Norwegian University
of Life Sciences

Master's Thesis 2017 30 ECTS
Faculty for science and technology
Professor Cecilia Marie Futsæther

Assessment of a diagnostic program for autodelineation of head and neck cancer based on PET/CT images

Martine Mulstad
Environmental physics and renewable energy

Acknowledgements

This Master's thesis was written during the autumn semester of 2017 and marks the completion of my Master's Degree in Environmental Physics at the Norwegian University of Life Sciences (NMBU). The work has been performed in cooperation with the Department of Medical Physics at Oslo University Hospital (OUS) and the Department of Physics at the University of Oslo (UiO).

First of all, would I like to offer a special thank you to my main supervisor, Professor Cecilia Marie Futsæther, Faculty of Science and Technology (REALTEK), NMBU for excellent guidance, support and feed-back during the course of this thesis. Secondly, I would like to thank Turid Torheim for developing the autodelineation program and for answering questions regarding this program. I am thankful for the opportunity to do research I find useful and important.

Further, I would like to thank my co-supervisor, Professor Eirik Malinen, Department of Physics, UiO, for the access and co-registration of the dataset, and also for being available to answer questions regarding the dataset. The oncologists and nuclear medicine physicists that performed the delineations of the cancerous voxels, used as the ground truth in the autodelineation program, deserves a big thank you for providing the foundation of this thesis.

Then I would like to thank PhD student Aurora Rosvoll Grøndahl and Master student Kari Helena Kvandal for help in the organizing and processing of our dataset, in addition to interesting and valuable discussions. Kari deserves an extra thank you for wonderful illustrations, used in this thesis, and support in the various steps in the process of writing this Master's thesis. I want to thank PhD student Stanislau Trukhan for giving me access and sharing the computer my models were run on. Finally, I want to thank my loved ones for their support, proof reading and encouragement during this Master's thesis.

Ås, 15.12.2017
Martine Mulstad

Abstract

Delineation of tumors and cancerous lymph nodes in medical imaging is a challenging, time-consuming and complex part of radiotherapy planning. A program for autodelineation of cervical cancer from MRI data was investigated to evaluate it's possible use on PET/CT images.

In this Master's thesis an autodelineation program developed to identify cervical cancer tumors from different types of MR images was investigated. This program classifies every voxel in MR image stacks as either cancerous or non-cancerous, using voxel intensities, spatial relationships and Fisher's Linear and Quadratic Discriminant Analysis (LDA and QDA).The aim of this thesis was to further develop the autodelineation program and adapt it to delineate head and neck cancers in PET/CT images.

The dataset used in this study consisted of 206 head and neck cancer (HNC) patients who had undergone ^{18}F-FDG PET and contrast-enhanced CT in conjunction with radiotherapy. All patients were treated at Oslo University Hospital (OUS), Norway between 31.10.2007 and 31.07.2015.

Contours delineating tumors and cancerous lymph nodes in the PET/CT images made by experienced oncologists and nuclear medicine physicists and are considered the ground truth. The contours were used to train and evaluate the autodelineation program. A total of twenty-four models were run for different combinations of classifiers, imaging modalities, spatial information and sorting of neighbors. The models were evaluated using the five performance measures Dice Similarity Coefficient (*DSC*), area under the ROC curve (*AUC*), κ-statistics, sensitivity and specificity.

Model evaluation revealed that there were large variations in delineation performance between patients especially for the *DSC* and κ values. Inclusion of the PET images in the models significantly improved model performance. Of the twenty-four models, a total of thirteen models, based on different combinations of either PET or CT + PET, gave an *AUC* larger than 0.90, *DSC* of 0.64-0.68 and κ of 0.56-0.62, indicating very good model performance and substantial agreement

between the ground truth and the model mask. Overlap between ground truth and the model delineation was significantly poorer when only CT images were included in the classification. In this case, DSC and κ were in the ranges 0.27-0.40 and 0.12-0.27, respectively.

There was a tendency for higher sensitivity for models based on CT (0.72-0.97) than for both PET alone (0.62-0.76) and PET in combination with CT (0.58-0.84). Thus, CT correctly classified more voxels as cancerous compared to PET and PET in combination with CT. On the other hand, CT had significantly lower specificity (0.26-0.50) than both PET (0.88-0.96) and CT + PET (0.84-0.96). As a consequence, the inclusion of PET (PET and PET + CT) images in the model resulted in a higher number of correctly classified voxels compared to CT images alone.

Including spatial information in the form of neighboring voxels significantly improved model performance, whereas sorting of the voxel neighbors in order of descending intensity had little effect. The choice of classifier had little effect on performance, except for delineation in CT images where QDA performed significantly better.

All images were cropped to remove artifacts surrounding the patient such as air and parts of the PET/CT unit. Removing image slices not containing cancerous voxels is recommended to further improve the balance between the classes. For classification based on CT images, there was a near linear relation between both κ and DSC and the fraction of voxels of cancerous regions, with κ and DSC increasing as the voxel balance between classes improved.

Overall, inclusion of PET images in the modeling was the dominant factor affecting model performance. As a comparison to the delineation model, tumor and lymph node contouring was attempted using a simple thresholding of the PET images, where voxels with intensities larger than a chosen threshold (SUV > 2.5) were defined as belonging to the cancerous structures. Similar performance measures to the delineation models were obtained, further emphasizing the dominance of the PET images for contouring.

Although PET had a significant effect on the performance measures, it was, however, prone to false positives and false negatives as the tracer 18F-FDG provides information about the glucose metabolism of different tissues. Non-cancerous tissue can have high glucose metabolism and in a few cases cancerous tissue can have low glucose metabolism.

With further testing and optimization, this autodelineation model has the potential of becoming a useful tool for physicians for contouring and assessment of different types of cancers based on a variety of different imaging modalities.

Contents

List of Abbreviations

CT Computed tomography

CTV Clinical target volume

^{18}**F-FDG** 2-Deoxy-2-[18F]fluoroglucose

^{18}**F** Flourine-18

GLUT Glucose transporters

GTC Gross tumor volume

HNC Head and neck cancer

IARC International Agency for Research on Cancer

ICRU The International Commission on Radiation Units and Measurements

LDA Linear Discriminant Analysis

LOR Line of response

OAR Organs at risk

OUS Oslo University Hospital

PET Positron emission tomography

PTV Planning target volume

RT Radiation therapy

SUV Standardized Uptake Value

TOF Time of flight principle

UICC The Union for International Cancer Control

QDA Quadratic Discriminant Analysis

Chapter 1

Introduction

Background

The International Agency for Research on Cancer (IARC) estimated a total of 14.1 million new cancer cases and 8.2 million cancer deaths worldwide in 2012 [1]. Due to the growth and aging of the population, the global burden of cancer is estimated to grow to 21.7 million new cancer cases and 13 million cancer deaths by 2030 [1]. Other factors, such as smoking, poor diet, physical inactivity and so forth, are expected to further increase the global burden of cancer [1].

According to the Cancer Registry of Norway, the cancer incidence has increased with approximately 3% after 2014, and this increase is equally represented for both sexes [2]. The probability of being diagnosed with a cancer before the age of 75, is approximately 36% in men and 30% in women [2].

Cancer is a disease that always begins in cells and occurs when abnormal cells divide in an uncontrolled way, as signals controlling how much and how often the cell divide are either faulty or missing [3]. Then these cells can start to multiple and grow into lump called a tumor [3]. The primary tumor is referred to as the volume of cancer cells where the cancer starts, and the first cancer cells can later potentially spread into other tissues [3]. Cancer is a heterogeneous disease as there are more than 200 different types of cancer [3].

Cancer is a common, complex disease with an increasing global burden and many influencing causes, and is thus an important and large research field with research being performed in multiple scientific disciplines. Only within the field of medical imaging, there are numerous research questions to investigate related to the preferred imaging modality for different cancer types, how to optimally detect and delineate the cancerous regions and considerations of different types of volume delineation of the cancerous regions regarding the use of radiation therapy [3, 4, 10, 43].

Aim of the Master's Thesis

In this Master's thesis the focus lies on delineation of tumors and cancerous lymph nodes in medical images, as this is a challenging, time-consuming and complex part of radiotherapy planning. With the estimated increasing global burden of cancer, the benefit of an autodelineation program to faster, more consistently and accurately detect cancerous regions from medical images would be beneficial and of major importance for the treatment of cancer. Since the implementation of the autodelineation program can affect the time between the scan and when treatment starts, it can be valuable especially for patients with aggressive tumor growth [4].

Torheim et al have developed a diagnostic tool for autodelineation of cervical cancer based on MRI scans [4]. This program classifies every voxel in MR image stacks as either cancerous or non-cancerous, using voxel intensities, spatial relationships and Fisher's Linear and Quadratic Discriminant Analysis (LDA and QDA). The aim of this thesis was to further develop the autodelineation program and adapt it to delineate head and neck cancer in PET/CT images.

Head and neck cancer is a rare form of cancer, and accounts for just over 2% of the total number of new cancer cases in Norway [5]. The dataset used in this study consisted of 206 head and neck cancer (HNC) patients who had undergone ^{18}F-FDG PET and contrast-enhanced CT in conjunction with radiotherapy. These patients were treated at Oslo University Hospital (OUS), Norway during the eight years between 31.10.2007 and 31.07.2015.

Build-up of the Master's Thesis

This thesis starts by explaining head and neck cancers, defining delineation volumes, display consideration related to the hybrid PET/CT scanner and going through the steps of supervised learning, in Chapter 2 (Theory). In Chapter 3, the washing and organization of the dataset is thoroughly described. When the dataset was quality assured, the autodelineation program could be tested and further developed. Thus, Chapter 4 and 5 consist of the modifications and results of the autodelineation program. Then the results are discussed and the autodelineation program is assessed, in Chapter 6. The last chapter is the conclusion, summarizing the finding of this study.

Chapter 2

Theory

2.1 Head and neck cancer

Head and neck cancer (HNC) accounts for just over 2 % of the total number of new cancer cases in Norway [5]. Head and neck cancer is a heterogenous group of cancer and is categorized by the area of the head and neck in which it begin [5, 6]. The head and neck areas are illustrated in Figure 2.1, and consist of paranasal sinuses, nasal cavity, oral cavity, tongue, salivary glands, larynx, and pharynx (including the nasopharynx, oropharynx, and hypopharynx).

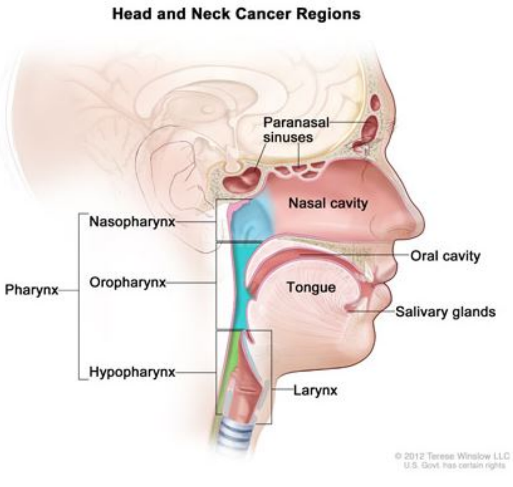


Figure 2.1: Head and neck cancer regions, illustrating the location of paranasal sinuses, nasal cavity, oral cavity, tongue, salivary glands, larynx, and pharynx (including the nasopharynx, oropharynx, and hypopharynx) [6]. For the National Cancer Institute © 2012 Terese Winslow LLC, U.S. Govt. has certain rights.

Squamous cell cancer is responsible for 90 % of all head and neck cancers. This type of cancer begins in the squamous cells lining the moist, mucosal surfaces inside the head and neck, such as inside the mouth, the nose and the throat [5,6]. Patients diagnosed with head and neck cancer in Norway have an average age of 64 years and the majority are male [5].

2.1.1 Causes of Head and Neck Cancer

There are different causes of head and neck cancer depending on the location of the HNC [5]. Tobacco and alcohol are the main risk factors for HNC and there are indications that they have a synergistic effect [5,7]. This means that the effect due to both alcohol and tobacco produces an effect greater than the sum of their individual effects.

Oropharyngeal cancer is the only cancer type in the head and neck region proven to be related to oncogenic forms of the human papillomavirus (HPV) [7,8]. However, HNC in other areas also indicate a relation to HPV [7, 8]. HPV is a group of more than 150 related viruses, where each HPV virus is given a unique number (called its HPV type) [8]. The large recent increase in incidences of oropharyngeal cancer is HPV-related, especially related to HPV type 16 [2, 9]. In a study from USA regarding oropharyngeal squamous cell, around 60 % of the people (in the study) have HPV 16 [9]. HPV-positive and HPV-negative oropharyngeal cancers are clinically and molecular distinct [9]. Studies have indicated that HPV-positive oropharyngeal cancer seems to be more responsive to treatment, such as chemotherapy and radiation, than HPV-negative disease [9–11].

There are also some additional causes of HNC for specific areas in the head and neck. Cancer in areas inside the mouth can be caused by bad dental hygiene and status [12]. Occupational exposure to nickel and dust from both hardwoods and leather products increases the risk of cancer in the areas inside the nose, the throat and in the sinuses [13, 14]. Previous exposure to ionizing radiation is the only known risk factor for cancer in the salivary glands [15].

2.1.2 Treatments

There are different kinds of treatment available for patients with HNC depending on a number of factors, such as stage and location of the cancer. The stage of the cancer is given by its TNM classification. The TNM classification is developed by the Union for International Cancer Control (UICC), and it is an anatomically based system that records the primary and regional nodal extent of the tumor and the absence or presence of metastases [16].

In general, the UICC TMN Classification is divided into the following three categories (individual aspects of the TNM):

- T describing the primary tumor site,
- N describing the regional lymph node involvement, and
- M describing the presence or otherwise of distant metastatic spread.

The TNM classification is described in detail for cancer in specific regions of the body, even within the head and neck area, in (the last/eight edition of) UICC's *TNM Classification of Malignant Tissue* [16].

Patients with tumors localized in the pharynx and larynx, are predominantly treated (at Oslo University Hospital (OUS)) with a 6-week course of external beam radiotherapy with concomitant administration of weekly chemotherapy (cisplatinum) [10]. Radiotherapy kills or damages cancer cells in specific areas (where the irradiation is aimed). Chemotherapy uses specific drugs that can also affect cancer cells located in other areas of the body [17]. The three main goals for chemotherapy are to cure, control and palliation the cancer and its side effects [17]. Surgery is another commonly used treatment, removing cancer cells located in the volume operated out of the patient's body [17].

2.2 Volume delineations used in radiotherapy planning

In the treatment planning and reporting processes, a number of different volumes, related to both tumor and normal tissue, have to be defined [18]. The delineation of these volumes must be performed before the radiation therapy part of treatment can begin, as the absorbed dose is dependent on the volume the radiation is aimed at [18]. The absorbed dose, D_T [Gy = J/kg], is given as

$$D_T = \frac{\epsilon_T}{m_T}, \tag{2.1}$$

where ϵ_T [J] is the total energy deposited in a mass m_T [kg] of the irradiated tissue [19].

The volumes are delineated on the basis of image stacks from various imaging modalities, for example Positron Emission Tomography (PET) and Computed Tomography (CT) [18]. In the delineation, the voxels (small volume elements of tissue) in the image stack are assigned as malignant or normal tissue, for example as cancerous and non-cancerous voxels, by the physician. All of the cancerous voxels would then make up the total cancerous volume for that patient. Target volumes would consists of the total cancerous volume and most likely also an extra edge of variable size. The most commonly used target areas are

- Gross tumor volume (GTV),
- Clinical target volume (CTV),
- Internal target volume (ITV), and
- Planning target volume (PTV).

The International Commission on Radiation Units and Measurements (ICRU) develops internationally accepted recommendations for all radiation units and measurements, for example by defining several target volumes for radiation therapy planning [20]. As a consequence, treatments in different clinics around the world would be based on the same recommendations, thus making it possible to compare clinical results and data.

GTV and CTV have an anatomic basis, being independent of the chosen irradiation technique and only influenced by oncological considerations [18]. Gross tumor volume (GTV) is the demonstrable area of gross malignant growth [18]. The GTV may include the primary tumor, metastatic regional lymph node(s) and distant metastasis [18]. Normally, GTV structures are defined for the primary tumor and nodes separately. To describe and report the GTV as accurate as possible is essential, as GTV is

- required for staging (according to the TNM classification),
- the minimal volume the adequate absorbed dose must be aimed at,
- evaluated through the course of treatment and these changes of GTV might be predictive of treatment outcome [18].

The clinical target volume (CTV) is a volume that contains the demonstrable GTV and also tissue, relevant for radiation therapy, that has a certain probability of being malignant [18]. A probability of disease higher than 5-10 % is normally assumed to require treatment, and this probability threshold is based on clinical experience [18]. This generally includes areas immediately surrounding the primary tumor and lymph nodes, in addition to areas where tumor infiltration or metastasis is likely to occur [18].

The ITV and PTV are geometric volumes, introduced to ensure that the absorbed dose delivered to a specific volume, with a clinically acceptable probability, matches the prescription constraints [18]. The internal target volume (ITV) consists of the CTV plus an internal margin. The internal margin take into account the uncertainties in size, shape, and position of the CTV within the patient (such as movement of internal structures due to the respiration cycle) [18]. The planning target volume (PTV) is based on the ITV and a set-up margin that accounts for geometrical uncertainties, such as patient position during the scan and alignment of the therapeutic beams during the treatment planning and sessions [18].

During radiation therapy (RT), it is desirable to reduced the amount of irradiation of (critical) normal tissue and radiosensitive organs, organs at risk (OAR), as much as possible - while ensuring adequate absorbed dose delivered to the malignant volumes [18]. This is because irradiation of these tissues could have substantial consequences, such as reduced functionality of the tissue/organ, and therefore might influence the treatment planning and/or the prescribed absorbed dose [18].

2.3 PET/CT

A PET/CT scanner is an important imaging modality in cancer research and diagnosis, and is useful when it comes to

- Diagnose cancer,
- Consider the effect of treatment,
- Distinguish benign changes from cancer,
- Distinguish scar tissue after operation from tissue with regrowth of cancer cells,
- Assess the prevalence of cancer, and
- Study the suspected spread of cancer [21].

PT/CT scanners combine a functional Positron Emission Tomography (PET) scanner and an anatomical Computed Tomography (CT) scanner into one integrated device [22]. In this device, the CT gantry is positioned parallel to the PET gantry, and it is critical that these gantries are aligned properly in all dimensions [23]. In addition, to image the patient at the exact same point in both modalities the separation between the two gantries must be adjusted for (through the position of the imaging table during scan) [23]. The PET/CT scanner can be used to acquire only CT scans, only PET scans or combined PET/CT scans [23]. The physics behind the underlying imaging modalities, PET and CT, are the basis for the hybrid PET/CT scanner.

2.3.1 Computed Tomography (CT)

Computed Tomography (CT) is an anatomical imaging technique that measures the X-ray attenuation through thin cross sections of the body [23]. In other words, the CT scanner measures the reductions in intensity, due to absorption or deflection, of the X-ray beam when it goes through a given cross section of the body [23]. The X-ray intensity beam, I_t, measured after traversing the material of thickness, Δx , is given as

$$I_t = I_0 e^{-\mu\ \Delta x}, \tag{2.2}$$

where I_0 is the X-ray intensity emitted from the X-ray source and μ is the linear attenuation coefficient of the specific material [23]. Since the information of the X-ray intensity without a body present, I_0, is known, it is possible to compute the sum of X-ray absorption along one line corresponding to a particular X-ray beam.

The X-ray intensity beam, I_t, through different cross sections of the patient, is measured using a rotating frame with am X-ray tube mounted on one side and a detector array on the opposite (the CT gantry), as illustrated in Figure 2.2. In a spiral CT the patient's body is scanned in a spiral path; the CT gantry is rotated while the patient, lying on the scanning table, is moved through the scanner.

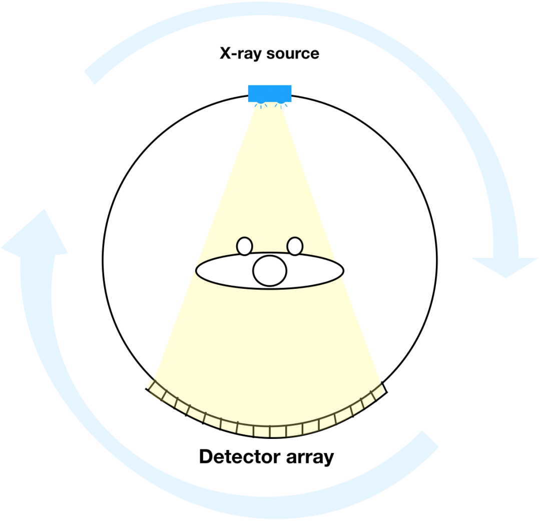
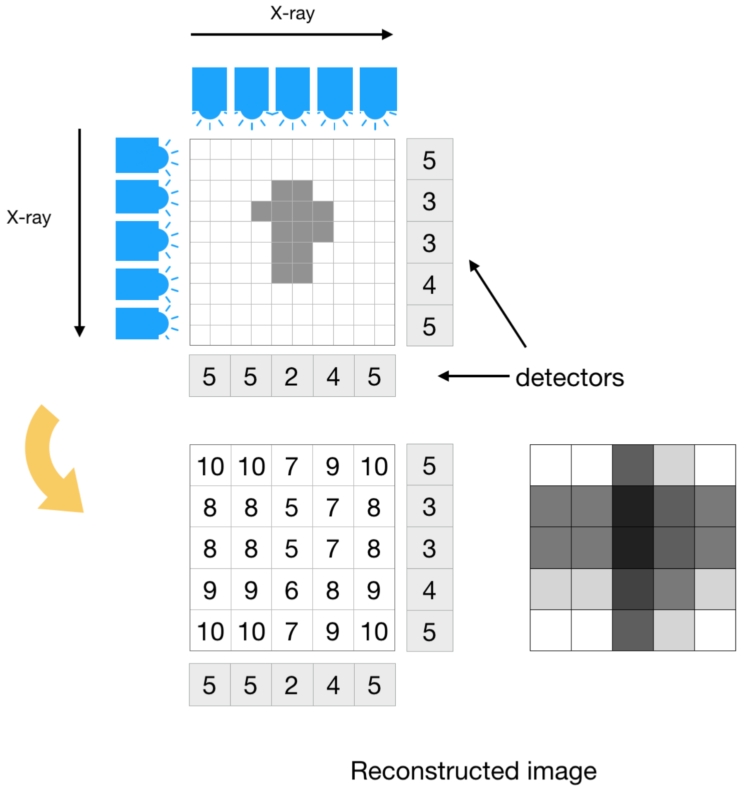


Figure 2.2: An illustration of the gantry of a CT scanner, consisting of an x-ray source and detector array (located at the opposite side) that rotates around the patient. With permission from Kari Helena Kvandal.

Forward projections to reconstructed CT image

During a CT scan, each detector in the (detector) array measures the X-ray intensity, I_t, that is transmitted through a given cross section of the patient's body [22]. In the upper left illustration in Figure 2.3, the X-ray absorption measured is illustrated, for two X-ray sources and detector orientations, as a number and a shaded box for each detector [22]. The detector's recordings are called the forward projection [22].



10	10	7	9	10	5
8	8	5	7	8	3
8	8	5	7	8	3
9	9	6	8	9	4
10	10	7	9	10	5
5	5	2	4	5	

Figure 2.3: The process from forward projection (upper left) through back projection (lower left) to reconstructed CT image (lower right) is illustrated for two X-ray sources and detector orientations. The gray scale of the tissue represents its absorption coefficient, where the grey color indicates no absorption (μ = 0) and the white color that there is some absorption ($\mu \neq 0$) of the X-ray beam. With permission from Kari Helena Kvandal.

When the forward projections have been measured for all the relevant angles, the process of back projection can begin. This process, illustrated in the lower left of Figure 2.3, reconstructs the original tissue's pattern of X-ray absorption based on the sum of each detector's absorption along the projected path [22]. The resulting matrix of local X-ray absorptions using a gray scale image leads to the reconstructed CT image (lower right in Figure 2.3). The gray scale of the tissue represents its absorption coefficient, where the grey color indicates no absorption ($\mu = 0$) and the white color that there is some absorption ($\mu \neq 0$) of the X-ray beam.

Each picture element (pixel) for a specific slice of the displayed reconstructed image, thus two-dimensional (2D), contain the pixel's CT value (the gray scale value) [22]. By combining pixels to form voxels (volume elements in the image stack), each voxel would result in an averaged CT value for the given slices used. When the CT values are normalized to the attenuation properties of water, it is referred to as the CT_{number} and is reported in Hounsfield units (HU) [22]. The CT_{number} [HU] is defined as,

$$CT_{number} = \frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \times 1000, \tag{2.3}$$

the percentage difference between the X-ray attenuation coefficient of a given voxel, μ_{tissue}, and that of water, μ_{water}, multiplied by the value 1000 [22]. While the pixel values (integers) are dependent on the X-ray intensities transmitted through cross sections of the body, the convention is to define water as 0 HU and air as -1000 HU independently of the X-ray spectrum [22]. The CT_{number} of human tissue depend upon the X-ray spectrum, and is negative for fat (-50 to -100 HU), very high for dense bone (> 1000 HU) and slightly positive for muscles and lymph nodes (40 to 50 HU) and blood (50 to 60 HU) [23].

CT contrast agent

CT contrast agents, for example Visipaque 320, have a rapid uptake in the body and are therefore injected shortly prior to the scan. The uptake of CT contrast agents is higher in malignant than normal tissue as the contrast agents leak faster through the more chaotic arterial network in the tumor and other malignant tissues [24–27]. The use of a contrast agent prior to scanning results in contrast enhancement and therefore larger differences in CT_{number} between normal and malignant tissue in the CT image. As a consequence, it is easier to differentiate between normal and malignant tissue when CT contrast agent is applied. Then classification based on CT_{number} would improve. There is a direct relation between the amount of contrast enhancement and contrast agent (level of iodine) injected into the patient [28]. Limiting factors are due to the fact that contrast agents are associated with risks, such as radiation exposure and potential allergic reactions [28].

2.3.2 Positron Emission Tomography (PET)

Positron Emission Tomography (PET) is a functional imaging technique that measures different types of body function depending on the radionuclide tracer used [22]. The most commonly used tracer in PET is 2-Deoxy-2-[18F]fluoroglucose (^{18}F-FDG) [23].

The tracer ^{18}F-FDG in the glycolysis process

Chemically, in the tracer ^{18}F-FDG, the positron-emission isotope Flourine-18 (^{18}F) is replaced by a hydroxy group (on C-2 rather than another carbon atom) in the glucose molecule [23]. The ^{18}F-FDG is a glucose analogue, and is, as a normal glucose molecule, transported into cells in need of glucose by a group of structurally related glucose transport proteins (GLUT) [29]. When the glucose and ^{18}F-FDG molecules are inside the cell, the first step towards glycolysis can begin [29]. Here the glucose and ^{18}F-FDG are phosphorylated by hexokinase [29, 30]. A phosphorylated normal glucose molecule will continue along the glycolytic pathway for energy production [29, 30]. In contrast, this is not possible for the phosphorylated ^{18}F-FDG molecule because the C-2 position in the molecule no longer contains an oxygen atom (OH was changed with ^{18}F in order to obtain ^{18}F-FDG) [30]. As a consequence, the tracer ^{18}F-FDG ends up being trapped inside the cell as ^{18}F-FDG-6-Phosphate [29,30].

Uptake of ^{18}F-FDG in different tissue

The use of ^{18}F-FDG, in PET, gives information on the uptake of glucose in different tissues. ^{18}F-FDG is useful in oncology as tumor cells are generally more metabolically active than normal tissue [23,29,30]. Reasons for this are that tumor cells

- have increased number of glucose transporters (especially GLUT 1 and GLUT 3),
- contain highly active hexokinase isoform (type I and II), and
- are often in a condition of relative hypoxia (lack of oxygen) activating the metabolic steps in the more energy demanding anaerobic glycolytic pathway [29,30].

Each of these three reasons result in enhanced glucose, and thus ^{18}F-FDG, uptake [29,30]. Necrotic (dead) tissue, that may be present, inside a tumor causes reduced tumor-to-nontumor ratio (based on the ^{18}F-FDG uptake of different tissues) [30].

The tracer ^{18}F-FDG is not cancer specific and will accumulate in all kinds of tissue consuming high levels of glucose, and thus having high metabolism [29, 30]. There is high uptake of ^{18}F-FDG in areas such as the brain and the heart (due to low cellular glucose-6-phosphatase), and also in sites of hyperactivity (muscular/nervous), active inflammation and in scar tissue [29,30]. Uptake of ^{18}F-FDG in non-cancerous tissue, would therefore interfere with the visualization of only the cancerous tissue in a PET scan. In order to reduce the availability of glucose transporters (GLUT) and lower the ^{18}F-FDG uptake in muscles (otherwise normally prominent), ICRU recommends that patients both fast and remain at rest prior to a PET scan [30].

Positron emission and annihilation

The ^{18}F-FDG-6-Phosphate, trapped inside the cell, is unstable due to radioactive Flourine-18 (^{18}F) with a half-life of 109.8 min. ^{18}F decays through β^+ decay (97 %). The last 3 % is due to electron capture (a parent nucleus captures one of its orbital electrons and emits a neutrino) [23]. During the β^+ decay of ^{18}F,

$$^{18}_{9}F \rightarrow ^{18}_{8} O + e^+ + v + energy, \tag{2.4}$$

a daughter nucleus, $^{18}_{8}O$, a positron, e^+, and a neutrino, v, is emitted [23]. Energy is released during the β^+ decay, in the form of kinetic energy of the released particles [23]. The positron, e^+, is the anti-particle of the electron (with the same mass, but exactly opposite charge), and the neutrino, v, has very little mass and interacts extremely weakly with matter [23].

The positron travels a short distance in tissue (up to 2 mm), slowing down due to interactions (ionization and excitation) with nearby atoms [23]. Only when the positron has lost almost all its energy can ir interact with a nearby electron [23]. This interaction between the positron (almost at rest) and the electron is called an annihilation, as illustrated in Figure 2.4. The annihilation forms two 511 keV γ photons, and this energy is equivalent to the rest mass of a electron/positron. This is a consequence of Albert Einstein's famous equation for the rest mass energy, E,

$$E = mc^2, \tag{2.5}$$

where m is here the total mass of the positron and electron ($m = m_{e^+} + m_{e^-}$) and c is the speed of light in vaccum [22,31]. Conservation of momentum dictates that if both the positron and electron were at rest, at the annihilation site, the two 511 keV γ photons would be emitted in exactly opposite directions [22,23,32].

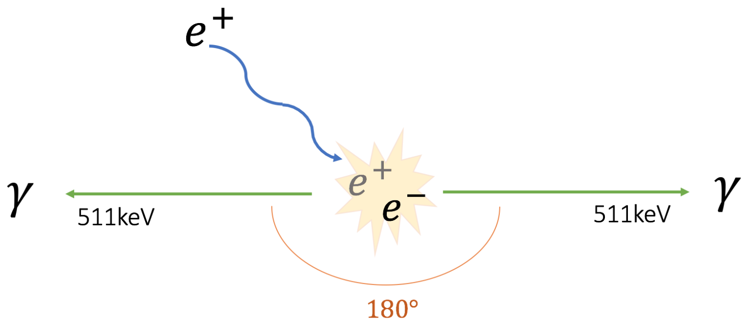


Figure 2.4: A positron, e^+, emitted from a nucleus annihilates, within 1 or 2 mm in tissue, with an electron, e^- , to form two 511 keV γ photons emitted in opposite directions. With permission from Kari Helena Kvandal.

The PET scanner design utilizes that the annihilation photons are created in pairs, have known energy (511 keV) and are traveling in opposite directions. The PET gantry consists of a circular configuration with (multiple) rings of detectors, surrounding the patient bed [31]. The detectors count the annihilation photon pairs that are hit in coincidence during a small time window, τ, hereby referred to as a coincidence event [31]. The scintillation detectors detect the incoming annihilation photons as they interact, by either the Photoelectric effect or Compton scattering, with the scintillator medium (for example Lutetium Oxyorthosilicate (LSO)) [31]. The resulting scintillation (optical light) photons are registered in the photon detector, creating electrical pulses [31]. Every interacting annihilation photon produces a single pulse in the detector [31]. The amplitude of the pulse is determined by the number of scintillation photons reaching the photon detector and any amplification inherent in the photon detector [31].

Figure 2.5 illustrates how the detectors can register scattered coincidences (left), random coincidences (middle) and true coincidences (right). When a photon interacts with tissue (due to Photoelectric effect/Compton scattering), the result would be reduced energy and changed direction of the photon. A random coincidence is detected when two unrelated γ photons enter a pair of detectors at the same time [32]. The line joining the two detected locations is referred to as the line of response (LOR). The LOR is illustrated, in Figure 2.5, as the dashed line for both scattered coincidences (left) and random coincidences (middle), but as solid line (as the path the photons take and the LOR is exactly the same) in the true coincidences (right). Ideally (in the case of true coincidence), the annihilation point should lie somewhere along the LOR.

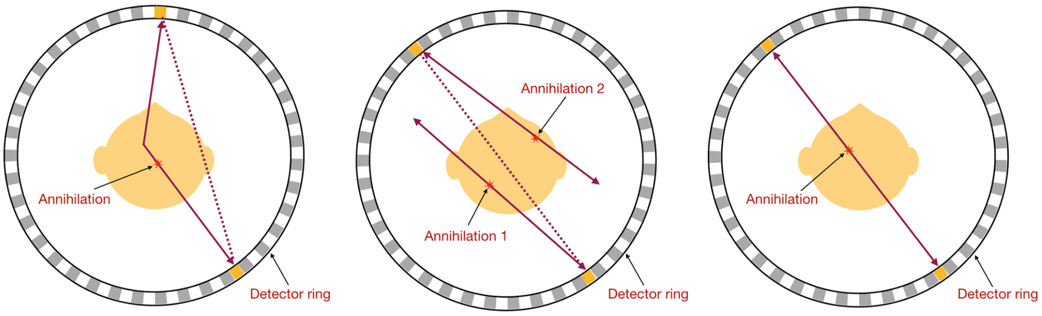


Figure 2.5: The detectors can register scattered coincidences (left), random coincidences (middle) and true coincidences (right). With permission from Kari Helena Kvandal.

While conventional PET seeks to determine along which LOR the annihilation has occurred, the TOF-PET seeks to also determine the position of the annihilation along the LOR [33]. The time of flight principle (TOF) makes a probability distribution of where the annihilation might have occurred along the line of response (LOR), by exploiting the measured difference in arrival time of the two annihilation photons to estimate the point of annihilation, as illustrated in Figure 2.6 [33,34].

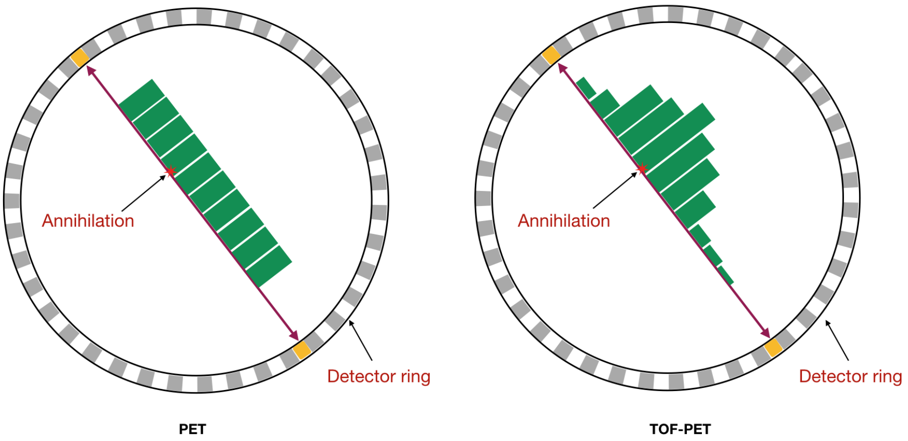


Figure 2.6: Conventional PET seeks to determine along which LOR the annihilation has occurred, giving equal probability distribution along the LOR inside the patient (left). TOF-PET (right) seeks to also determine the position of the annihilation along the LOR, using Time of flight (TOF) principle, exploiting the measured difference in arrival time of the two annihilation photons to estimate the point of annihilation. This resulted in a probability distribution of where the annihilation might have occurred. With permission from Kari Helena Kvandal.

The raw data from a PET scan is a list of counts of the coincidence events obtained along each LOR [23]. The distribution of counts along any direction is a projection of the distribution of radioactivity [23]. The data must be corrected for attenuation effects as well as for accidental and multiple coincidences, dead-time losses and scattered radiation [34]. Then image reconstruction can be performed, and in PET iterative algorithms based on filtered backprojection form the image from all the acquired LORs [34].

Standardized Uptake Value (SUV)

The standardized uptake value (SUV) is the PET value corresponding to the voxel intensity (for each voxel in the PET image stack). The SUV is a simple, semi-quantitative measure of the radioactivity, normalized to the patient's weight and the injected amount of radioactivity. The SUV is defined as

$$SUV = \frac{C}{\frac{A}{W}}, \tag{2.6}$$

where C is the radioactivity [mCi/mL] measured within a region of interest, A is the amount of tracer injection into the patient [mCi] and W is the weight of the patient [g]. The SUV becomes unitless under the assumption that 1 g of body weight is equal to 1 mL, which is the case for water.

As the SUV is sensitive to

- the time between tracer injection and scan,
- whether the patient has been fasting (prior to scan), and
- the patient's weight and body mass index,

misleading SUV can often occur and it is therefore a measure that needs to be treated with caution [23].

SUV would have a value of 1, if the injected tracer was completely and uniformly distributed throughout the body after injection, and if there was no excretion [23]. As previously discussed, this is not the case as ^{18}F-FDG would generally be higher in cancers, but also in other tissues consuming high levels of glucose. SUV is used to assess ^{18}F-FDG in oncology as a measure to separate malignant from benign tissues (in most cases) [23]. In addition, the SUV would be useful for monitoring the response to cancer treatment since higher SUV may be associated with more aggressive tumors [23].

Since SUV, under certain conditions, shows promising differentiation between cancerous and non-cancerous regions, a PET thresholding would be interesting to perform [23]. Several different PET thresholding methods exist for segmentation of target volumes. In the simplest method, the absolute PET thresholding, an absolute

SUV value (for example 2.5) is used as the condition for the segmentation of the cancerous regions.

Consideration of PET/CT

There are multiple advantages with a hybrid PET/CT scanner relative to separate PET and CT scanners. Figure 2.7 displays the same slice of the image stacks for the imaging modalities CT (left), PET (middle) and CT + PET (right, and PET overlayed in the red channel). CT clearly yield good anatomical information, due to the high spatial resolution of around 1x1x2 mm^3. PET on the other hand, had a lower spatial resolution of around 3x3x2 mm^3. However, the functional PET scanner provides a display of the tracer uptake (glucose metabolism), which is useful to distinguish the cancerous regions as they clearly light up (displayed in yellow) in Figure 2.7 [23].

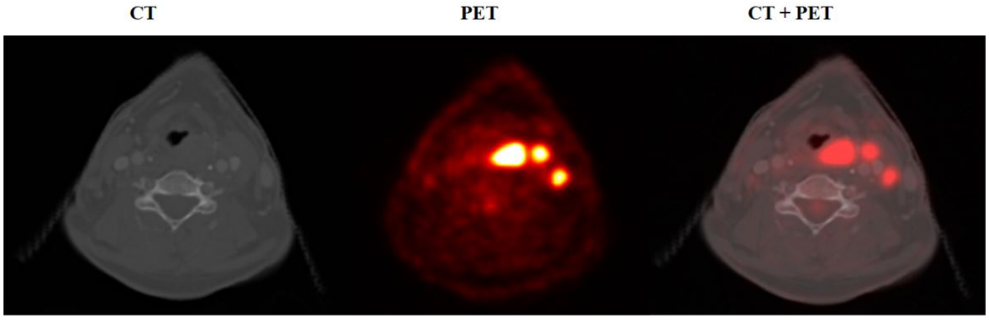


Figure 2.7: Images are displayed for patient number 50 and slice number 111, for the imaging modalities CT (left), PET (middle) and CT + PET (right, and PET overlayed in the red channel).

The lower spatial resolution of PET (compared to CT) is due to

1. the distance the positron travels in tissue before annihilation,
2. the small deviations from 180 °angle difference of the two γ photons created in the annihilation, and
3. the dimensions of the detector crystals (uncertainty in the exact location at which the γ photon first strikes the detector (especially for Compton scatterings inside the crystal, as they would be multiple) increases with the thickness of the crystal) [34].

The CT data improves the spatial resolution by generating accurate localization maps and accurate attenuation correction maps [23]. It is the high photon flux in CT that leads to improved accuracy and reduced noise levels for attenuation measurements [23].

Information from both the PET and CT scanner can make diagnoses more accurate (Figure 2.7, right), by reducing the number of

- false negatives, since CT can detect tumors that might not show up on PET as metabolically active, and
- false positives, as PET can distinguish between malignant and benign tumors that have the same appearance in CT [22].

There are also challenges with the hybrid PET/CT scanner unique for this fused imaging modality. While CT images are taken at specific times (a snapshot image), PET images are acquired over a longer time interval [23]. As a consequence, respiratory motion would then be an intrinsic component of the PET images, and can lead to blurring, due to the averaging over the respiratory cycle [23]. In contrast, CT would then give an image of a specific part of the respiratory cycle, and therefore it would be difficult to achieve perfect registration between the PET and CT images, as structures in these two images might have slightly change position [23].

2.4 Supervised learning in MATLAB

Supervised learning is an approach within machine learning. In a supervised learning algorithm, there is a known set of input data and known output responses of the data [35]. The supervised learning algorithm trains a model to generate reasonable predictions for the response to new data based on evidence in the presence of uncertainty [35]. Supervised learning uses both classification and regression techniques to develop predictive models [35]. In this thesis, the focus lies on using classification techniques because the medical images can be categorized into certain categories, such as cancerous and non-cancerous regions.

In MATLAB there are many Statistics and Machine Learning Toolbox™ algorithms for supervised learning. Most of these use the following steps for obtaining a predictor model:

1. Prepare data,
2. Choose an algorithm,
3. Fit a model,

4. Choose a validation method,

5. Examine fit and update until satisfied, and

6. Use fitted model for predictions.

The first step in preparing the data is to look for outliers and missing data, and decide how to deal with these. The input data matrix, **X**, contains all the observations the model is based upon [35]. Each row and column in **X** represents one observation and predictor, respectively [35]. Therefore, each element in **Y** represents the response to the corresponding row of **X** [35]. Rows containing missing values, so-called `NaN` values in MATLAB, are ignored.

When choosing an algorithm there are trade-offs to consider, such as the speed of training, memory usage, predictive accuracy on new data and interpretability [35].

The model is fitted using the chosen classification algorithm. Choosing the right model takes time, and there are multiple considerations to take [35]. As seen in Figure 2.8, the same data can be grouped, or classified, in several ways leading to an underfit of the data (left), a decent fit (middle) or an overfit of the data (right). Simple models can lead to underfitting of the data due to an oversimplification, for example through inappropriate assumptions [35]. On the other hand, models can be too flexible, leading to an overfit of the data by modeling minor variations caused by noise [35].

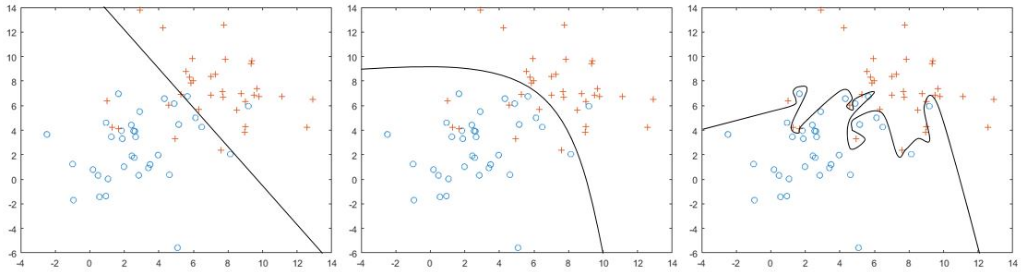


Figure 2.8: A model is fitted to the same dataset using different classification algorithms, leading to underfitting the data (left), decent fitting of the data (middle) or overfitting of the data (right).

The accuracy of the fitted model has to be examined, as a measurement of how well the model performs both on the given dataset and new datasets [35]. One of the main methods to examine the accuracy of the resulting fitted model is to examine the cross-validation error [35]. Cross-validation is a model validation technique based on dividing the dataset into a training set, to train the model, and a test set, to evaluate the predicted model [36]. Since the model is tested against data that was not being used during the modeling process, the cross-validation give an indication of how it will perform on new datasets [35]. Larger datasets tend to yield models that generalize well for new datasets [35].

After validating the model, the fit should be examined and updated until the fitted model is satisfactory for the specific purpose. In this step, the model can be fitted using slightly different model parameters in order to obtain better accuracy, computing speed and being less memory demanding [35]. The model could be fitted using a different classification algorithm and potentially also more classes. Another option, is introducing a cost function to reflect the consequences of oversensitive detection (more false positives) compared to undersensitive detection (more false negatives). For example, it is better to have oversensitive cancer detection than undersensitive cancer detection, as the consequences for an undetected cancer tumor are often far worse than a falsely detected cancer tumor for a patient.

In the last step, the fit is examined and updated until the fitted model is satisfactory. MATLAB has the build-in function `predict(obj,Xnew)`, that predicts the classification response, $Y_{predicted}$, for the fitted model *obj* and the new input data X_{new}.

In this thesis the classification algorithm used is the discriminant analysis, both linear and quadratic, through the build-in MATLAB function `fitcdiscr`. The function `fitcdiscr` offers high prediction speed, high interpretability, low memory use for linear discriminant analysis, but large memory use for quadratic discriminant analysis [35]. A requirement for using this function is that the predictor is numerical and not categorical. In general, the discriminant analysis classifiers are robust and do not exhibit overtraining (when the number of predictors is much less than the number of observations) [35].

Chapter 3

Dataset

3.1 The dataset

3.1.1 Background of the dataset

Oslo University Hospital (OUS) treats around 60% of the head and neck cancer (HNC) patients in Norway [10]. To be able to better understand the complexities of HNC it is essential to have a solid dataset to analyze. OUS has collected data from HNC patient records from the Department of Oncology and also these patient's radiotherapy plans from the Department of Medical Physics [10]. The Department of Oncology, OUS, has collected data consisting of both clinical factors, such as age, sex, stage, nodal status and HPV, and follow-up data, such as local and regional control and metastases [10]. During the eight years, between 2007 and 2015, data have been collected from 256 HNC patients treated at OUS. Due to the extensive dataset in terms of a high number of patients, many clinical factors, follow-up data and digital radiotherapy plans, this material comprises one of the largest cohorts of this sort worldwide [10].

The clinical use of ^{18}F-FDG PET in radiotherapy planning of HNC was implemented at OUS in 2007 [10]. The dataset used in this Master's thesis is from a local retrospective study of HNC patients who have undergone ^{18}F-FDG PET and contrast-enhanced CT in conjunction with radiotherapy [10]. This dataset has not been processed before, and it had to be washed and quality assured before it could be analyzed. After processing and quality assurance of this dataset there were 206 HNC patients left. The reduction of this dataset is described in detail in Section 3.2. The dataset for these 206 HNC patients was collected over a period of around eight years, from 31.10.2007 to 31.07.2015.

3.1.2 The PET/CT scanning

Oslo University Hospital (OUS) has the combined PET/CT scanner Siemens Biograph 16 PET/CT. This scanner ran in three dimensional (3D) PET/CT mode for all patients in this dataset. During the PET/CT scan all the patients were wearing a radiation mask.

Before injection of the tracer ^{18}F-FDG the patients had to fast for a minimum of six hours. The time from the tracer injection to the scanning started was 60 ± 10 minutes for all the patients in the dataset. This can result in differences of up to 20 minutes, which can affect the magnitude of the standardized uptake value (SUV) for the patients. The SUV value given in the dataset is adjusted by the patient's body weight.

The patients were given the same amount (100 ml) of CT contrast agent (Visipaque 320) without taking weight considerations. This leads to the situation where a patient weighing 50 kg would be injected with the same amount of contrast agent as someone with twice the weight. Larger patients have larger blood volumes than smaller patients, and as a consequence the contrast agent administered into the blood compartment dilutes more in the larger patient than in the smaller patient [28]. This would therefore lead to a smaller contrast agent concentration in the blood for the larger patient, which would lower contrast enhancement [28]. As a result, differences in CT voxel intensities between patients can be due to the weight differences. For future data collection, it is an idea to differentiate the amount of CT contrast agent injected to the patient based on their weight, in order to achieve similar contrast enhancement.

The CT contrast agent was injected around one minute before the scan due to rapid uptake in the body. CT contrast agent uptake is higher in tumors than in normal tissue because the contrast agent leaks faster through the more chaotic arterial network in the tumor [24–27]. To convert to Hounsfield units the value 1024 must be subtracted from the CT_{number} given in the dataset.

3.1.3 From `DICOM`-images to a co-registered dataset

Oslo University Hospital (OUS) uses the program IDL (Harris Geospatial Solutions, Broomfield, Colorado, USA) to process the medical images. The DICOM-images from each scan were converted to `uint16` PET and CT image stacks, and text files containing relevant, anonymized data about each scanning. These image stacks were smoothed in IDL with a Gaussian filter. Then the co-registration was performed, by registering the PET and CT image stacks on a common image stack with isotropic voxels of size $1 \times 1 \times 1$ $\mathtt{mm}^3$. The registration was performed with linear interpolation (using the command `congrid` in IDL). Originally, the voxel size of the PET and CT images differed, both between the imaging modalities and also for patients within the same imaging modality.

All the patients in the dataset were anonymized by replacing date of birth and initials with the patient number, P$\star\star\star$, in the folder name and file name. This made it impossible to identify the patients in the dataset. All of these processes in IDL were performed by Professor Eirik Malinen, Department of Physics, University of Oslo.

3.1.4 Primary tumor and lymph nodes contouring

Contour masks identifying (potentially) cancerous regions were created. The structures with names containing *PET* was contoured by the nuclear medicine physicist, while those that did not contain *PET* was contoured by the oncologist.

It was initially decided to only include primary tumors contoured by the oncologist. If no primary tumor contour was provided by the oncologist it was decided to use primary tumor drawn in by the nuclear medicine physicist (if it existed). The reason for this is that the oncologist might have agreed with the nuclear medicine physicist regarding the location of the primary tumor. In this case the oncologist might just have drawn in a clinical target volume (CTV) and not a gross tumor volume (GTV). For the patients with more than one tumor, these were checked and analyzed in detail to decide i) which tumor to choose as the primary tumor, or ii) potentially use both as the primary tumor (as this can occur).

For the lymph nodes all contours were used; both those from the oncologist and the nuclear medicine physicist. This resulted in a UNION-volume between the different lymph nodes that was larger than the lymph node volume contoured by either the oncologist or the nuclear medicine physicist.

The chosen contouring was considered the ground truth in the model. This is naturally not accurate, as there are variations in the contouring, both between the oncologist and nuclear medicine physicist per patient and also intra-variances between different oncologists and nuclear medicine physicists [38].

3.1.5 Biopsies only of tumors

Biopsies of the tumors had been taken, but not of the lymph nodes. Therefore, one can be sure that the contoured tumors contain cancer cells. Since biopsies of the lymph nodes were not performed, one can not be sure whether the contoured lymph nodes contained cancer cells. For instance, infectious lymph nodes would also be displayed as metabolically active in a PET/CT image stack [29, 30]. It is therefore important to take biopsies of the lymph nodes in order to rule out or verify the existence of cancer cells in all contoured structures. In practice, however, is this a demanding and invasive procedure for the patient, especially in the situation with multiple lymph nodes. A compromise here, could be to take a biopsy of only one lymph nodes (the most likely cancerous lymph node), to be able to assess if the cancer has spread from the primary tumor to the lymph system. This is especially important as cancer cells can spread through the lymph system to other parts of the body [39].

3.1.6 A better and more quality assured dataset

The dataset was generated in two rounds. In the first round, there were 256 patients. Of these, 226 patients had a patient folder with three separate text files containing information about the performed PET/CT scanning in addition to the PET and CT image stacks. In total, there were 30 so-called missing patient folders. There are several reasons for patients not having a folder. Patients were excluded if there were problems with the co-registration between PET and CT. Patients with incorrect SUV or CT_{number} would also be excluded. There could also be an error of some kind in the process from the `DICOM`-images to an anonymized, co-registered dataset.

As a number of challenges with the dataset were detected, it was decided to generate the dataset again. In this second round, all the data was generated simultaneously, to assure no discrepancies in the number of patients between the different files in the dataset. The PET/CT images were checked by Professor Eirik Malinen after the co-registration [26]. For some patients this co-registration was poorly executed, as the images were shifted in relation to each other. All patients with discovered errors, in any of the files in the dataset, were removed [26]. The new dataset consisting of 226 patients is thereby a better and more quality assured dataset than the previous one.

3.1.7 Files in the dataset

Each of the 226 patients in the new dataset have a patient folder containing the following files

1. P***_struk.txt (henceforth referred to as the struk-file): name of the GTV-structures and the voxel indices,
2. P***_ info.txt (henceforth referred to as the info-file): contain information about the dimensions of the images and the maximum PET and CT value,

3. P⋆⋆⋆_ bilde.jpg (henceforth referred to as the jpg-image): is a fused PET and CT image with structures drawn in for three different slices through the patient.

In addition to these three files, the dataset also consists of PET and CT image stacks from the PET/CT scan for each patient.

An example of the jpg-image is seen in Figure 3.1. The cancerous regions shown in these slices are drawn in, as seen by the white lines, by the oncologist and the nuclear medicine physicist. This figure also displays the names of the GTV-structures, which are given in the struk-file, together with the voxel indices for all the cancerous regions for each patient. Notice, in Figure 3.1, that the oncologist detected three lymph nodes and one primary tumor, and that the nuclear medicine physicist detected four lymph nodes and one primary tumor. The structure called *GTV union* is a structure consisting of both the primary tumor and the lymph nodes.

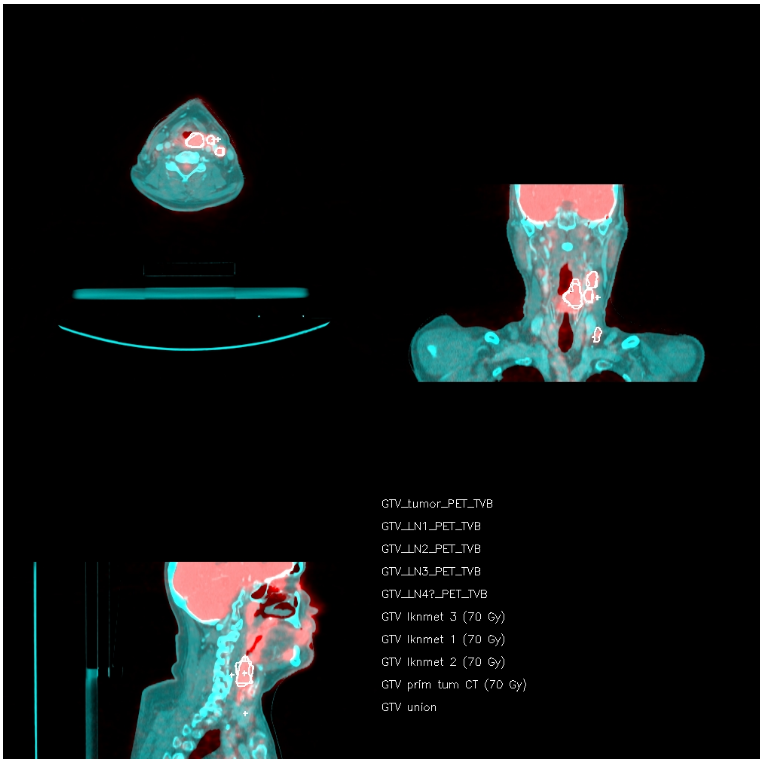
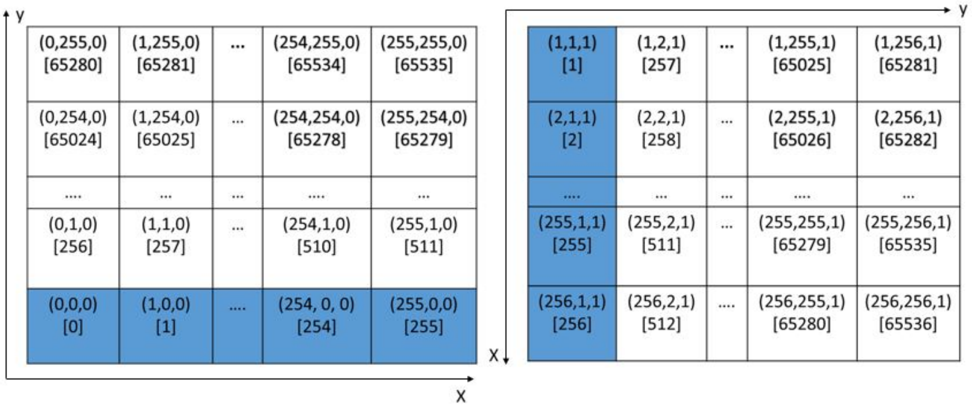


Figure 3.1: A fused PET and CT image with cancerous regions drawn in for three different slices through patient number 50.

3.2 Washing and organization of the dataset

3.2.1 Conversion from IDL to MATLAB

The smoothed, co-registered PET and CT images were imported into MATLAB. MATLAB and IDL have different indexing, which is illustrated in Figure 3.2. In this figure the indexing in IDL and MATLAB are given for a matrix with the dimensions 256×256×*nz*; where *nz* is the first slice of the matrix.



(0,255,0) [65280]	(1,255,0) [65281]	...	(254,255,0) [65534]	(255,255,0) [65535]
(0,254,0) [65024]	(1,254,0) [65025]	...	(254,254,0) [65278]	(255,254,0) [65279]
....	...	...	....	...
(0,1,0) [256]	(1,1,0) [257]	...	(254,1,0) [510]	(255,1,0) [511]
(0,0,0) [0]	(1,0,0) [1]	....	(254, 0, 0) [254]	(255,0,0) [255]

(1,1,1) [1]	(1,2,1) [257]	...	(1,255,1) [65025]	(1,256,1) [65281]
(2,1,1) [2]	(2,2,1) [258]	...	(2,255,1) [65026]	(2,256,1) [65282]
....	...	...	....	...
(255,1,1) [255]	(255,2,1) [511]	...	(255,255,1) [65279]	(255,256,1) [65535]
(256,1,1) [256]	(256,2,1) [512]	....	(256,255,1) [65280]	(256,256,1) [65536]

*Figure 3.2: A matrix with dimensions 256×256×*nz*, where* nz *is the first slice of the matrix, is illustrated for the programs* IDL *(left) and* MATLAB *(right). The voxel conversion from* IDL *to* MATLAB *requires that the value one be added to all the voxel indices, and all images and masks (based on the contoured voxel indices) to be rotated by 90 degrees clockwise.*

In IDL, the origin (0,0,0) is placed in the lower left corner and has the voxel index of zero (Figure 3.2). The voxel indexing proceeds to the right and then upward row by row. In contrast, the origin (1,1,1) in MATLAB has a voxel index of one and a location in the upper left corner. MATLAB counts the voxel indices by going through every row in a column and then does the same for the next column, as is illustrated in Figure 3.2. It is thereby apparent that the voxel conversion from IDL to MATLAB requires that the value one be added to all the voxel indices of the images and the masks, based on the contoured voxel indices, to be rotated by 90 degrees clockwise.

In the PET and CT image stacks, it was necessary to check if all the patients had their nose pointing upwards as this will ensure the same anatomical location of the voxels as the created, contoured masks. An example of PET and CT images is illustrated in Figure 3.3. Notice that the brain is very metabolically active from the PET image; this is because the brain consumes relatively large amounts of sugar compared to normal tissue [29, 30]. Normally symmetrically, metabolically active areas in the head and neck area are benign, despite the high SUV [23]. In a few cases, malignant tissue can hide behind these symmetric active areas [23].

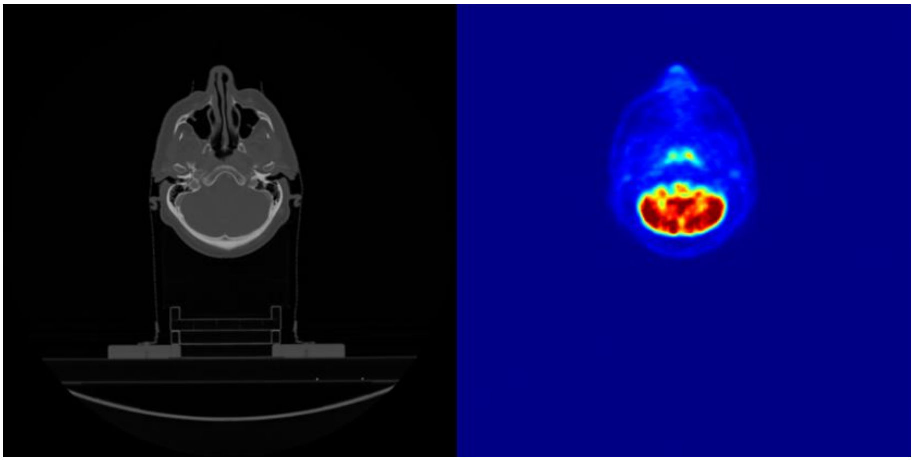


Figure 3.3: The CT (left) and PET (right) image stacks for slice number 205 for patient number 50. Notice that the nose is pointing upwards, and that the brain is significantly more metabolically active than the normal tissue.

3.2.2 Patients injected with a contrast agent

It was decided to only include patients given CT contrast agent, as these contrast agents results in higher relative intensities between normal and malignant tissue. The contrast agent results in an increase in contrast enhancement in the CT image, and the contrast enhancement is directly related to the amount of contrast agent injected into the patient [28]. As a consequence, even if the CT image stacks were auto-scaled (to a mean of zero and standard deviation of one) the voxel intensities (for the specific structures) in the CT image would be different depending on whether the patient was injected with CT contrast agent or not. To ensure consistency in CT_{number} between patients, all patients not given contrast agent were excluded from further analysis. Of the 226 patients in the new dataset that have all the files (described in subchapter 3.1.7), there are 210 patients left after the condition of applied CT contrast agent was met.

3.2.3 Naming consistency

The washing and organization of the dataset became a time-consuming task during this thesis, especially considering the lack of consistency when naming the GTV structures. This is because the dataset was collected over almost eight years and several different oncologists and nuclear medicine physicists were involved in the contouring. For instance, twenty conditions were used to find the primary tumor as different oncologists and nuclear medicine physicists referred to the tumor by different names. In order to find the names of all the lymph nodes, eight conditions were needed; searching for names containing *kn*, *lk*, *LK*, *LN*, *ln*, *Knute*, *l.k* and *lg*.

3.2.4 Zero or two primary tumors

Nine patients were found not to have a primary tumor. This can, according to Eirik Malinen [26], be because

- the primary tumor is unknown (patients can have cancer-infiltrated lymph nodes without a known place of origin),
- the primary tumor has been surgically removed prior to operation (but this is very rarely done), or
- the primary tumor was not given the correct name (and it therefore not selected out from the database).

Twenty-seven patients had a primary tumor drawn in by the nuclear medicine physicist, but not the oncologist.

Eight patients had two primary tumors (P038, P051, P182, P186, P202, P205, P206 and P246). The decision on which masks to use as the primary tumor were taken after investigating the masks and the PET/CT images (jpg-file). For example, in the case of patient number 202 and 206, one of the masks of the primary tumor gave an indication of non-connected structures; which had to be a combination of tumor and lymph node(s). These structures were used in the planning of radiation therapy to decide where the radiation should be aimed [26]. For six out of the eight patients with two primary tumors, one tumor was selected as the primary tumor. In the case of the two last patients with two primary tumors, P051 and P186, it was not possible at present time to determine which was the primary tumor. This was the case (undetermined primary tumor) also for the two patients, P235 and P238, having one detected tumor. Even though the four patients, P051, P186, P235 and P238, had information about the lymph nodes, they were excluded from the dataset due to uncertainties regarding the primary tumor (as described above). This exclusion is important as an unclassified primary tumor would result in misclassification as non-cancerous region in the images and would thereby negatively affect the classification model.

Forty-two patients had no contoured lymph nodes. Patients with only lymph node(s) and no tumor were kept in the dataset as long as no tumors were drawn in their PET/CT image. For one patient, P177, the given name of the primary tumor did not correspond with the twenty conditions used to extract the primary tumors in the MATLAB code. The primary tumor, for this patient, was added to the dataset.

GTV-structures, for both primary tumor(s) and lymph node(s), have been added and removed for selected patients in the dataset (as described above). In the end, twelve and forty-five patients did not have a primary tumor and lymph node(s), respectively. All of the 198 patients with a primary tumor now had one and only one primary tumor.

3.2.5 Organization of the dataset

Patients injected with CT contrast agent were selected out from the patients in the new dataset. All of these patients had all the files: struk-file, info-file, jpg-image, PET and CT image stacks (described in 3.1.7). Organizing the dataset focused on collecting all the information from all the patients together into one table or structure for the two different files, the struk-file and the info-file, respectively. For the info-file, a matrix containing patient number, image dimensions in x, *nx*, y, *ny*, and z, *nz*, *maxSUV* and *maxCT* for all patients was created.

A pipeline from P$\star\star\star$_struk.txt to the total mask, based on the contouring of both primary tumor and lymph nodes, is illustrated in Figure 3.5. For the struk-file, a cell array with a hash table in each cell was created. The hash table mapped the keys, the name of the GTV structures, to the values, the corresponding voxel indices (Figure 3.4). The cell index corresponds to the index in the array consisting of all the patients in the dataset. This organization of the cancerous regions made it possible to find the values to the desired keys for each patient. The cell array with the hash tables made it convenient and simple to extract information about the relevant GTV structures.

MATLAB structures for primary tumors and lymph nodes were created so that each structure contained the patient number, the name of the GTV structure and the corresponding voxel indices for the primary tumors and lymph nodes, respectively.When the MATLAB structures for primary tumors and lymph nodes were created, the value one was added to all the voxel indices, to take care of the difference in starting voxel index in MATLAB compared to IDL. The created masks, based on the voxel indices for the primary tumors and lymph nodes, were rotated by 90 degrees clockwise. These steps are essential in order to make the transition from IDL to MATLAB correct, as discussed in Section 3.2.1.

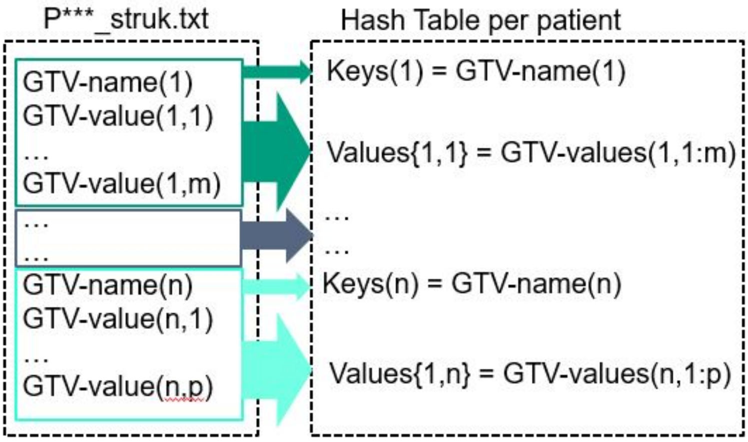


Figure 3.4: From P⋆⋆⋆_struk.txt file to a hash table structure. The hash table mapped the keys, the name of the GTV structures, to the values, the corresponding voxel indices. In this example, m and p designate the number of voxels for the first and last (n) GTV structure (either primary tumor or lymph nodes), respectively.

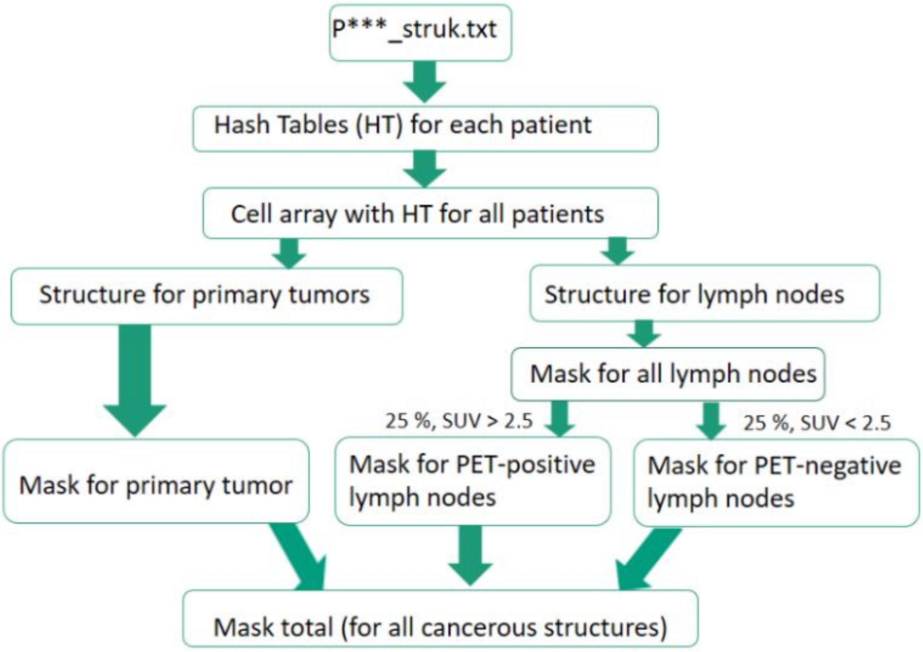

P***_struk.txt
Hash Tables (HT) for each patient
Cell array with HT for all patients
Structure for primary tumors
Structure for lymph nodes
Mask for all lymph nodes
25 %, SUV > 2.5
25 %, SUV < 2.5
Mask for primary tumor
Mask for PET-positive lymph nodes
Mask for PET-negative lymph nodes
Mask total (for all cancerous structures)

Figure 3.5: This is an illustration of the pipeline from the P⋆⋆⋆_struk.txt-file to the contoured masks. On the basis of the P⋆⋆⋆_struk.txt file, hash tables were created for all patient and later gathered in a cell array. Then MATLAB *structures for the primary tumor and lymph nodes are created. Based on the information of voxel indices in the* MATLAB *structures all the masks, showing primary tumor and lymph nodes contours, were created.*

Lymph nodes can be either PET-positive or PET-negative depending on how metabolically active they are, given by their standardized uptake value (SUV). The separation of these groups of lymph nodes was based on testing whether over a quarter of the voxels in a lymph node had a SUV value greater or less than 2.5 [26]. The PET-positive and PET-negative lymph nodes were given a unique value between 0 and 1; the values 0.8 and 0.5 were chosen, respectively. This gave greater flexibility for future classification of different structures, in addition to proving valuable information for the detection of overlapping structures.

All the masks, for primary tumors and lymph nodes (both PET-positive and PET-negative), were added together to form one mask for each patient (henceforth called total mask). For distinct and non-overlapping regions, the voxel intensities of the total mask should consist of the value 0.5 in case of PET-negative lymph nodes, the value 0.8 in case of PET-positive lymph nodes, the value 1 for primary tumors and the value 0 for normal tissue and background. Therefore in the case of non-overlapping structure within the total mask, there are maximum four unique values (0, 0.5, 0.8 and 1) and a maximum value of one.

An overlap for the lymph nodes was expected, as these are contoured by both the oncologist and the nuclear medicine physicist. In total for all the masks, there were 205 incidences regarding a total of twelve new, distinct values (different than the four expected unique values 0, 0.5, 0.8 and 1). More than half of these structures (116/205) were due to an overlap between two different PET-positive lymph nodes, and there were fourteen instances of overlap between a single PET-positive and PET-negative lymph node. Overlaps between both types of lymph nodes and the primary tumor also exist, and these were most likely due to misnaming of clinical target volumes. In these cases, the voxel indices for the primary tumor also include voxel indices for the lymph nodes, as illustrated in Figure 3.6.

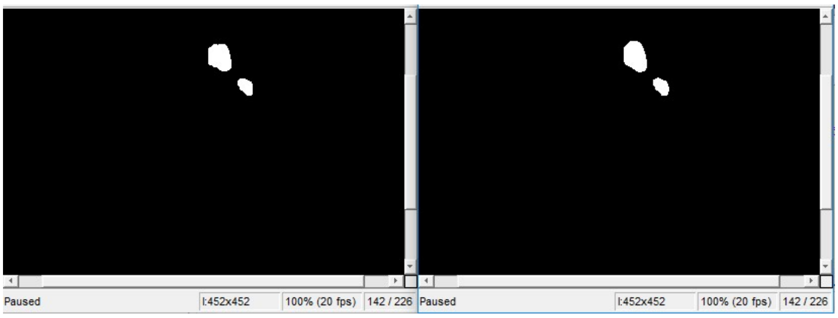


Figure 3.6: The masks for primary tumor (left) and PET-positive lymph nodes (right), for patient number 18 and slice number 142 is illustrated. Here the structures seem to have a high degree over overlap and calculations show that 11525 voxels (out of a total of 25742304 voxels in the image stack) overlap between primary tumor and PET-positive lymph nodes.

Visual inspection of this figure clearly shows overlap between primary tumor and PET-positive lymph nodes, and calculations (of the entire image stack **X**) give an overlap of 11525 voxels (out of a total of 25742304 voxels). For forty-nine instances, multiple structures overlap without there being a clear indication of the structures involved in the overlap (on the basis of the voxel intensity values). All of the overlapping structures needs to be quality assured before it is possible to perform more than a binary classification of the structures involved.

3.2.6 Cropping of the masks, PET and CT images

A cropping, of the total masks and the PET and CT images stacks, results in a higher percentage of cancerous voxels in the image stacks and therefore in a better balance between the classes. Therefore, two different types of croppings were performed, as illustrated in Figure 3.7. The two main cropping methods are local and global cropping.

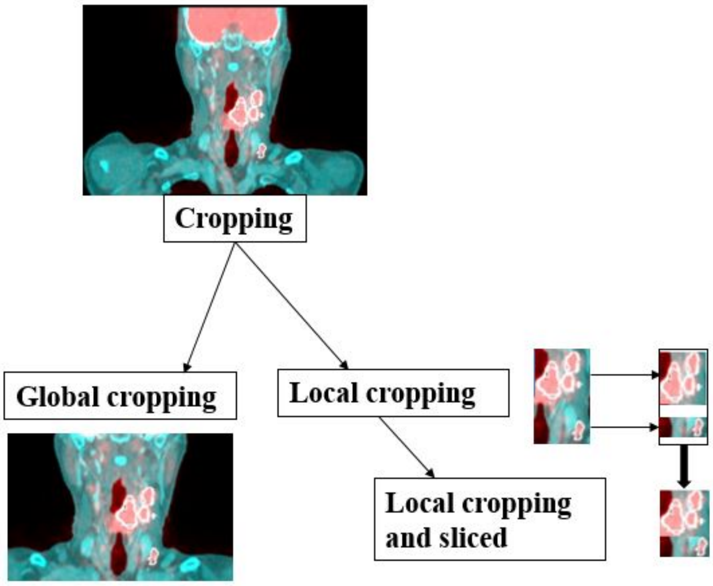


Figure 3.7: The total masks and the image stacks can be cropped in different ways. The first choice is whether to do a local or global cropping, and the second choice is whether to remove slices not containing primary tumor or lymph nodes.

An voxel index in the image stack is converted into it's coordinates in the x-, y- and z-plane. The voxel indices with voxel intensities larger than zero gives the location of the cancerous regions in the image stack, and these x, y and z coordinates form the basis of the cropping. The local cropping is based on the minimum and maximum of these x, y and z coordinates for each patient. The global minimum and maximum were calculated by finding the global minimum and maximum in the x-plane, y-plane and z-plane from all patient's local minimum and maximum x, y and z coordinates. The PET and CT images were cropped in the exact same way as the mask for the two cropping methods.

The global cropping was based on a box with the size of the biggest difference between minimum and maximum values in the x- and y-plane. For each patient this box was placed in the middle of the local minimum and maximum in both the x- and y-plane. An edge of 1 mm was added around this box, to make sure that the voxel indices of all cancerous regions were inside the box, for all patients. For each patient the local minimum and maximum z-plane were used for the cropping in the z-direction. The local cropping was based on local minimum and maximum in all three planes.

The relation between the number of voxels of the cancerous region and the total number of voxels in the image stack were approximately 3% for global cropping, and approximately 9% for local cropping. The method with local cropping is preferred as this roughly triples the relation between the number of voxels of cancerous region and the total number of voxels. The method, *Local cropping and sliced* (illustrated in Figure 3.7), is a way to further reduce the dataset by excluding z-slices where there are no cancerous regions. The relation between the number of voxels of cancerous region and the total number of voxels were approximately 13% for local cropping with z-slices without cancerous regions sliced out of the image stacks. There were large variations regarding the fraction of cancerous voxels among patients, in the range of 3% to 59%, for locally cropped and sliced image stacks.

3.3 Software and computer

The software MATLAB (version R2017a, The Mathworks Inc., Natick, MA) with the packages Statistics and Machine Learning Toolbox™ and Image Processing Toolbox™ was used in this thesis. MATLAB was used to import and organize the dataset, make calculations, perform LDA and QDA classifications and statistical analysis.

The models were run, in MATLAB, on the computer LENOVO ThinkStation P910 with the processor Intel® Xeon® E5-2600 v4 (with 20 cores, 3.5 GHz per CPU) and the operating system Ubuntu. This computer had a RAM of 125.8 GiB, which allows for processing and analysis of large datasets. An external Toshiba hard-disk, with 1.81 TB memory, were used to store the large dataset and all the scripts, functions and results from the MATLAB programming.

Chapter 4

Modifications of the autodelineation program

4.1 Background of the autodelineation program

A tumor autodelineation program, which was based on MR images of cervical cancer, was developed by former PhD student at NMBU, Turid Torheim [4]. This autodelineation program builds on classification methods in supervised learning, see Section 2.4 on page 17. The aim of the autodelineation program is to classify the voxels in the image as tumor or not tumor [4]. The steps in the autodelineation program are illustrated in Figure 4.1.

Input and output to the program are both images. The input images are medical images from the scanning, while the output images are showing the predicted cancerous regions either as cancer probability maps or as binary maps (cancer/not cancer). The steps in the autodelineation program consist of image pre-processing, image unfolding, voxel classification, post-processing of resulting predictions and performance measure calculation.

In Torheim's PhD Thesis the input images were MR images of cervical cancer and the delineation was performed by two radiologists [4]. These delineations were considered the ground truth, and were used as the basis for training and assessment of the classification [4]. If necrotic areas (dead tissue) in the tumor were present, they were included in the delineation [4]. The MR images were dynamic, consisting of several time steps, and static (only one time step) [4]. Torheim used images for 78 patients for three different MR representations (T2w (static), T1w (static) and DCE-MRI (dynamic)), and she downscaled the T2w and T1w images to match the size of the DCE-MR images [4].

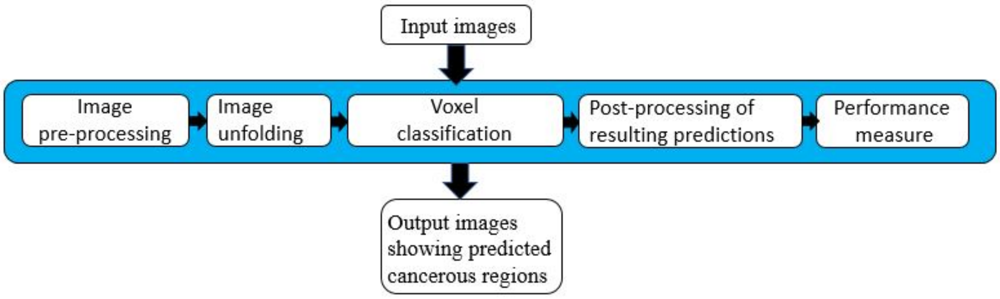

Input images
Image pre-processing
Image unfolding
Voxel classification
Post-processing of resulting predictions
Performance measure
Output images showing predicted cancerous regions

Figure 4.1: The build-up of the autodelineation program. Input to the program is the medical images from the scanning. These images are pre-processed to compensate for intensity differences among patients. Then the images are unfolded. The classification of each voxel is performed on the basis of the unfolded mask and the chosen classification algorithm. The mask is based on the voxel indices given by contouring by the physician. Insignificant noise is removed by post-processing the resulting predictions. In the final step the program's performance is evaluated by comparing the output images, showing the predicted cancerous regions, with the masks based on the physician's contouring.

Then these images were cropped and further reduced, by only including the image slices in which the radiologist's identified tumor tissue, as shown in Figure 4.2(a). By cropping and slicing the images, the balance between the two voxel classes (Torheim used in her paper), tumor and non-tumor, improved.

The already cropped input images were pre-processed by auto-scaling the images to a mean of 0 and a standard deviation of 1. The auto-scaling is performed to compensate for intensity differences between patients (for the T2w and T1w images) and to maintain the intensity increase between time steps (for DCE-MR) [4].

The next step in the autodelineation program is image unfolding, as illustrated in Figure 4.2 (b), (c) and (d). The voxels in the image stacks (for T2w, T1w and DCE-MRI) are defined by their position in the image, the x, y and z coordinate, and a specific voxel index. The image stack is unfolded with increasing voxel indices, from voxel index 1 to voxel index n (n is the total number of voxels in an image stack). To obtain information about the spatial relationships between voxels, the eight closest neighbors were included, in the unfolding, for T2w and T1w images (illustrated in Figure (4.2(c))). The voxel intensity is, for T2w and T1w images, represented in each voxel. For DCE-MRI, the 13 first time steps were included in the unfolding process, and here the voxel intensity represents its time-intensity curve. The constructed data matrix $\mathbf{X}$ consists of the unfolded T2w and T1w images (both with eight neighbors) and the unfolded DCE-MRI images (with 13 time steps), as seen in Figure 4.2(d).

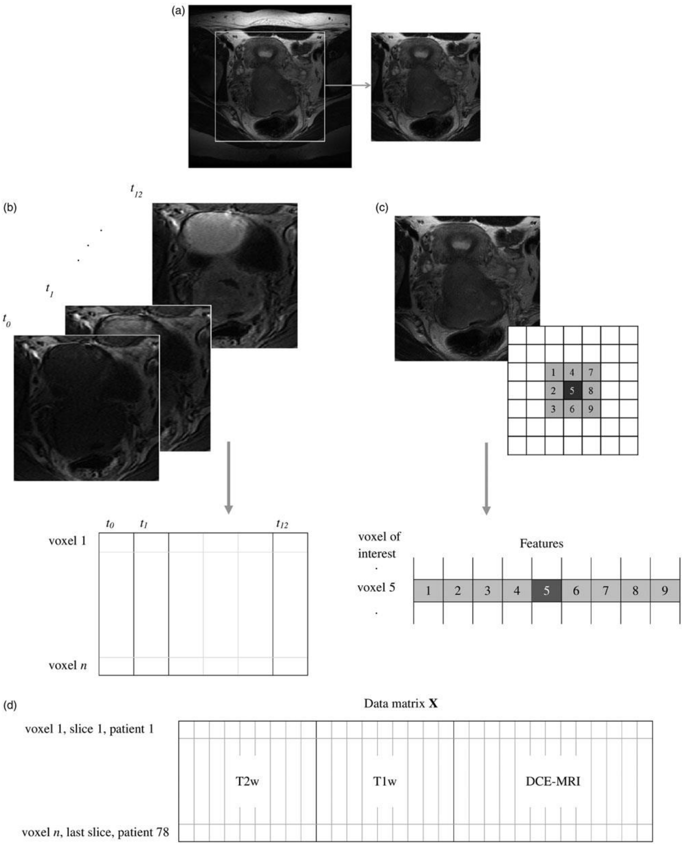


Figure 4.2: Conversion of the MR images into a data matrix **X** *used for voxel classification. (a) Edges were removed from all images. (b) For the DCE-MRI series, each voxel was represented by the intensities of the 13 first time steps in the image series. (c) The T2w or T1w images were represented by either only their intensity, or by their intensity (dark gray) and the intensities of the eight closest neighbor voxels (light gray). (d) The data matrix* **X** *for the model based on T2w images with eight neighbors included, the T1w images with eight neighbors included and the 13 first DCE-MRI time steps. With permission from Turid Torheim, from her PhD thesis [4].*

Then the voxel classification can begin. The voxels were classified using Fisher's Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) [4]. Fishers's LDA and QDA seeks to find a line, linear or quadratic, respectively, that maximizes the separability among two (or more) classes. The models (based on either LDA and QDA) were trained using the radiologist's contours, to classify all the voxels in the data matrix **X** as tumor or non-tumor.

Classification produced the probability maps, which gave the probability of each voxel belonging to the tumor. The probability threshold were set to 50 %; all voxels with probabilities (of the voxel being a tumor) higher than 50 % were assigned as tumor. The probability maps were converted into a binary image on the basis of the chosen probability threshold. These binary images were post-processed by an in-plane morphological operation that removed voxel volumes consisting of less than 100 voxels [4].

The model's performance, on new sets of images, was estimated using leave-one-out cross-validation [4]. The images of one patient were removed, and the LDA or QDA model was trained on the remaining *N*-1 patients, as illustrated in Figure 4.3. Then the resulting model was used for segmentation of the images from the left-out patient [4]. This process was repeated until all patients were left out once. A benefit with this validation method is the similarity with the situation in a clinic; a model trained on images from earlier patients will be used for the diagnosis of a new patient [4].

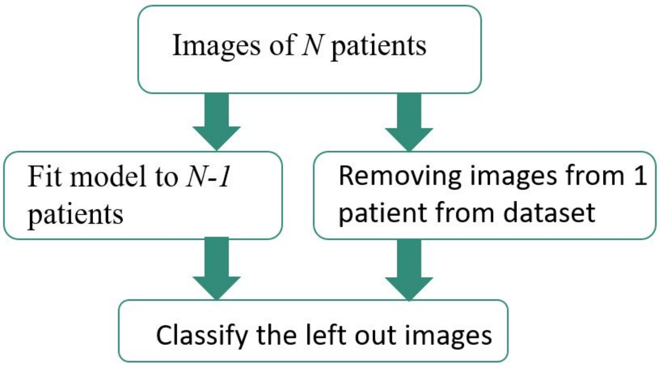

Images of *N* patients
Fit model to *N-1* patients
Removing images from 1 patient from dataset
Classify the left out images

Figure 4.3: Leave-one-out cross-validation. The model was trained on N - *1 patients, and the resulting model was used for segmentation of the images of the left-out patient. With permission from former PhD student at NMBU, Turid Torheim.*

In the final step, the program's performance is evaluated by comparing the mask based on the radiologist's contouring (ground truth) with the mask based on the LDA/QDA classification model (the program's output image). The overlap between these two masks was evaluated using Dice similarity coefficient (*DSC*) and the Kappa statistics *K* [4]. In addition, the assessment of the model performance for classifying each of the two voxel groups (tumor and non-tumor) was also based on the the area under the receiving operator curve (*AUC*), sensitivity (S_e) and specificity(S_e) [4]. N-way ANOVA (with significance level p < 0.05) was used to test the main effect and two-factor interaction of different model parameters, such as image type and spatial information [4].

4.2 Modifications of the autodelineation program

4.2.1 Input images

In this thesis, the PET/CT images of head and neck cancer, both PET, CT and PET/CT images stacks, are used as the input images to the autodelineation program. These images were locally cropped and either sliced and not sliced, as discussed in detail in Section 3.2.6. In addition, these images could be unfolded with zero neighbors and eight neighbors (both sorted and unsorted).

4.2.2 Preprocessing

The first step of the program is to pre-process the input images in order to compensate for intensity differences between patients and the two imaging modalities. The chosen pre-processing method is auto-scaling with an average of 0 and a standard deviation of 1. The auto-scaling had to be performed for the PET + CT image stacks since the intensity differences between these two imaging modalities are significantly different. By auto-scaling the CT images, subtracting the value 1024 from the CT_{number} given in the dataset in order to obtain Hounsfield units is no longer necessary, as demonstrated in Appendix A, for Equations 1 to 6.

Standardization or auto-scaling is a method yielding z scores,

$$z_{scores,i} = \frac{x_i - E(x)}{\sigma(x)}, \tag{4.1}$$

where *E(x)* and $\sigma(x)$ is the expected value and standard deviation, respectively, of the total sample variable array x. x consists of the sample variables x_i, for i = 1, 2, ...,n-1, n. In this Thesis, the x would correspond to the image stack(s) containing all the voxel intensities, and x_i would be the voxel intensity of a specific voxel i. In the MATLAB computation of the z scores, potential `NaN` values are omitted using the option `'omitnan'` inside the z scores function.

4.2.3 Unfolding

In the unfolding step the images are folded out to a matrix consisting of the intensity value to all the voxels of all the patients in a systematic manner. For zero neighbors this matrix consists either of two columns, when PET/CT data is used as input, or only one column, when either PET or CT is used. Then the images are unfolded out to a matrix where each row consist of the intensity value to a specific voxel in the image and to a specific patient. For PET and CT images, a row in the matrix will consist of the SUV and the CT_{number}, respectively. In the case of PET/CT data, the first column of the matrix will consist of the CT_{number} and the second column of the SUV.

When the unfolding includes the 8 nearest neighbors, the unfolding is the same as illustrated in Figure 4.2 on page 37. Then instead of one column, there are nine columns for each imaging modality. Therefore, there are eighteen columns for the combination of both imaging modalities, PET/CT. The (nine or eighteen) columns are either sorted in descending order (from first to last column) or unsorted.

4.2.4 Classification

The classification of each voxel is performed on the basis of the unfolded mask and the chosen classification algorithm. The mask is based on the contouring performed by the oncologist/nuclear medicine physicist of primary tumor and lymph nodes (located in the head and neck), and represented by their voxel indices in the image stacks. The applied classification algorithms are Fisher's Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

For the head and neck dataset it will be beneficial to include the cancer infiltrated lymph nodes in addition to the primary tumor. The main reason for this is that the lymph nodes need to be considered in the treatment plan, as cancer cells can spread through the lymphatic system to other regions of the body [39]. The lymph nodes can be PET-positive or PET-negative depending on how metabolically active they are, which their SUV give an indication of (as discussed in Section 2.3.2). The separation of these groups of lymph nodes is based on whether over a quarter of the voxels in the lymph nodes have a SUV higher than 2.5 or not, respectively [26]. This leads to the following possibilities for the number of classes:

- Two classes
  1. Lymph nodes (both PET-positive and PET-negative) and the primary tumor
  2. Everything else
- Three classes
  1. Primary tumor
  2. Lymph nodes (both PET-positive and PET-negative)
  3. Everything else

- Four classes
    1. Primary tumor
    2. PET-positive lymph nodes
    3. PET-negative lymph nodes
    4. Everything else

The classification step of the program therefore does not only depend upon the chosen classification method, but also on how many classes the classification should be build on. Due to the challenges with the washing of the dataset and the time limitation of this Master's Thesis, as described in the Section 3.2.5, only a binary (two class) problem is performed. Multiple structures need to be manually checked for each of the 206 patients in the dataset, making sure that, for example, the GTV-structures were given the correct name and that overlapping structures are assigned to the correct class.

LDA and QDA

The models (based on either LDA and QDA) are trained using the oncologist's and nuclear medicine physicist's contours, to classify all the voxels in the data matrix, **X**, into the two classes

1. Primary tumor and lymph nodes,
2. Everything else.

Fishers's linear and quadratic discriminant analysis (LDA and QDA) seeks to find a boundary, linear or quadratic, respectively, that maximizes the separability among two (or more) classes. This is equivalent to finding the combination of predictors giving the maximum separation between the centers of the data (between-class variance) and at the same time minimizing the variation within each class of the data (within-class variance) [40]. Mathematically, LDA and QDA minimizes the total probability of misclassification by assuming that the data can be truly separated by hyperplanes or quadratic surfaces, respectively [40].

Fisher's LDA determines linear combinations of the predictors to maximize the signal-to-noise ratio [40]. When **B** represent the between-group covariance matrix and **W** represent the within-group covariance matrix, the LDA problem seeks to find the value of b that maximizes [40]

$$\frac{b'\mathbf{B}b}{b'\mathbf{W}b}. \tag{4.2}$$

The eigenvector, or linear discriminant, corresponding to the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ is the solution to this optimization problem (given in Equation 4.2) [40]. The same optimization in new directions, uncorrelated with the previous discriminants, leads to the subsequent discriminants [40].

While LDA assumes the same covariance matrix for all the classes, the covariance matrix for each class is estimated separately in QDA [40, 41] As a consequence, if there is reason to believe that there are different within-class variations, and the classes therefore should have their own covariance matrix, QDA should theoretically perform better than LDA. With QDA the decision boundaries become quadratically curvilinear in the predictor space [40]. As all the class-specific matrices are utilized in QDA, the inverse of the matrices must exist, the number of predictors must be less than the number of cases within each class, and there must not be any collinearity between these in-class predictors.

4.2.5 Postprosessing

Post-processing is needed to remove insignificant noise in the classification results. This is done by applying an in-plane morphologic operation on the binary images to remove all elements containing less than 10 voxels. It was decided to remove elements smaller than 10 instead of 100, as lymph nodes and not just primary tumor, are now included in the contouring.

4.2.6 Performance measure

In the final step, the program's performance is evaluated by comparing the program's images, the output images, with the masks based on the oncologist's/nuclear medicine physicist's contouring. This is the results and validation step of the program. The probability threshold was set to 50 %, converting the probability maps into binary output images.

The oncologist/nuclear medicine physicist analyzes medical images from a patient to diagnose the patient as being sick or healthy. In this case being sick is equivalent of having cancer tumor(s) and (potentially) cancerous lymph nodes, and healthy refers to the patient not having any cancerous regions. To be diagnosed as sick is referred to as being *positive*, while *negative* refers to being diagnosed as healthy. The terms *true* and *false* refers to a correct and an incorrect diagnosis, respectively. Therefore, in clinical diagnostics there are four possible outcomes [34], as illustrated in Table 4.1.

Table 4.1: Four possible outcomes when analyzing medical images

	Actual class	
Predicted class	*Sick*	*Healthy*
Sick	True positive (*TP*)	False positive (*FP*)
Healthy	False negative (*FN*)	True negative (*TN*)

A true positive means that the patient is sick and is also diagnosed to be sick, while a a false positive is when the patient is healthy, but is diagnosed as sick. When a patient is diagnosed as healthy, the patient is either healthy, a true negative, or sick, a false negative.

These four possible outcomes can in turn be exploited in further analysis. The accuracy, A_c, is the number of correct diagnoses divided by the total number of diagnoses

$$A_c = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.3}$$

where *TP* is true positives, *FP* is false positives, *TN* is true negatives and *FN* is false negatives. The sensitivity, S_e, is the number of true positives (*TP*) divided by the sum of the true positives (*TP*) and false negatives (*FN*) [34]

$$S_e = \frac{TP}{TP + FN}. \tag{4.4}$$

The specificity, S_p is the number of true negatives (*TN*) divided by the sum of the number of true negatives (*TN*) and false positives (*FP*) [34]

$$S_p = \frac{TN}{TN + FP}. \tag{4.5}$$

A commonly used analyzing technique is based on the receiver operating characteristic (ROC) curve [34].The ROC curve plots the sensitivity on the vertical axis and one minus specificity on the horizontal axis [34]. The ideal ROC curve has a true positive fraction of 1 and a false positive fraction of 0. The area under the ROC curve, *AUC*, therefore gives a quantitative measure of the quality of the diagnostic procedure. In Figure 4.4, the area under the ROC (*AUC*) curve is given for CT + PET as input images, that are locally cropped and sliced, with 8 neighbors (sorted). The dashed line, in Figure 4.4, is a non-informative diagonal ROC curve, with sensitivity = 1 - specificity. In the case of sensitivity = specificity = 0.5, this would be as good as a flipping a coin [42]. The ideal ROC curve maximizes the integral (*AUC*) with a value of 1 [34]. This is equivalent to the ideal ROC curve having an ideal value of 100 percent for the three measures accuracy, sensitivity and specificity [34].

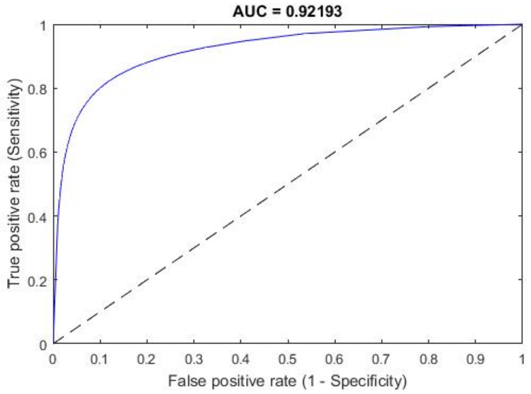


Figure 4.4: AUC, the area under the ROC curve, given for CT + PET as input images, that are locally cropped and sliced, with 8 neighbors (sorted).

Both the Dice similarity coefficient (*DSC*) and the κ statistics compares the binary masks (output images) of the autodelineation program with the contoured masks (based on the contouring of the oncologist/nuclear medicine physicist). The Dice similarity coefficient (*DSC*) is a performance measure used to evaluate the overlap between the contoured masks (ground truth) and the binary masks (produced by the autodelineation program). *DSC* is defined as

$$DSC = \frac{2TP}{FP + FN + 2TP}, \tag{4.6}$$

where *TP* are the true positives, *FP* are the false negatives and *FN* are the false negatives. *DSC* can have values between 0 and 1, where 0 and 1 indicates no and perfect overlap, respectively.

The κ statistics for comparison between observers find the agreement between the contoured masks (the ground truth) and the binary masks produced by the autodelineation program. The κ is calculated as

$$\kappa = \frac{2(TP \cdot TN - FP \cdot FN)}{(TP + FN)(FP + TN) + (TP + FP)(FN + TN)}, \tag{4.7}$$

where all the variables are defined above. A κ value of 1 indicates perfect agreement, and κ value larger than zero indicates agreement better than chance.

ANOVA

To test whether the mean of different performance parameters are significantly different, the analysis of variance, ANOVA, was performed [42]. In ANOVA, the null hypothesis is that there is no significant difference among the groups, while the alternative hypothesis states that there is at least one significant difference among the groups [42].

When using ANOVA, the following three assumptions need to be checked if

1. the samples are drawn from a normally distributed population,
2. the samples are independent of each other, and
3. the variance among the groups should be approximately equal [42].

After the assumptions of the ANOVA are tested, the calculation of the *F*-value and the *p*-value is performed. If the *p*-value is less than 0.05, the null hypothesis is rejected with 95 % confidence, concluding with the alternative hypothesis that at least one group mean is different.

Chapter 5

Results

The autodelineation program was based on the contouring of cancerous regions performed by the oncologist and nuclear medicine physicist. The input images to the model could have any combinations of the following parameters,

- imaging modality (CT, PET or CT + PET),
- spatial information, zero or eight neighbors (and whether the neighbors were sorted or not), and
- size of the image stacks (sliced/not sliced).

The classification algorithm chosen was Fisher's Discriminant Analysis, and the model was tested using both the Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Twenty four different models were run, as seen in Table 5.1.

In addition, a PET-thresholding was performed using an absolute SUV of 2.5 as the chosen threshold limit. This gives an indication of how well a simple PET-thresholding would perform.

5.1 Performance measures of the models

All the models, given in Table 5.1, had *DSC* values between 0.27 to 0.68 and κ values between 0.12 and 0.62, indicating low but better than chance agreement with the contoured ground truth masks. The *DSC* and κ values were lowest for the CT images, with the values 0.27-0.4 and 0.12-0.27, respectively, indicating a slight to fair agreement. The PET images, with *DSC* values (0.65-0.68) and κ values (0.57-0.61), and the CT + PET images, with *DSC* values (0.53-0.68) and κ values (0.47-0.62), indicate moderate agreement. It is therefore evident that PET images and PET + CT images show a significant increase in the performance measures *DSC* and κ compared to only CT images.

Table 5.1: Performance measures for LDA and QDA classification models based on different combinations of feature vectors derived from PET, CT and CT + PET, given with two significant figures.

Images	**Neighbors**	**Sliced**	**Method**	**DSC**	κ	**AUC**	**Sens**	**Spec**
PET	0	Yes	LDA	0.65	0.59	0.90	0.62	0.95
CT	0	Yes	LDA	0.27	0.12	0.58	0.72	0.44
CT + PET	0	Yes	LDA	0.53	0.47	0.87	0.58	0.93
PET	8	Yes	LDA	0.68	0.61	0.92	0.70	0.93
CT	8	Yes	LDA	0.40	0.27	0.71	0.88	0.47
CT + PET	8	Yes	LDA	0.68	0.62	0.92	0.71	0.93
PET	8	Yes	QDA	0.66	0.58	0.91	0.76	0.88
CT	8	Yes	QDA	0.39	0.25	0.76	0.92	0.39
CT + PET	8	Yes	QDA	0.66	0.58	0.91	0.83	0.85
PET	8	No	LDA	0.68	0.61	0.92	0.70	0.93
CT	8	No	LDA	0.40	0.26	0.70	0.88	0.48
CT + PET	8	No	LDA	0.68	0.61	0.92	0.71	0.94
PET	8 (unsorted)	Yes	LDA	0.65	0.60	0.91	0.63	0.96
CT	8 (unsorted)	Yes	LDA	0.35	0.20	0.61	0.75	0.50
CT + PET	8(unsorted)	Yes	LDA	0.66	0.60	0.91	0.63	0.96
PET	8 (unsorted)	Yes	QDA	0.65	0.57	0.90	0.75	0.88
CT	8 (unsorted)	Yes	QDA	0.39	0.25	0.76	0.92	0.39
CT + PET	8(unsorted)	Yes	QDA	0.64	0.56	0.90	0.84	0.84
PET	0	Yes	QDA	0.65	0.59	0.89	0.64	0.94
CT	0	Yes	QDA	0.36	0.20	0.70	0.97	0.26
CT + PET	0	Yes	QDA	0.66	0.58	0.91	0.75	0.90
PET	8	No	QDA	0.66	0.58	0.91	0.76	0.88
CT	8	No	QDA	0.38	0.25	0.76	0.93	0.38
CT + PET	8	No	QDA	0.65	0.58	0.91	0.83	0.85

DSC: Dice similarity coefficient, Sens: Sensitivity, Spec: Specificity, AUC: area under the receiving operator curve and κ: κ statistics. Values are averaged over all patients. LDA: Linear Discriminant Analysis, QDA: Quadratic Discriminant Analysis.

Models using CT showed a tendency for higher sensitivity and significantly lower specificity than PET alone or PET combined with CT. For CT the sensitivity was between 0.72 and 0.97, compared to the sensitivity for PET (0.62-0.76) and CT + PET (0.58-0.84). CT models had a specificity between 0.26 and 0.50. In contrast, the specificity was above 0.84 for all models run using PET (0.88-0.96) and CT + PET (0.84-0.96). The AUC was lower for CT (0.58-0.76) than for PET (0.89-0.92) and CT + PET (0.87-0.92), indicating that the models performed better when PET images (either alone or in combination with CT) were included as the input images.

5.2 The effect of different factors

For the imaging modality PET all of the five performance measures wre significant, and thus it was evident that the model was heavily influenced by PET. It was therefore interesting to test the effect of each imaging modality separately for the effect of different model factors, without PET present as the dominating factor, in the N-way ANOVA. This was especially beneficial for CT, as this imaging modality contributed significantly less to the model in comparison with PET (and PET in combination with CT). It was then possible to test how the factors in the model, spatial information, sorting, slicing and classifier, affected the model performance in only CT images (or only PET or only CT + PET images).

5.2.1 Effect of spatial information and classifier

The effect of including spatial information (zero or eight sorted neighbors) and the classifier choice (LDA/QDA) on model performance was investigated. The p values (from the N-way ANOVA) are displayed in the Tables 5.2, 5.3, 5.4 and 5.5, testing for the effect of spatial information and classifier choice both including and excluding PET and CT as factors. The input images to all of these models were locally cropped and sliced.

In Table 5.2, PET and CT included as factors, it is evident that the performance measures DSC, κ and sensitivity improved significantly when neighborhood information was included. The choice of classifier had a significant effect on the performance measures sensitivity and specificity. PET had significant effect on all performance measures, specificity were the only significant performance measure for CT. As seen in the Tables 5.3, 5.4 and 5.5, the effect of neighbors and classifier were significant for sensitivity (all), specificity (PET and CT), and DSC (CT).

It was beneficial to include spatial information for all model combinations. For the models with eight sorted neighbors, there were minimal differences in the performance measures due to classifier choice. However, models based on zero neighbors had higher specificity, but lower sensitivity for the QDA classifier compared to the LDA classifier. The classifier QDA performed better for DSC, κ and AUC than LDA for CT images, especially for zero neighbors. In contrast, for PET and CT + PET, there were minimal effect of the choice of classifier. Therefore, it is recommended to include spatial information in models, as neighborhood information improved model performance.

Table 5.2: p values (from the N-way ANOVA) for the effects of the factors classifier choice (LDA/QDA) and spatial information (zero or eight sorted neighbors), including imaging modality (CT and PET) as a factor.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
CT	-	-	-	-	0.04
PET	<0.0001	<0.0001	0.0003	0.0009	<0.0001
Neighbors	0.003	0.007	-	0.002	-
Classifier	-	-	-	0.011	0.0001

Table 5.3: p values (from the N-way ANOVA) for the effects of the factors classifier choice (LDA/QDA) and spatial information (zero or eight sorted neighbors) based on only PET images.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
Neighbors	-	-	-	0.03	0.003
Classifier	-	-	-	0.04	0.0012

Table 5.4: p values (from the N-way ANOVA) for the effects of the factors classifier choice (LDA/QDA) and spatial information (zero or eight sorted neighbors) based on only CT images.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
Neighbors	<0.0001	-	-	0.03	0.003
Classifier	0.0001	-	-	0.04	0.0012

Table 5.5: p values (from the N-way ANOVA) for the effects of the factors classifier choice (LDA/QDA) and spatial information (zero or eight sorted neighbors) based on only CT + PET images

.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
Neighbors	-	-	-	0.02	-
Classifier	-	-	-	0.02	-

DSC: Dice similarity coefficient, Sens: Sensitivity, Spec: Specificity, AUC: area under the receiving operator curve and κ: κ statistics. Values are averaged over all patients. Non significant p-values are given by '-', using the significance level p = 0.05, for visual purposes.

5.2.2 Effect of sorting neighbors and classifier

The effect of sorting the eight neighbors and the classifier choice (LDA/QDA) on model performance was investigated. Table 5.6 displays the p values for the effects of the factors image type, classifier choice (LDA/QDA) and sorting of the eight neighbors on the classification performance measures. The ANOVA tables for only CT and only PET showed no significance of sorting the eight neighbors and the classifier choice (LDA/QDA) for any of the five performance measures. Table 5.7 displays the p values for the effects of the factors classifier choice (LDA/QDA) and sorting of the eight neighbors on the performance measures. The input images to all of these models were locally cropped and sliced.

Table 5.6: p values from N-way ANOVA for the effects of the factors image type, classifier choice (LDA/QDA) and sorting of the eight neighbors on the classification performance measures.

Factor	**DSC**	**κ**	**AUC**	**Sens**	**Spec**
CT	-	-	-	-	-
PET	<0.0001	<0.0001	0.0003	0.004	<0.0001
Sorted	0.04	-	-	-	-
Classifier	-	-	-	0.001	<0.0001

Table 5.7: p values from N-way ANOVA for CT + PET, testing for the effects of the factors classifier choice (LDA/QDA) and sorting of the eight neighbors on the classification performance measures.

Factor	**DSC**	**κ**	**AUC**	**Sens**	**Spec**
Sorted	-	-	-	-	-
Classifier	-	-	0.02	-	-

DSC: Dice similarity coefficient, Sens: Sensitivity, Spec: Specificity, AUC: area under the receiving operator curve and κ: κ statistics. Values are averaged over all patients. Non significant p-values are given by '-', using the significance level p = 0.05, for visual purposes.

In Table 5.6, PET and CT included as factors, it is evident that sorting of the eight neighbors were only significant for the performance measures *DSC*. The classifier choice had a significant effect on the two performance measures sensitivity and specificity. For LDA, the sensitivity was higher when the neighbor were sorting, while the specificity was higher for unsorted neighbors. In the case of QDA, there was minimal effect of sorting (of the eight neighbors) on specificity and sensitivity. PET had a significant effect on all performance measures, while CT had no significant effect on any of the performance measures. It is evident, from Table 5.7, that the only significant factor for CT + PET was classifier choice (LDA/QDA) on the performance measure *AUC*. Therefore, sorting of the neighbors had no to minimal effect on model performance.

5.2.3 Effect of slicing and classifier

The slicing of the image stacks reduced the total number of voxels and improved the balance between the two classes. In this study, the slicing of the locally cropped images increased, in total, the fraction of cancerous voxels in the image stacks from approximately 9% to 13%. The effect of slicing and classifier choice on model performance was thus investigated. The results from the N-way ANOVA testing slicing, classifier choice, CT and PET on model performance is displayed in Table 5.8. The effect of slicing and classifier choice, for the individual imaging modalities CT, PET and CT+PET, is displayed in the Tables 5.9, 5.10 and 5.11, respectively. All these models were based on image stacks that were unfolded with eight sorted neighbors.

In Table 5.8, including CT and PET as factors, there was no significant effect of slicing for any of the performance measures. PET was, as always in this study, significant for all performance measures. The choice of classifier had an significant on effect on all performance measures except *AUC*.

In contrast, in the Tables 5.9, 5.10 and 5.11, it is clear that the factor slicing had significant effect on the performance measures for the individually tested imaging modalities. For PET and PET in combination with CT, the factor slicing had significant effect on all performance measures, while CT had significant effect on *DSC*, κ and *AUC*. The classifier (LDA/QDA) had significant effect on *DSC* (PET and CT) and sensitivity (CT + PET).

Thus, the effect of slicing was found to be minimal in this study, as there were no significant effect of slicing for any of the performance measures when factors of imaging modality, PET and CT, were included. Since testing for individual imaging modalities showed that slicing had significant effect on performance measures and better balance between the classes was desirable, it was recommended to perform slicing. The slicing increased the fraction of cancerous voxels in the image stack.

Table 5.8: p values from N-way ANOVA for the effects of the factors image type, classifier (LDA/QDA) and slicing of the image stacks on the classification performance measures.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
CT	-	-	-	-	-
PET	<0.0001	<0.0001	<0.0001	0.0001	<0.0001
Slicing	-	-	-	-	-
Classifier	<0.0001	<0.0001	-	0.0007	<0.0001

Table 5.9: p values from N-way ANOVA for PET, testing for the effects of the factors classifier (LDA/QDA) and slicing of the image stacks on the classification performance measures.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
Slicing	0.009	0.006	0.008	0.009	0.002
Classifier	0.05	-	-	-	-

Table 5.10: p values from N-way ANOVA for CT, testing for the effects of the factors classifier (LDA/QDA) and slicing of the image stacks on the classification performance measures.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
Slicing	0.005	0.03	0.009	-	-
Classifier	0.02	-	-	-	-

Table 5.11: p values from N-way ANOVA for CT + PET, testing for the effects of the factors classifier (LDA/QDA) and slicing of the image stacks on the classification performance measures.

Factor	**DSC**	κ	**AUC**	**Sens**	**Spec**
Slicing	0.010	0.006	0.03	<0.0001	0.003
Classifier	-	-	-	0.0005	-

DSC: Dice similarity coefficient, Sens: Sensitivity, Spec: Specificity, AUC: area under the receiving operator curve and κ: κ statistics. Values are averaged over all patients. Non significant p-values are given by '-', using the significance level p = 0.05, for visual purposes.

5.2.4 Visualization of the ROC curve

Figure 5.1 displays the *AUC* for all the models with eight sorted neighbors for both classifiers (LDA/QDA) and all imaging combinations. The three models based on PET for LDA and CT + PET for both LDA and QDA resulted in *AUC*s over 0.90. The combination of CT and PET for LDA had slightly (but not significantly) larger *AUC* than the other two models. The model combination resulting in the lowest *AUC* was again CT and LDA, were a small part of the ROC curve is even lower than the by-chance line for sensitivities and false positive rates smaller than 0.05.

Figure 5.2 displays the *AUC* for all the models with zero neighbors for both classifiers (LDA/QDA) and all imaging combinations. It was evident that the combination of PET and LDA gave the highest *AUC* for models with zero neighbors. In contrast, the combination of PET and QDA gave somewhat (but not significantly) lower *AUC*. The worst model combination here was CT and LDA, were part of the ROC curve is even lower than the by-chance line when the sensitivity and false positive rate is low (smaller than 0.3). Thus, this model performs no better than randomly expected for low sensitivities and false positive rates. Therefore, if the model is run for CT, there would be improvement in *AUC* by switching from the linear LDA (0.58) to the non-linear QDA (0.70) classification algorithm.

Figure 5.3 displays the *AUC* for all the models with eight unsorted neighbors for both LDA and QDA and all imaging combinations. The model combinations based on PET using the LDA classifier and PET in combination with CT for both classifiers (LDA and QDA), resulted in *AUC* larger than 0.90. The model combination yielding the lowest *AUC* was again obtained using the LDA classifier and the CT images.

Figure 5.4 displays the *AUC* for all the models based on locally cropped and sliced image stacks, unfolded with eight sorted neighbors, and both classifiers (LDA and QDA) for all imaging combinations. The models based on inclusion of PET (either PET alone or PET in combination with CT) all yielded an *AUC* larger than 0.90, while the *AUC*s for the CT based models were less than 0.76.

Of the twenty-four models, a total of thirteen models based on different combinations of either PET or CT + PET, gave an *AUC* larger than 0.90, with *DSC* of 0.64-0.68 and κ of 0.56-0.62, indicating a very good model performance and substantial agreement between the ground truth and the binary mask produced by the model. The model combinations based on the LDA classifier, inclusion of both sliced and unsliced PET images (PET and PET + CT), resulted the in same *AUC* of 0.92 indicating excellent performance of these four models. For these four models the *DSC* (0.68) and κ (0.61-0.62) were high, indicating a substantial agreement between the ground truth and the mask produced by the autodelineation program.

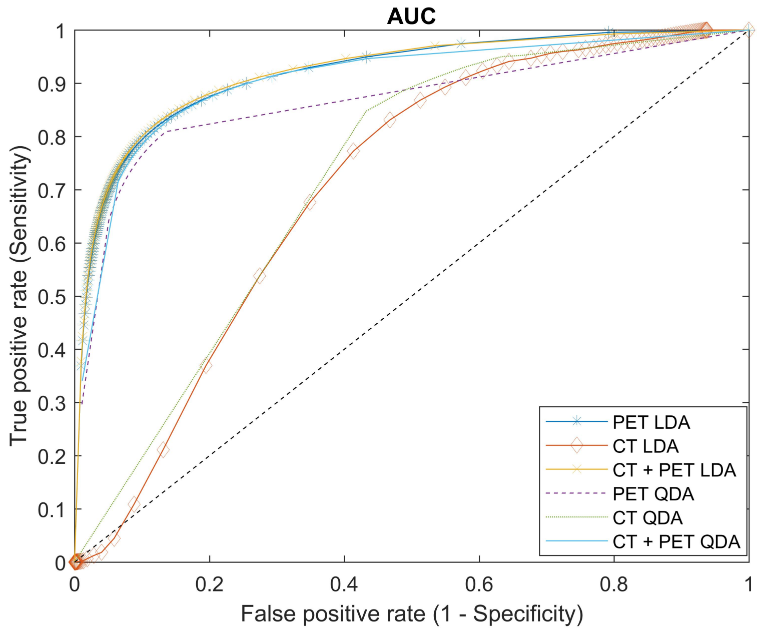


*Figure 5.1: Area under the curve (*AUC*) for models with eight sorted neighbors, sliced, LDA and QDA classifiers, and for all imaging modalities.*

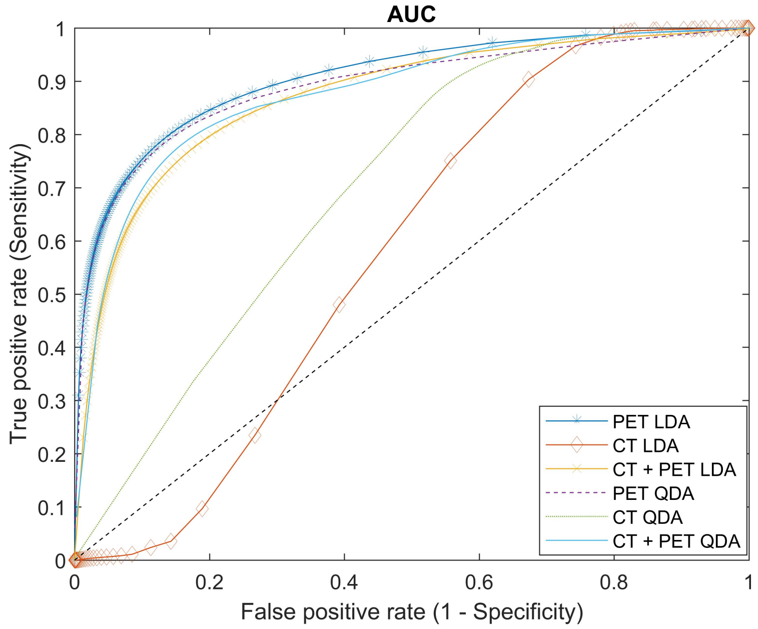


*Figure 5.2: Area under the curve (*AUC*) for models with zero neighbors, sliced, LDA and QDA classifiers, and for all imaging modalities.*

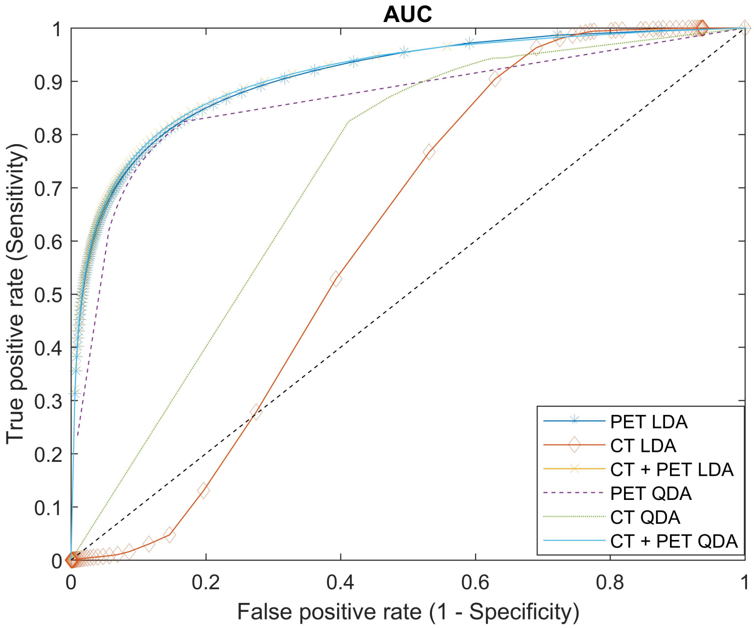


Figure 5.3: Area under the curve (AUC) *for models with eight unsorted neighbors, unsliced, LDA and QDA classifiers, and for all imaging modalities.*

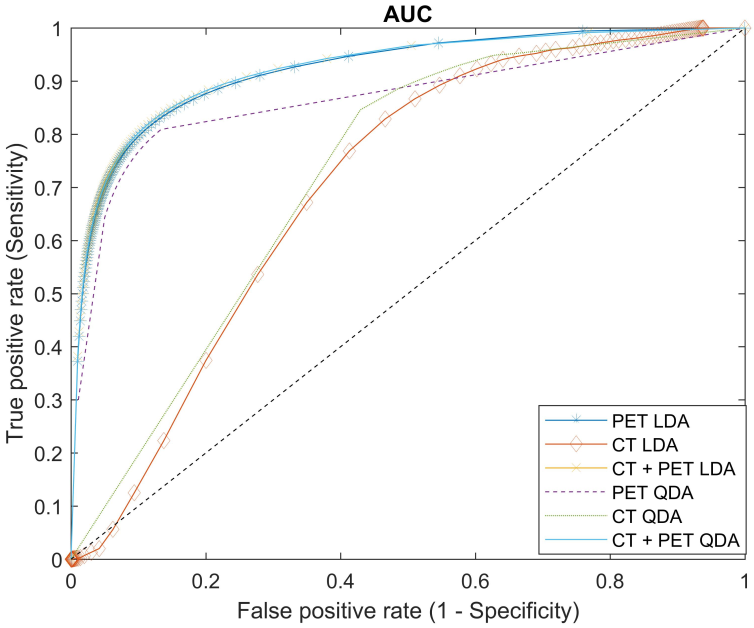


Figure 5.4: Area under the curve (AUC) *for models with eight sorted neighbors, unsliced, LDA and QDA classifiers, and for all imaging modalities.*

5.2.5 Performance plots of the best models

There was clear differences in performance measures among patient, especially for κ and DSC, also within the same model combination. The two model combinations yielding the absolute highest AUC, both the sliced and unsliced combined CT and PET images, unfolded with eight sorted neighbors, using the LDA classifier, are displayed in the performance plots in Figures 5.5 and 5.6. In the performance plot for the sliced images, four arbitrary ground truth masks (with the visually largest amount of cancerous regions, displayed in white) are also displayed.

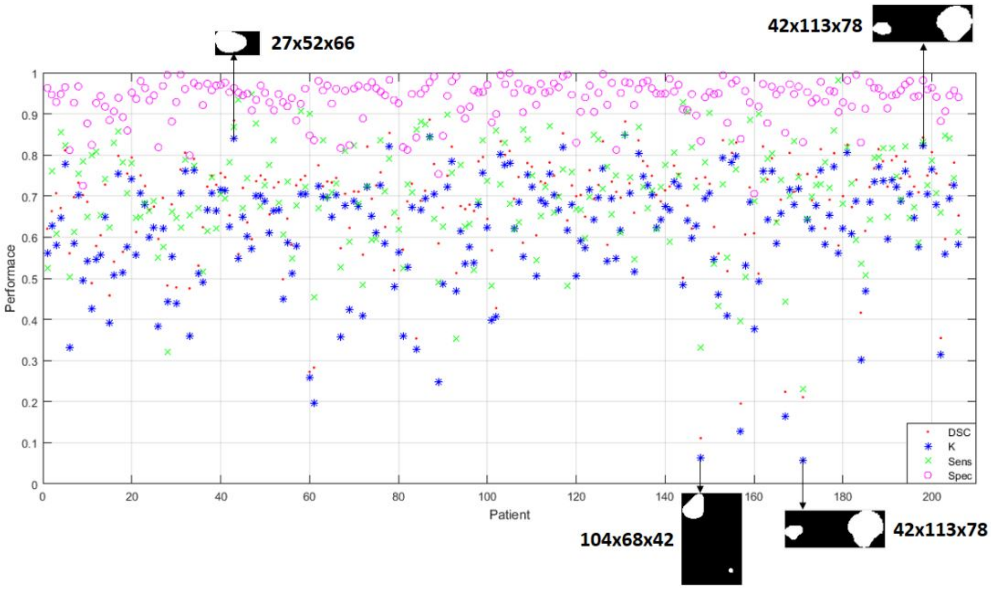


Figure 5.5: Performance measures for the model combination of sliced CT + PET images, eight sorted neighbors and LDA classifier. The four ground truth masks are added to visually inspect the amount of cancerous regions (white) and non-cancerous regions (black).

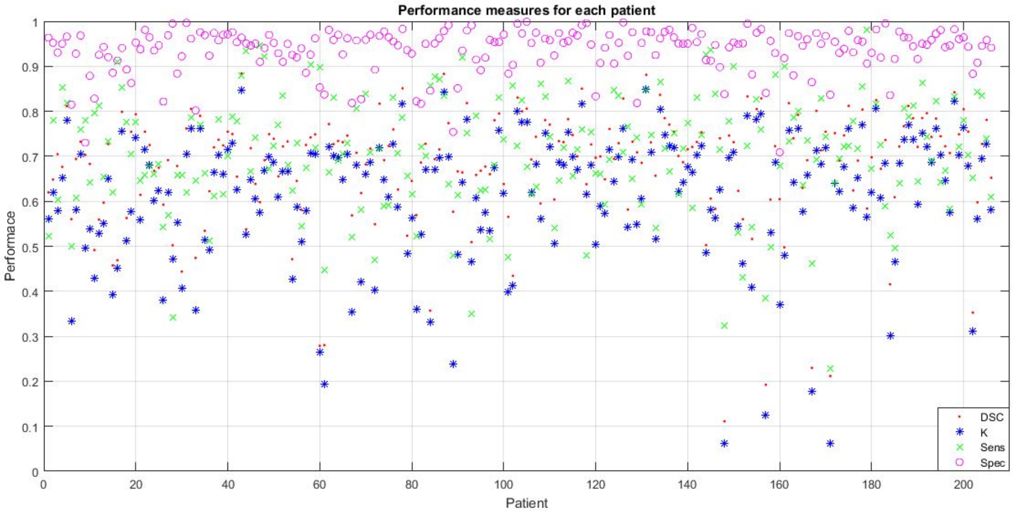


Figure 5.6: Performance measures for the model combination of unsliced CT + PET images, eight sorted neighbors and LDA classifier.

5.3 Dependencies of κ and *DSC* on class balance

As the variations of κ and *DSC* values were substantial between patients, these two performance measures were tested for their relation with the fraction of cancerous regions and the number of cancerous voxels, as these two parameters also displayed variation between patients. These relations are illustrated in Figures 5.7, 5.8 and 5.9 for all imaging modalities (CT and PET combined, PET only and CT only, respectively). For these plots, the first row display κ and *DSC* as a function of the relative fraction of cancerous regions and the second row display the κ and *DSC* values as a function of the number of cancerous voxels.

For the PET + CT case, Figure 5.7, small relative fraction of cancerous regions and number of cancerous voxels had a large range in both κ and *DSC* values. Higher relative fractions of cancerous regions and number of cancerous regions had smaller variations in κ and *DSC* values. Therefore, there was no evident, clear relation between the variable κ and *DSC* values and either the relative fraction of cancerous regions or the numbers of cancerous voxels for the CT + PET images. Almost identical tendencies for these same four plots for PET + CT is found in the corresponding plots for PET, as seen in Figure 5.8.

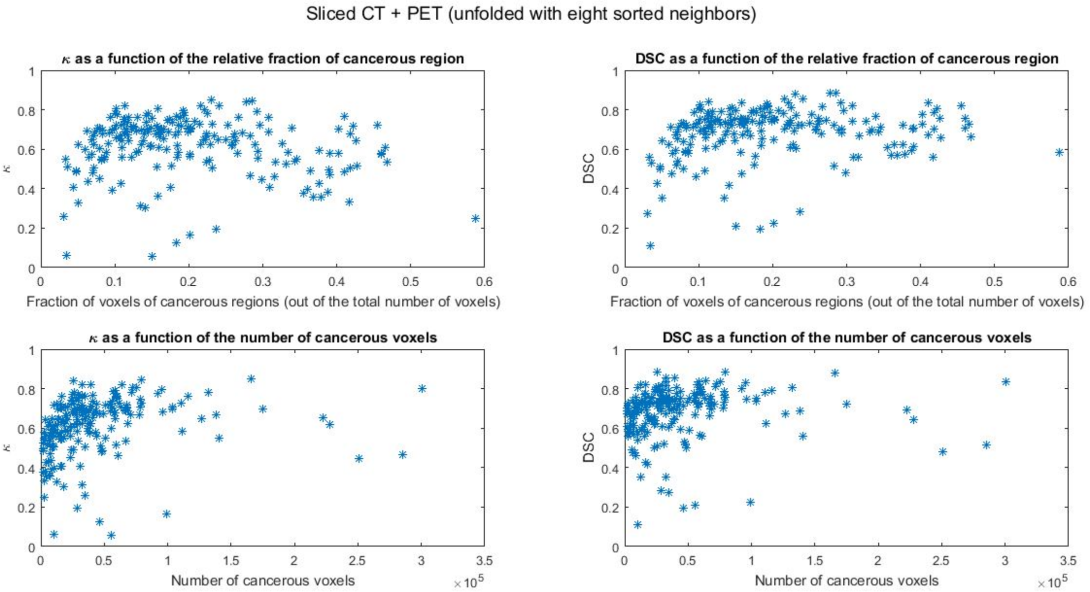


Figure 5.7: κ (left) and DSC (right) as a function of the relative fraction of cancerous regions (row 1) and number of cancerous regions (row 2), for CT + PET image stacks (sliced and eight sorted neighbors), respectively.

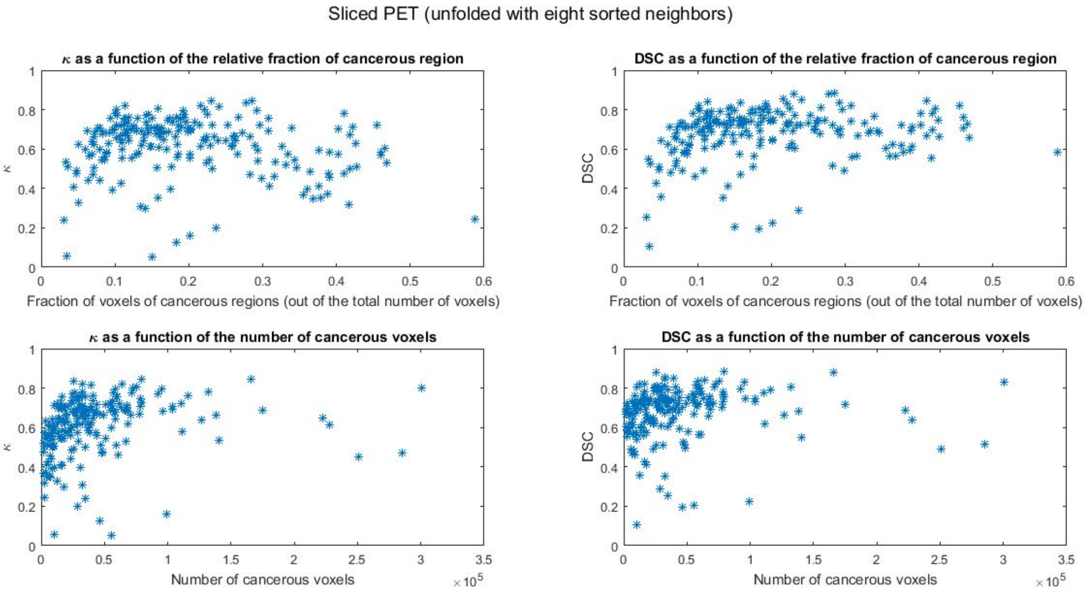


Figure 5.8: κ (left) and DSC (right) as a function of the relative fraction of cancerous regions (row 1) and number of cancerous regions (row 2) for PET image stacks (sliced and eight sorted neighbors), respectively.

In contrast, the CT plots (Figure 5.9) were notably different on all the PET and CT + PET cases. In the plots of κ and DSC plotted against the number of cancerous voxels there were larger variations in κ and DSC for smaller numbers of cancerous voxels, but this might simply be due to the larger amount of patients with smaller number of cancerous voxels. In the plot of κ and DSC against the fraction of voxels of cancerous regions there was a near linear tendency. Thus, κ and DSC were almost linearly dependent on the fraction of voxels of cancerous regions. Higher κ and DSC was desirable and would was achieved for CT image stacks with higher fraction of cancerous regions, and thus methods to improve the fraction of cancerous voxels in an image stack seems promising.

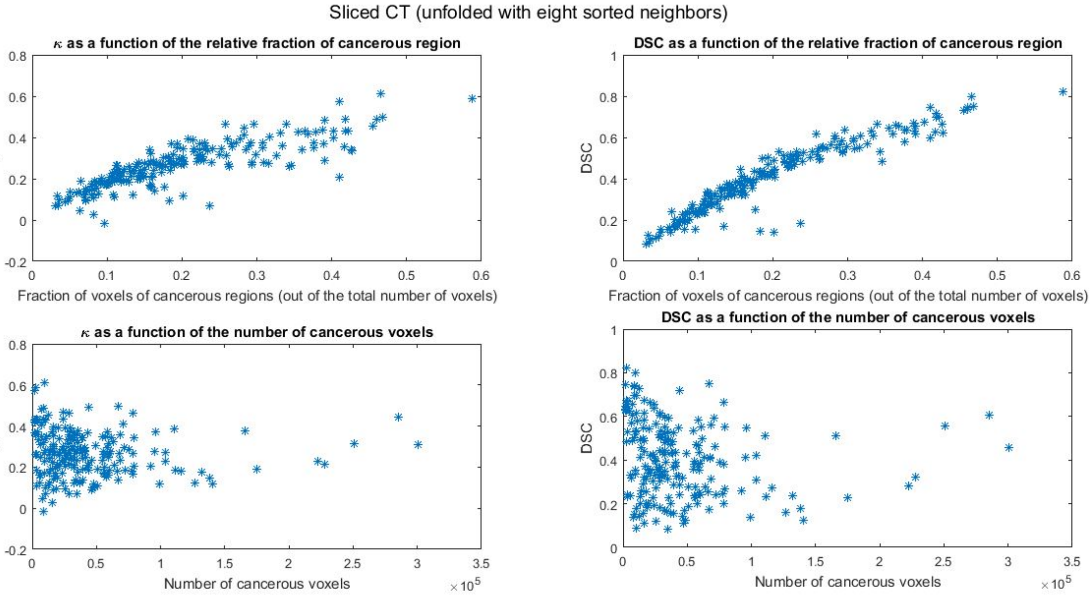


Figure 5.9: κ (left) and DSC (right) as a function of the relative fraction of cancerous regions (row 1) and number of cancerous regions (row 2) for CT image stacks (sliced and eight sorted neighbors), respectively.

5.4 Visualization of input and output masks

Another interesting aspect to investigate, is how the model was affected by non-cancerous regions with high SUV and cancerous regions with low SUV. The Figures 5.10 and 5.11 display the unsliced and sliced CT + PET images (first column), the ground truth mask (second column), probability maps (third column) and LDA binary mask (forth column), respectively. All of the input models were unfolded using eight sorted neighbors and classified using LDA.

Figure 5.10 displays image sliced from four patients, all having unsliced CT + PET images, unfolded with eight sorted neighbors and classified by LDA. The agreement between the contoured mask (ground truth) and the LDA binary mask were good for two of the patients, with *DSC* of 0.88 (first row) and 0.72 (second row), and low for the patient given in the third and forth row with a *DSC* of 0.11 and 0.21, respectively. In the CT + PET image, the CT is displayed as grey while PET is displayed in the red color shading.

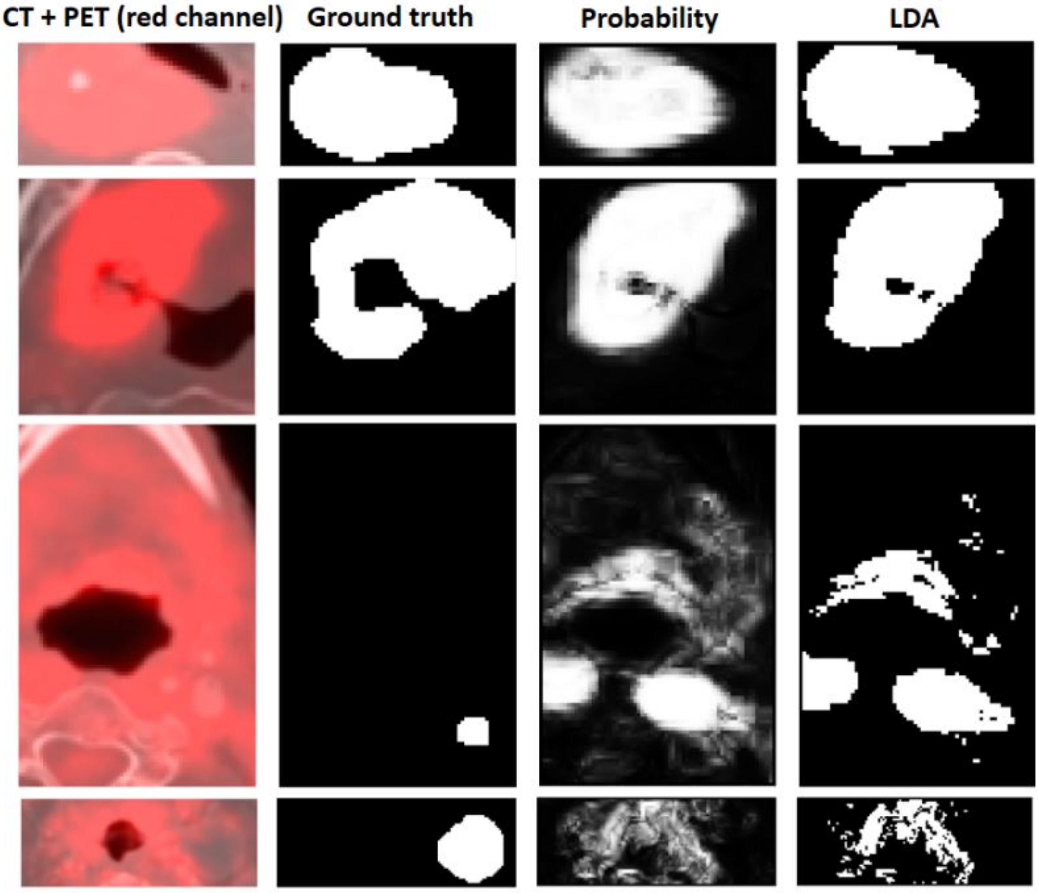


Figure 5.10: The CT + PET (PET overlayed in the red channel) image in the first column and the ground truth mask in the second column for image slices from three different patients. The LDA model was based on CT + PET, unfolding using eight sorted neighbors and auto-scaling as a preprocessing method, and resulted in the probability maps (third column) and LDA output image (forth column). In the probability maps, black and white indicate 0% and 100% predicted probability of the voxel belonging to cancerous tissue, respectively. A probability threshold of 50% was used to convert the probability maps into LDA output images. The agreement between the contoured mask (ground truth) and the LDA binary mask, given for the Dice similarity coefficent (DSC*), was 0.88 (first row), 0.72 (second row), 0.11 (third row) and 0.21 (fourth row).*

For the two first rows, the red color is highly defined in regions similar to the binary image representing the ground truth and also to the binary image produced by the autodelineation program using the LDA classification algorithm. From these images, it is clear that there was a high overlap between the contoured mask (ground truth) and the LDA binary mask, and thus achieving high *DSC* of 0.88 (first row) and moderately high of *DSC* 0.72 (second row). In contrast, for the third and forth row there was little overlay between the contoured mask (ground truth) and the LDA binary mask, resulting in the low *DSC* values of 0.11 and 0.21. In the CT + PET images for the third and forth row, large portions of the image are displayed in red, indicating areas of high metabolism. Due to the symmetric distribution of the metabolically active areas, these are not cancerous as seen in the image representing the ground truth, but rather due to other glucose-demanding causes, such as infections. This led to a high number of false positives as many non-cancerous regions were classified as cancerous. Also, the actual cancerous regions were not classified as cancerous, thus causing a high number of false positives.

In Figure 5.11, the sliced CT + PET images (first column), the ground truth mask (second column), probability maps (third column) and LDA binary mask (forth column) are illustrated for image slices from three different patients. The agreement between the contoured mask (ground truth) and the LDA binary mask are good for two of the patients, with *DSC* of 0.80 (first row) and 0.84 (third row), and low for the patient given in the second row with a *DSC* of 0.21.

For the patient in the second row, it is clear that there was regions in the tissue with high glucose metabolism that the oncologist/nuclear medicine physicist has not contoured as cancerous tissue. This would then lead to a false positive, as non-cancerous tissue are classified as cancerous in the LDA model. On the other hand, the largest cancerous region displayed in the mask considered the ground truth (created by the contouring of the oncologist/nuclear medicine physicist) was not displayed in the CT + PET image. As a consequence, this patient also has regions that are cancerous, but are not classified as cancerous in the LDA model (false negative). In conclusion, this patient had both types of misclassification (false positive and false negative). It is therefore evident that the agreement between the contoured mask (ground truth) and the LDA binary mask was low for this patient. This is both seen by the differences between the mask showing the ground truth and the LDA created mask, and by the low *DSC* and κ values.

Patients with high degree of either false positives, false negatives or both types of misclassification would therefore have low agreement between the ground truth mask and the output mask for the model. It is also worth mentioning that the deeper red color in the CT + PET image (first row and first column) indicates that there are areas with higher metabolism within the tumor, and thus this tumor is heterogeneous with regard to voxel intensity (in PET image).

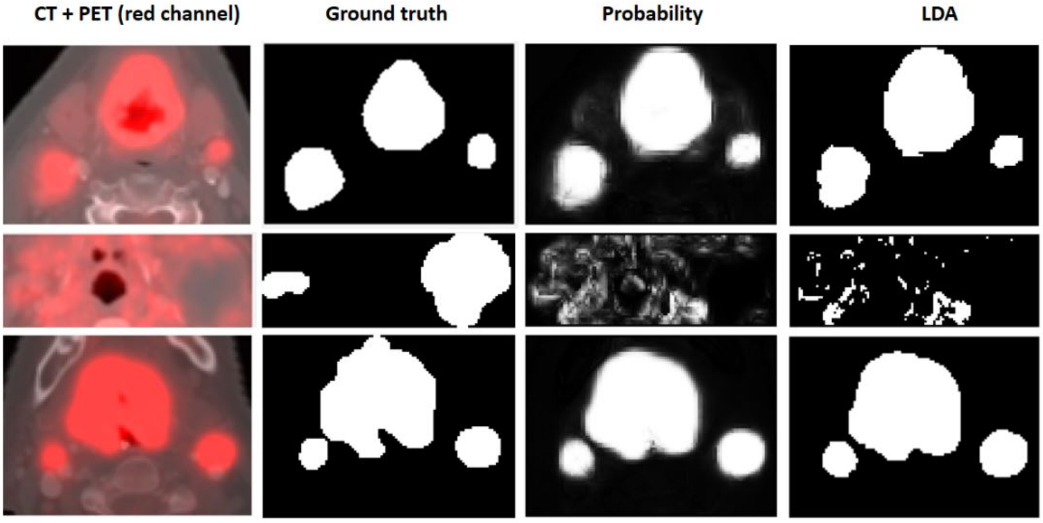


Figure 5.11: The CT + PET (PET overlayed in the red channel) image in the first column and the oncologists'/nuclear medicine physicists' mask in the second column for image slices from three different patients. The LDA model was based on CT + PET, unfolding using eight neighbors (sorted) and auto-scaling as a preprocessing method, and resulted in the probability maps (third column) and LDA output image (fourth column). In the probability maps, black and white indicate 0% and 100% predicted probability of the voxel belonging to cancerous tissue, respectively. A probability threshold of 50% was used to convert the probability maps into LDA output images. The agreement between the contoured mask (ground truth) and the LDA binary mask, given for the Dice similarity coefficent (DSC*), was 0.80 (first row), 0.21 (second row) and 0.84 (third row).*

5.5 PET, CT and PET + CT images

It is clear that PET images influence the voxel classification model heavily, when all the imaging modalities are used to test different performance measures. To get a visual understanding of how the autodelineation model performs for the CT and PET separately as well as combined, the classification results are illustrated in Figure 5.12.

From the CT images (first row), it is clear that there was little overlap with the binary mask representing the ground truth in the actual CT image, the probability map image and the LDA binary mask, as given by the low *DSC* (0.15) and κ (0.14). This model combination achieved an *AUC* of 0.71, indicating fair model performance. The sensitivity was very high (0.99), as there are few cancerous voxels classified as non-cancerous. In contrast, the specificity was very low (0.41). This is evident through the large fraction of non-cancerous voxels being classified as cancerous in the LDA binary mask for CT.

Both the PET and CT+ PET display overlap between the input images, the probability map images and the LDA binary masks compared to the mask representing the ground truth, as indicated through the high *DSC* (0.64) and κ (0.62) values. The PET and CT+ PET both achieves an *AUC* of 0.92, sensitivity of 0.74 and specificity of 0.97, indicating good to excellent model performance.

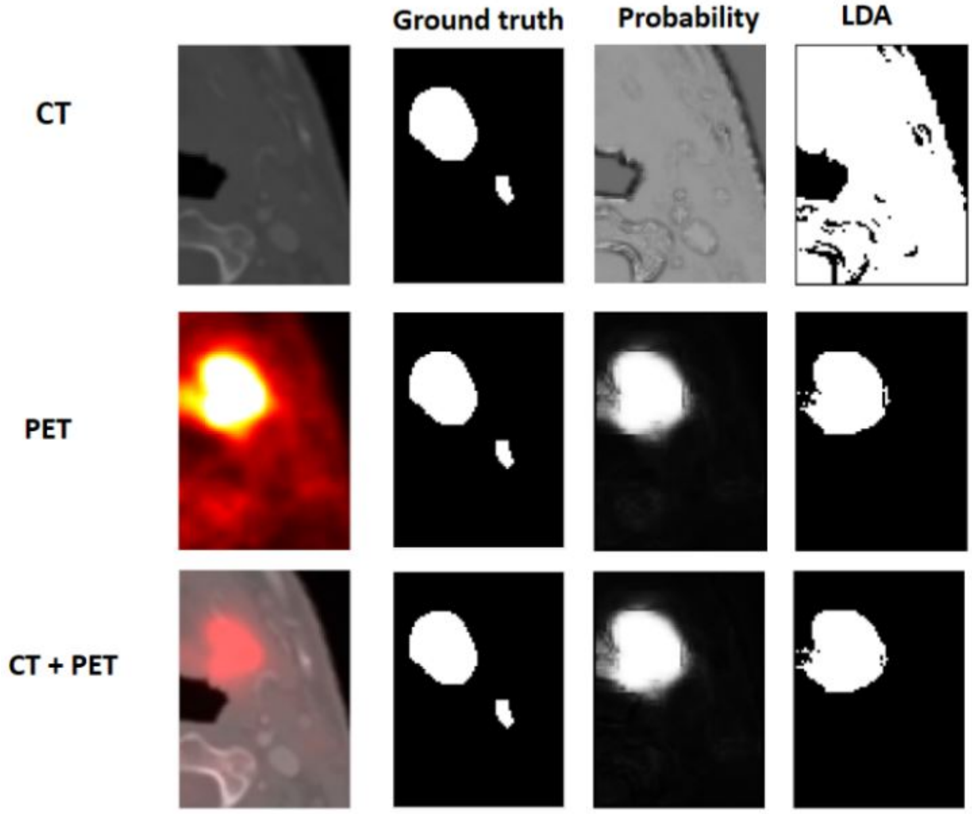


Figure 5.12: The CT, PET and CT + PET (PET overlayed in the red channel) sliced images in the first column and the ground truth masks in the second column for image slices from the same patient. The LDA model was based on unfolding of eight sorted neighbors and auto-scaling as a preprocessing method, and resulted in the probability maps (third column) and LDA output image (fourth column). In the probability maps, black and white indicate 0% and 100% predicted probability of the voxel belonging to cancerous tissue, respectively. A probability threshold of 50% was used to convert the probability maps into LDA output images. The agreement between the contoured mask (ground truth) and the LDA binary mask, given for the Dice similarity coefficent (DSC), was 0.15 (first row), 0.64 (second row) and 0.64 (third row).

5.6 PET thresholding

Including the PET images in the classification significantly improved all the performance measures. Thus, it is interesting to find out how well a fast, simple absolute PET thresholding would determine the cancerous regions compared to the more time consuming autodelineation model.

All the voxels in the PET image stack was tested on whether the voxel intensity were larger than 2.5 or not. This test resulted in a binary image mask where voxels with intensities larger than 2.5 were given the value one, while the voxels with intensities smaller than 2.5 were given the value zero. This segmentation of cancerous and non-cancerous tissues was performed on the basis of the locally cropped and sliced PET image stack.

In Figure 5.13, the performance measures *DSC*, κ, sensitivity and specificity are plotted for each of the 206 patients in the dataset.

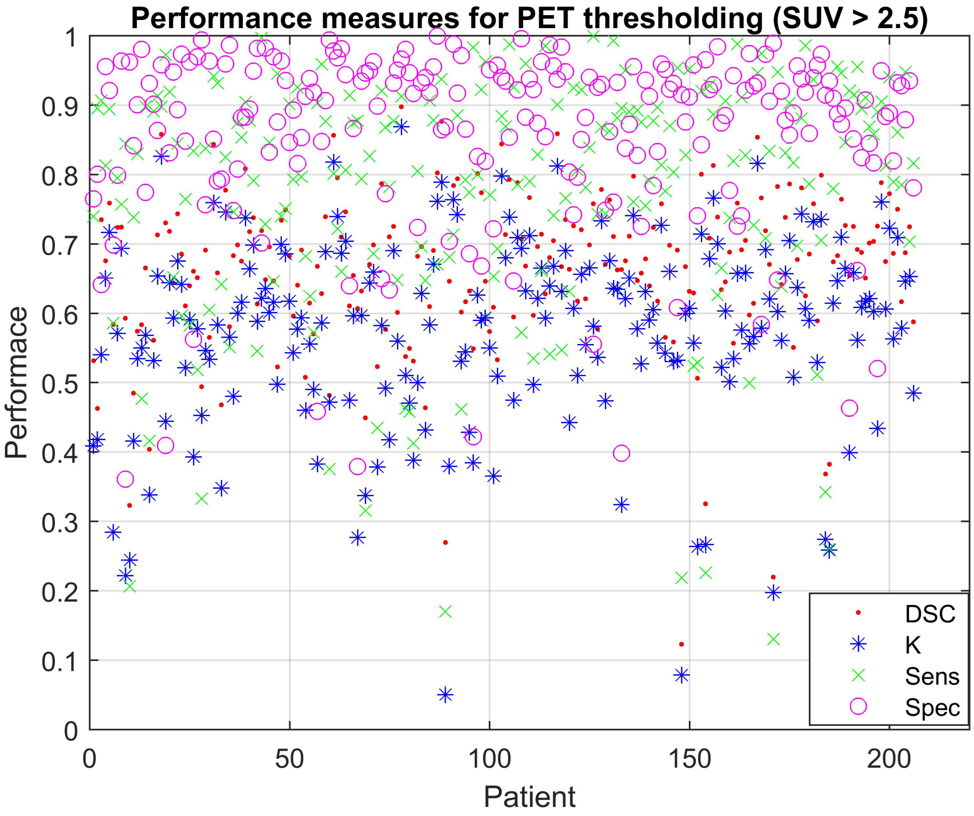


Figure 5.13: Performance measures for simple PET thresholding, segmenting into cancerous regions for SUV > 2.5.

This PET thresholding resulted in a *DSC* of 0.66, κ of 0.58, sensitivity of 0.77 and specificity of 0.86, indication a good overlap between the binary image resulting from the PET thresholding and the ground truth mask. All the performance measures from the PET thresholding were within the intervals, for all the 24 model combinations as seen in Table 5.1. This indicates that a simple PET thresholding is a promising tool as a fast estimate of the location of the cancerous regions within the patient.

Chapter 6

Discussion

A delineation program was developed to automatically identify cancerous regions within the head and neck. This program was originally developed for tumor delineation for different types of MR images of cervical cancer. It has now been tested and modified for CT, PET and CT + PET image stacks of head and neck cancers. All of these image stacks were locally cropped to improve the balance between the two voxel classes (cancer/no cancer). For each patient the location of the cancerous regions in the image stack, their x, y and z coordinates, were found and the minimum and maximum of these x, y and z coordinates formed the basis for the local cropping. The locally cropped image stacks could also be sliced by removing z-planes without any cancerous voxels. Thus, the slicing of the image stacks further reduced the size of the image stack and improved the balance between the classes. All of the image stacks were auto-scaled to a mean of zero and a standard deviation of one.

The image stacks were unfolded with either zero neighbors or eight neighbors (sorted/unsorted), to test the effect of spatial information in the model. The classification algorithms Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were implemented to classify voxels in the CT, PET or CT + PET image stacks as cancerous or non-cancerous. This classification resulted in probability maps and binary LDA/QDA masks which were then compared to the ground truth, binary masks based on the oncologist's/nuclear medicine physicist's contouring, yielding the performance measures DSC, AUC, κ, sensitivity and specificity. A total of twenty-four models were evaluated to explore the effects for different parameters on model performance.

6.1 Overview of the effect of imaging modality

All the models tested had *DSC* values between 0.27 and 0.68 and κ values between 0.12 and 0.62, indicating a low to substantial agreement between the contoured masks (ground truth) and the binary masks resulting from the autodelineation program. Models based on only CT had significantly lower for *DSC* and κ values than PET and CT + PET models, that both had moderate to substantial agreement with the contoured masks. Therefore, it is evident that PET had a significant contribution to the voxel classification model and that information regarding cancerous and non-cancerous regions are, in general, better captured by the PET image stacks than the CT image stacks, as illustrated in Figure 5.12.

PET contributed significantly to the classification model, and the models based on PET or CT + PET images provided the highest overall performance measures. Of the twenty-four models, a total of thirteen models gave an *AUC* larger than 0.90 and thus had very good performance. Models with the classifier LDA, spatial information (eight sorted neighbors) and either PET or PET + CT images (as input to the model) resulted in an *AUC* of 0.92, *DSC* of 0.68 and κ of 0.61-0.62. Thus, these four models had excellent performance and the masks produced by the autodelineation program had substantial agreement with the ground truth.

There were indications of higher sensitivity for models using CT (0.72-0.97) compared to than PET alone (0.62-0.76) or PET in combination with CT (0.58-0.84). As sensitivity, or true positive rate, gives the proportion of cancerous voxels correctly identified as such, this performance measure is influenced by the number of false positives and the number of true positives [23]. The number of false positives was substantial in the PET image stacks for some patients and slices, particularly when there were non-cancerous regions with high glucose metabolism, as seen in the Figures 5.11 and 5.10. Because ^{18}F-FDG PET measures the glucose metabolism of different tissue, this imaging modality would display metabolically active areas due to cancer and other causes in the same manner, causing false positives (non-cancerous regions classified as cancerous) [23]. On the other hand, some cancers have low glucose metabolism, resulting in false negatives (cancerous regions classified as non-cancerous) [23].

The false positives and false negatives were naturally problematic for the autodelineation program, and resulted in poorer model performance. Therefore, the performance measures of the models varied for individual patients in relation to the fraction of false positives and false negative, as visually displayed in the Figures 5.11 and 5.10. Similar voxel intensities (SUV) in the PET images do not necessarily indicate a tumor, but could result from a variety of glucose-demanding causes, such as infection and muscle movement. Symmetric regions with high metabolism are in most cases due to other causes than cancer, and the model might perform better if a test for symmetrical areas is implemented [23]. Instead of a program based entirely on autocontouring, a semi-automated program for contouring would be preferred in this situation [23]. In a semi-automated program, the oncologist/nuclear medicine physicist can apply knowledge that is challenging to implement in a machine learning algorithm, such as typical spread of cancers in certain areas, relevant information about the patients (for example, scar tissue after an operation would result in higher SUV) and aspects regarding human anatomy and so fourth [23].

CT had significantly lower specificity (0.26-0.50) than both PET (0.88-0.96) and PET in combination with CT (0.84-0.96). This means that in CT images there was a higher proportion of non-cancerous voxels correctly identified as such (non-cancerous), compared to the PET and CT + PET images. As already discussed, PET provided more false positives than CT, due to high ^{18}F-FDG uptake in all tissue with high glucose metabolism, and thus had more non-cancerous voxels incorrectly classified as cancerous. Therefore, there were fewer correctly identified non-cancerous voxels for PET compared to CT.

The *AUC*, the area under the ROC curve plotted for sensitivity as a function of false positive rate (1 - specificity), was lower for CT (0.58-0.76) than for PET alone (0.89-0.92) and PