncologist' tumor delineation. In addition, the images were aimed to look more or less like real patient data. For further research on autodelineation of the ANCARAD dataset, a more excessive elastic deformation should be conducted.

The oncologist' target volumes were often located in the center of the image. Therefore, flipping the images horizontally did not provide

the model with sufficient variations of the delineated area (as may be indicated by observing Figure 3.8). The effect of flipping the images alone was not inspected, and this should be explored further in any future experiment including flipping as an image augmentation technique.

Neural augmentation
The augmented images were used to double the training set, as proposed by Wang and Perez [45] and Simard et al. [46]. Wang and Perez [45] also used a different set-up for the augmentation, by including the image augmentation within the loop of the neural net, opening the opportunity for the network to learn the augmentations which best decrease the classification loss [45]. This technique, named 'neural augmentation', should be explored further to boost the model performance, and utilize the image augmentations even further.

7.4.4 Effect of an increased dataset

Sun et al. [53] demonstrated that the performance of their Deep Learning model grew logarithmically as training data increased. The PET/CT dataset increased the Dice performance for the PET/CT experiment excluding all regularization and data cleaning substantially, as shown in Table 6.4. However, inspecting the results in Table 6.3 yields that the PET/CT experiment using the PET/CT dataset, performed inferior to the PET/CT experiment using the MRI dataset even though the MRI dataset had fewer image slices and patients as shown in Table 5.3. Consequently, this might indicate that the autodelineations had reached a peak mean Dice performance per slice relative to the oncologists target volumes for this dataset.

7.5 Experiments

For inspecting the best combination of imaging modalities for the purpose of automatic delineation 63 different experiments could be conducted based on the possible different combinations of the modalities. The experiments chosen for this thesis took in to consideration that not all patients had MRI scans. It is common practice to limit the number of examinations of the patient only to what is considered necessary. Combining a T2W image and a PET image could give a better autodelineation, but would require that both a MRI and a PET examination is performed on the patient.

7.6 Deep Learning in Radiology

Time consumption
Running Convolutional Neural Networks for the purpose of classifying images are tasks that require a large amount of processing power. Each image slice in the dataset contains 55 696 (236 x 236) voxels, or 55 696 variables, which need to be analyzed and evaluated. The sentiment image segmentation is therefore a time consuming task, depending on the available processing power and is probably the main limitation of the proposed model.

The time consumed in this project ranged between two to four hours per experiment (including model training and prediction), depending on the experiment set-up and the dataset used. For the PET/CT experiment with 36 patient, the time consumed for running the pretrained model on a single, arbitrary patient from the validation set, was approximately one minute. In comparison, Rusten et al. [9] reported that the oncologists completed the delineation of the tumors in 20 minutes on average per patient. Ergo, the autodelineation could potentially save at least 15 minutes of work per patient for the oncologists, assuming that the oncologists would still use 5 minutes per patient to re-evaluate the proposed autodelineations.

Time consumption in co-registering
However, the time consumed for co-registering the imaging modalities should also be taken into consideration. The time consumed while co-registering was not measured, and one can therefore not make any accurate assumptions on how much time the co-registering would be per patient.

Nonetheless, the PET/CT experiment used the PET and CT images as input, and the co-registering of these modalities may be redundant since they are generated during a single examination. Table 6.1 presented that the experiments using the PET, CT and T2W individually, also generated excellent Dice performances (> 0.85). Using a single modality as input to the model could decrease the practicality of the co-registering, and may be redundant in such cases as well. Yet, experiments without, or with minimal, co-registering should be considered explored for any future research regarding the proposed autodelineation model. Additionally, measurement of the time consumed for the co-registering per patient would be useful.

Challenges regarding the use of AI
Another common challenge regarding deep learning algorithms is overfitting. Generally, the neural networks are prone to overfit the training data and perform poorly on unseen data [38]. However, by inspecting Figure 6.1, the model network proposed in this thesis did not seem to overfit significantly.

Other issues one should take into consideration is the credibility of the proposed autodelineations. In most cases, it is difficult to explain the basis of the delineations provided. Moreover, legal aspects of using medical data and deciding who has the responsibility of the proposed autodelineations should be discussed. The 'Ethics guidelines for trustworthy AI" [17], requires that proper oversight needs to be ensured while developing an AI system by, for instance, a human-in-command approach. For autodelineation of tumors, such an approach could mean that a radio oncologist uses the autodelineation as a guideline for delineations, and is responsible for the final delineation made.

Lastly, there is the question of privacy and security of medical data. Using AI on sensitive, personal data could be risky as it may violate the health data law [61]. However, using the medical data for the purpose of research, as in the ANCARAD study, the medical data can be utilized to generate a network model. Once the model is saved, the data used to generate the network model is redundant for any future use of the model. As a result, the model can be utilized to run predictions on new, unseen patients independent of sensitive medical data. Yet, database reconstruction attacks (DRAs), as explained in [62] should be explored.

7.7 Limitations of the dataset

7.7.1 The dataset

Size of the dataset
The provided data for the ANCARAD study of 85 patients with PET and CT scans and 36 patients with MRI scans could potentially limit the performance of the proposed autodelineation program. Insufficient data is a common challenge in the medical industry due to privacy concerns [45]. The main problem is often that the model does not generalize well after

training, and therefore might perform poorly on new, unseen data [45]. An attempt to avoid biased results was conducted by doubling the training data with augmented data. This increased the Dice performance. Yet, how well the augmented data can represent realistic subjects and increase the generalization of the model is not definite.

A larger dataset can potentially provide results that are more realistic. This will, however, demand more computational power (hardware) and the computation time will increase depending on the size of the data and the available processing power. There is no definite answer to what amount of data is significant, but generally one can state that more data, or a larger dataset, will result in a more robust and less biased performance for a deep learning algorithm [38]. As demonstrated by Sun et al. [53], the performance of the model could increase logarithmically with an increased pre-training dataset.

When inspecting the effect of an increased dataset, by running the original PET/CT experiment on the PET/CT dataset as opposed to the MRI dataset, the Dice performance did not increase considerably. This indicates that the model may have reached a limit for how well the autodelineations can be made compared to the oncologists target volume delineation.

7.7.2 The target delineation volumes

The autodelineation can only be as good as the target volumes provided by the oncologist. When evaluating the performance of the proposed model, it was assumed that the oncologist' delineations provided the ground truth. However, there are examples where the target volume excludes tissue which appears to be cancerous tissue.

Figure 7.1 is an example of a delineation that may confuse the algorithm, as the delineated area does not entirely match the PET signal (Figure 7.1 b)). In addition, it may be challenging to distinguish the tissue surrounding the target volume from the tissue withing the target volume in the CT image (Figure 7.1 a)).

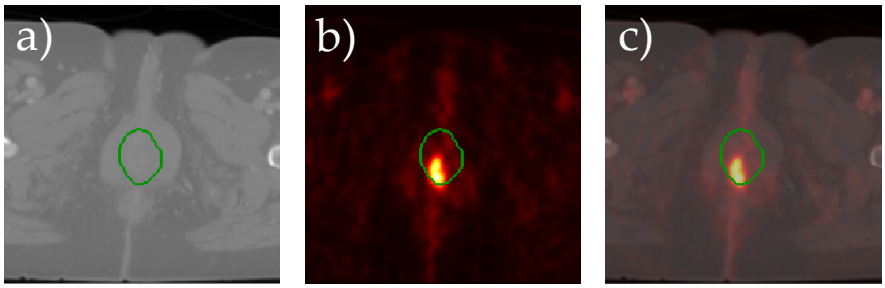


Figure 7.1: Illustration of an oncologist' target volume that may confuse the autodelineation model. The target volume is presented in a) CT, b) PET and c) fused PET/CT image, of patient 'M124', slice 18.

The reasoning behind the target volume delineations is not always apparent for the model, since the oncologist might have additional information about the patient, such as information from physical examinations. Oncologists have access to clinical information, including findings from previous examinations [8]. These parameters are not available for the network, which limits its predicting potential and may confuse the network during the training phase.

It was assumed that the cancerous tumor did not change significantly in-between the scans and was located in more or less the same area. However, the anorectal region consists of soft tissues and despite the co-registering there will be motion artifacts in between the imaging modalities since the medical scans are never generated at the exact same time. Consequently, even the delineations made by the oncologists, which are assumed to be the ground truth, will not match the voxel tissues sufficiently for all the imaging modalities.

Precise delineation of target volumes are considered the weakest link and the largest source of uncertainty in radiotherapy planning [8], [10]. Additionally, as presented in Section 2.5, inter- and intra-observer variability of the delineations are likely to be present [7].

Still, the variability might make the algorithm more robust as a result of increased diversity in the dataset [38] and the algorithm can learn more of the different possibilities for making delineations. However, this requires a substantially increased dataset in order to give the network an opportunity to learn the numerous deviations of the delineations.

Multiple ground truths
Another approach is to provide multiple ground truth delineations to the model, thereby increasing the variability of each contouring. Javaid et al. [63] explored this by having additional, augmented ground truth delineations during the training phase, and thereby increased the robustness of their dilated U-Net model for automatic delineation of multi-organ segmentation in CT images. This approach could be explored further on the ANCARAD dataset to see if the network performance could be be increased further.

7.7.3 Artifacts

Co-registering is mainly based on the bone structures available in the image, which in this case is the hip bones [55]. This is, to the author's knowledge, our best estimate.

Nonetheless, shape and location of the soft tissues can vary substantially since the patient cannot lie in the exact same manner from one examination to another. This effect is reduced through the co-registering [55], but is challenging, if not impossible, to eliminate completely. Consequently, a high intervariability is more likely to be present between image sequences generated by two different imaging modalities, as demonstrated by Rusten et al. [9].

Moreover, the GTV is often not properly aligned for all modalities, which makes it difficult for the model to learn the characteristics of tumour tissue. This might also be the case in Figure 7.1.

ADC maps
Two ADC maps, named 'ADC' and 'Perf', were created in order to capture both slow moving and fast moving water nuclei present in the tumor tissue of interest (see Section 2.4). The images generated as 'Perf' ADC maps, were often unclear and hard to interpret. This may explain the lower Dice performance of 'Perf' in Table 6.1 compared to the other experiments.

Ideally, the ADC maps should be customized for each MRI scan. But, aiming for a more consistent co-registration with as little manual

modifications as possible [55], it was decided to base the ADC maps of all the patients on the same b-values. If the ADC maps had been customized, the images generated might have contained other or more information, and thus the network could potentially have performed better for this particular imaging modality.

7.8 Suggestion for future improvements

7.8.1 Tuning hyperparameters

First of all, the hyperparameters used in this project were mainly based on the conclusions reached in Moe's MSc thesis [18]. However, the windowing did not match precisely, and the remaining hyperparameters should also be examined further to see if the Dice performance of the model can be improved. Conceivably, each experiment has its own uniquely combination of hyperparameters which optimizes the performance for that particular modality or combination of modalities.

Therefore, any further research on autodelineation of the ANCARAD dataset should work with tuning hyperparameters. The choice of loss function, activation function and batch sizes according to each experiment should be reconsidered for a more fair and adequate comparison of the experiments.

SGDR+momentum
Moe recommended in his MSc thesis [18], that future tumor delineation experiments should use the SGDR+momentum as optimizer for the autodelineation model. This optimizer was, however, not used in the experiments conducted for this thesis. Based on the findings of Moe [18], the author would recommend using a SGDR+momentum optimizer for any future automatic tumor delineation experiments, since the SGDR+momentum theoretical advantages over Adam [64].

7.8.2 Lymph nodes

The lymph node structures were included in the dataset during the co-registering to open up for the opportunity of reducing the occurrence

of false positives as they light up in the PET image sequences. Yet, the structures were not used since the autodelineation model did not get confused by the high PET signals, as observed for patient 'M007', slice 31 in Figure 6.3.

7.8.3 Two-phase classification

As presented in Section 4.1.2, only about 50 % of the slices had an oncologist' target volume. In addition to the reported effect of the data cleaning in Section 6.2, the high performances are likely due to the exclusion of image slices without an oncologist' target volume delineation. However, in a real life scenario one might not be certain whether the slice actually contains cancerous tumors or not, and it is likely that most of the slices presented to the network do not contain cancerous tumors.

As a suggestion for further improvements, one could introduce a two-phase learning network, where the model first can state whether or not the image slice contains an area where anal cancer could potentially be present and in a second phase make delineations on the image slices labelled as high probability of containing cancerous tumors. This would mean that a medical image of the stomach or the thighs would be opted out before the second phase, where the tumor autodelineation would take place.

For this approach, one would need significantly more patient data, including patients without anal cancer. This would also increase the computation time, since the network might have to process each image twice. Another shortcoming is if the performance of the first phase is poor. This would result in many false negatives, since the model network could end up autodelineating image slices without an oncologist' target volume delineation.

A similar approach was explored by Kim et al. [65], where they used two-phase learning to boost the performance of their semantic segmentation model. This resulted in an effective enhancement of object localization on a challenging dataset.

Chapter 8

Conclusion

The U-Net autodelineation architecture proposed in the MSc thesis by Yngve Mardal Moe [18] has been explored and a modified version was used to delineate cancerous tumors of patients with anal cancer included in the ANCARAD dataset. In this thesis, the autodelineation was conducted on images generated from the medical imaging modalities PET, CT and MRI. Furthermore, the autodelineation of the images from these modalities, and combinations of some of the modalities, were compared, to evaluate which imaging modality provided the highest Dice performance relative to the corresponding target volume delineations provided by the radio oncologists. Moreover, the effects of regularization, such as image augmentation and Dropout activation, and data cleaning on the Dice performance of the model were explored.

The experiments showed an overall good potential for autodelineating medical images of the anorectal region of AC patients. The Dice performance for the experiments exploring the different modalities, all exceeded the calculated baseline Dice performance for the ANCARAD dataset, which was 0.701. In addition, the experiments' Dice performance were comparable to the inter observer Dice coefficients obtained in the study of Rusten et al. [9]. The experiment using PET and CT was concluded to show the highest performance for the purpose of autodelineation, with a Dice coefficient of 0.885 ± 0.164. However, the model did not seem to be able to evaluate image slices where there was no oncologist target volume. In this case, the performance decreased by 50.3 %. All of the autodelineations generated by the model were located

in the region of the anal canal and did not seem to get distracted by the high PET signals from the bladder and lymph nodes.

In conclusion, autodelineating cancerous tumors of AC patients using a deep learning approach shows excellent results and should be explored further. For future development and research the dataset should include more image data of the anorectal region of patients that are AC free, to train the model to distinguishing between a healthy anal canal tissue from cancerous tumor tissue.

Bibliography

[1] T. Johannessen, ed. *Analkreft (norwegian)*. Norsk Helseinformatikk. Mar. 29, 2019. URL: https://nhi.no/sykdommer/magetarm/endetarm/analkreft/ (visited on 04/08/2019).

[2] O. Dahl and Ø. Fluge. "Analkreft (norwegian". In: *Tidsskr Nor Legeforen* 128.2 (2008), pp. 198–200.

[3] M. G. Guren. *Anal Cancer Radiotherapy (ANCARAD)*. English. Study Description. U.S. National Library of Medicine. Sept. 10, 2013. URL: https://clinicaltrials.gov/ct2/show/NCT01937780 (visited on 02/19/2019).

[4] P. M. Harari, S. Song and W. A. Tome. "Emphasizing Conformal Avoidance Versus Target Definition for IMRT Planning in Head-and-Neck Cancer". In: *International Journal of Radiation Oncology Biology Physics* 77 (3 July 1, 2010), pp. 950–958. DOI: 10.1016/j.ijrobp.2009.09.062.

[5] L. V. Loureiro, L. V. Maia et al. "Waiting time to radiotherapy as a prognostic factor for glioblastoma patients in a scenario of medical disparities." In: *Arquivos de Neuro-Psiquiatria* 73 (2 Sept. 19, 2014), pp. 104–110. DOI: 10.1590/0004-282X20140202.

[6] J. Bresnick. *Arguing the Pros and Cons of Artificial Intelligence in Healthcare*. Sept. 17, 2018. URL: https://healthitanalytics.com/news/arguing-the-pros-and-cons-of-artificial-intelligence-in-healthcare (visited on 04/17/2019).

[7] E. Weiss and C. F. Hess. "The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy." In: *Strahlenther Onkol.* 179 (1 Jan. 1, 2003), pp. 21–30. DOI: 10.1007/s00066-003-0976-5.

[8] S. Gudi, S. Gosh-Laskar et al. “Interobserver Variability in the Delineation of Gross Tumour Volume and Specified Organs-at-risk During IMRT for Head and Neck Cancers and the Impact of FDG-PET/CT on Such Variability at the Primary Site”. In: *Journal of Medical Imaging and Radiation Sciences* 48 (2 June 1, 2017), pp. 184–192. DOI: 10.1016/j.jmir.2016.11.003.

[9] E. Rusten, B. L. Rekstad, C. Undseth, G. Al-Haidari, B. Hanekamp, T. P. Hellebust, E. Malinen et al. “Target volume delineation of anal cancer based on magnetic resonance imaging or positron emission tomography”. In: *Radiation Oncology* 12.1 (147 Sept. 6, 2017). DOI: 10.1186/s13014-017-0883-z.

[10] C. F. Njeh. “Tumor delineation: The weakest link in the search for accuracy in radiotherapy”. In: *Journal of Medical Physics* 33 (4 2008), pp. 136–40. DOI: 10.4103/0971-6203.44472.

[11] A. Vial, D. Stirling et al. “The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review”. In: *Transl Cancer Res* 7 (3 2018), pp. 803–816. DOI: 10.21037/tcr.2018.05.02.

[12] L-C. Chen and Y. Zhu. *Semantic Image Segmentation with DeepLab in TensorFlow*. Mar. 12, 2018. URL: https://ai.googleblog.com/2018/03/semantic-image-segmentation-with.html (visited on 03/11/2019).

[13] X. Shen, H. Gao et al. “High-Quality Correspondence and Segmentation Estimation for Dual-Lens Smart-Phone Portraits”. In: *CoRR* abs/1704.02205 (Apr. 7, 2017).

[14] Q. Ren and R. Hu. “Multi-scale deep encoder-decoder network for salient object detection”. In: *Neurocomputing* 316 (Nov. 17, 2018), pp. 95–104. DOI: 10.1016/j.neucom.2018.07.055.

[15] M. Drozdal, E. Vorontsov et al. “The Importance of Skip Connections in Biomedical Image Segmentation”. In: *CoRR* abs/1608.04117 (Aug. 14, 2016).

[16] P. Fischer O. Ronneberger and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *CoRR* abs/1505.04597 (May 18, 2015).

[17] High-Level Expert Group on Artificial Intelligence European Union. *Ethics guidelines for trustworthy AI*. Report. European Union, Apr. 8, 2019. Chap. I: Foundations of Trustworthy AI, pp. 14–22. 36 pp.

[18] Y. M. Moe. "Deep learning for the automatic delineation of tumours from PET/CT images." English. MA thesis. Norwegian University of Life Sciences (NMBU), Feb. 28, 2019.

[19] J. Kenneth Shultis and Richard E. Faw. *Fundamentals of Nuclear Science and Engineering*. CRC Press, 2016. ISBN: 1498769292.

[20] E. Lin and A. Alessio. "What are the basic concepts of temporal, contrast and spatial resolution in cardiac CT?" In: *Journal of Cardiovascular Computed Tomography* 3 (6 Nov. 1, 2009), pp. 403–408. DOI: 10.1016/j.jcct.2009.07.003.

[21] J. McKenzie and S. Goergen. *Computed Tomography (CT)*. Aug. 31, 2017. URL: https://www.insideradiology.com.au/computed-tomography (visited on 02/15/2019).

[22] L. W. Goldman. "Principles of CT: Multislice CT." In: *Journal of Nuclear Medicine Technology* 36.2 (May 15, 2008), pp. 57–68. DOI: 10.2967/jnmt.107.044826.

[23] J. Lilley. *Nuclear Physics: Principles and Applications (Manchester Physics Series Book 44)*. Wiley, 2013. ISBN: 0-471-97935-X.

[24] K. Kvandal. "Eksplorativ analyse av PET/CT-bilder av hode/hals-kreft med fokus på prediksjon av behandlingsutfall og HPV-status (Norwegian)". Norwegian. MA thesis. Norwegian University of Life Sciences (NMBU), May 15, 2017.

[25] Y. Watanabe. "Derivation of linear attenuation coefficients from CT numbers for low-energy photons." In: *Physics in Medicine and Biology* 44 (9 Oct. 1, 1999), pp. 2201–11. DOI: 10.1088/0031-9155/44/9/308.

[26] T. Kimpe and T. Tuytschaever. "Increasing the number of gray shades in medical display systems–how much is enough?" In: *J Digit Imaging* 20 (4 Dec. 29, 2006), pp. 422–432. DOI: 10.1007/s10278-006-1052-3.

[27] H. Knipe, A. Murphy et al. *Windowing (CT)*. Ed. by H. Knipe. 2019. URL: https://radiopaedia.org/articles/windowing-ct?lang=us (visited on 02/16/2019).

[28] R. L. Wahl, ed. *Principles and Practice of PET and PET/CT*. LWW, Nov. 1, 2008. ISBN: 9780781779999.

[29] E. Safaie, R. Matthews and R. Bergamaschi. "PET scan findings can be false positive". In: *Techniques in Coloprotocology* 19 (6 May 5, 2015), pp. 329–330. ISSN: 1128-045X. DOI: 10.1007/s10151-015-1308-3.

[30] P. E. Kinahan and J. E. Fletcher. "PET/CT Standardized Uptake Values (SUVs) in Clinical Practice and Assessing Response to Therapy". In: *Seminars in Ultrasound, CT and MRI* 31 (6 Dec. 1, 2010), pp. 496–505. DOI: 10.1053/j.sult.2010.10.001.

[31] D. V. Sahani M. A. Blake J. Slattery and M. K. Kalra. "Practical issues in abdominal PET/CT". In: *Appl Radiol.* (Nov. 3, 2005).

[32] D. C. Preston. "Magnetic Resonance Imaging (MRI) or the Brain and Spine: Basics". 2006.

[33] M. Mascalchi, M. Filippi et al. "Diffusion-weighted MR of the brain: methodology and clinical application." In: *Radiol Med* 109.3 (Mar. 19, 2005), pp. 155–97.

[34] P. B. Kingsley and W. G. Monahan. "Selection of the Optimum b Factor for Diffusion-Weighted Magnetic Resonance Imaging Assessment of Ischemic Stroke". In: *Magnetic Resinance in Medicine* 51 (5 Apr. 26, 2004), pp. 996–1001. DOI: 10.1002/mrm.20059.

[35] D-M Koh and A. R. Padhani. "Diffusion-weighted MRI: a new functional clinical technique for tumour imaging". In: *Br J Radiol.* 79 (944 May 30, 2006), pp. 633–635. DOI: 10.1259/bjr/29739265.

[36] "4.Definition of Volumes". In: *Journal of the International Commission of Radiation Units and Measurements* 10.1 (1 Apr. 1, 2010), pp. 41–53. DOI: 10.1093/jicru/ndq009.

[37] S. Nikolov, S. Blackwell et al. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy". In: *CoRR* abs/1809.04430 (Oct. 12, 2018).

[38] S. Raschka and V. Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition*. Packt Publishing, Sept. 1, 2017. ISBN: 9781787125933.

[39] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2011. ISBN: 9780387310732.

[40] D. P. Kingma and J. L. Ba. "Adam: A Method for Stochastic Optimization". In: ICLR 2015 - The 3rd International Conference for Learning Representations. May 7, 2015.

[41] N. Srivastava, G. Hinton et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *Journal of Machine Learning Research* 15 (June 1, 2014), pp. 1929–1958.

[42] A. A. Taha and A. Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC Med Imaging* 15 (29 July 9, 2015). DOI: 10.1186/s12880-015-0068-x.

[43] J. Hua and X. Gong. "A Normalized Convolutional NeuralNetwork for Guided Sparse Depth Upsampling". In: *International Joint Conference on Artificial Intelligence (IJCAI-18)* (July 13, 2018), pp. 2283–2290. DOI: 10.24963/ijcai.2018/316.

[44] F. Ma and S. Karaman. "Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image". In: *CoRR* abs/1709.07492 (Sept. 21, 2017). arXiv: 1709.07492.

[45] L. Perez and J. Wang. "The Effectiveness of Data Augmentation in Image Classification using Deep Learning". In: *CoRR* abs/1712.04621 (Dec. 13, 2017).

[46] D. Steinkraus P. Y. Simard and J. C. Platt. *Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis*. Vol. 2. I.E.E.E.Press, Aug. 3, 2003, p. 958. ISBN: 0-7695-1960-1.

[47] L. C. Chen, G. Papandreou et al. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs". In: ICLR. 2015. DOI: https://arxiv.org/pdf/1412.7062.pdf.

[48] J. Long, E. Shelhamer and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *CoRR* abs/1411.4038 (Nov. 14, 2014). arXiv: 1411.4038.

[49] W. Wang and J. Shen. "Deep Visual Attention Prediction". In: *IEEE Transactions on Image Processing* 27 (5 May 27, 2018), pp. 2368–2378. ISSN: 10577149. DOI: 10.1109/TIP.2017.2787612.

[50] J. Yang, B. Price et al. "Object Contour Detection with a Fully Convolutional Encoder-Decoder Network". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 27, 2016). ISSN: 1063-6919. DOI: 10.1109/CVPR.2016.28.

[51] *Approvals for medical and health research*. Regional Committees for Medical and Health Research Ethics. July 22, 2015. URL: https://www.med.uio.no/english/research/phd/application/how-to-apply/approvals-for-medical-and-health-research.html (visited on 01/22/2019).

[52] *Søk om data fra NPR*. Norwegian. Helse direktoratet. 2018. URL: https://www.helsedirektoratet.no/english/norwegian-patient-registry (visited on 01/22/2019).

[53] C. Sun, A. Shrivastava et al. "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *CoRR* abs/1707.02968 (2017). DOI: 10.1109/iccv.2017.97. eprint: 1707.02968.

[54] *MICE TOOLKIT. Medical Image analysis made easy*. MICE Toolkit Software. 2019. URL: https://www.micetoolkit.com/ (visited on 02/12/2019).

[55] C. K. Kaushal. "Termpaper". English. Dec. 6, 2018.

[56] MICE Toolkit. *User manual*. MICE. Mar. 1, 2018. URL: https://www.micetoolkit.com/wp-content/uploads/2018/03/MICE_user_manual_v1.0.4.pdf (visited on 02/02/2019).

[57] Andrew Collette. *Python and HDF5: Unlocking Scientific Data*. O'Reilly Media, 2013, pp. 5–7. ISBN: 1491944994.

[58] *Spyder. The Scientific Python Developed Environment*. 2018. URL: https://www.spyder-ide.org/ (visited on 05/10/2019).

[59] F. J. Reh. "Pareto's Principle-The 80-20 Rule: the National Magazine of Business Fundamentals CFM". English. In: *Business Credit* 107.7 (July 2005). Copyright - Copyright National Association of Credit Management Jul/Aug 2005; Last updated - 2014-05-19; SubjectsTermNotLitGenreText - United States–US, p. 76.

[60] S. Y. Lai D. M. Cognetti R. S. Weber. "Head and Neck Cancer: An Evolving Treatment Paradigm". In: *Cancer. Author manuscript; available in PMC 2009 Sep 24* 113 (S7 Sept. 17, 2008). DOI: 10.1002/cncr.23654.

[61] *Helseregisterloven*. Norwegian. report 8. Apr. 10, 2019.

[62] J. M. Abowd S. Garfinkel and C. Martindale. "Understanding Database Reconstruction Attacks on Public Data". In: *Communications of the acm* 62.3 (Mar. 1, 2019). DOI: 10.1145/3287287.

[63] U. Javaid, D. Dasnoy and J. Lee. "Segmentation of CT images with AI: compensating annotation uncertainties using contour augmentation". In: ESTRO 38. (Apr. 26, 2019). PO-1016. Apr. 26, 2019. DOI: 10.3252/pso.eu.ESTRO38.2019.

[64] I. Loshchilov and F. Hutter. "SGDR: Stochastic Gradient Descent with Restarts". In: *ICLR 2017 conference submission* abs/1608.03983 (2016).

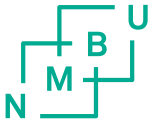
[65] D. Kim, D. Cho et al. "Two-Phase Learning for Weakly Supervised Object Localization". In: *ICCV* (Oct. 22, 2017), pp. 1 3534-3543. DOI: 10.1109/ICCV.2017.382.

Appendix A

Patient numbers

Table A.1: Conversion from patient number to the patient ID.

Patient number	Patient ID
1	′M003′
2	′M007′
3	′M009′
4	′M012′
5	′M015′
6	′M018′
7	′M020′
8	′M023′
9	′M026′
10	′M027′
11	′M030′
12	′M031′
13	′M033′
14	′M038′
15	′M044′
16	′M045′
17	′M047′
18	′M049′
19	′M052′
20	′M053′
21	′M055′
22	′M061′
23	′M064′
24	′M066′
25	′M067′
26	′M068′
27	′M070′
28	′M074′
29	′M087′
30	′M089′
31	′M096′
32	′M098′
33	′M100′
34	′M101′
35	′M105′
36	′M110′



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway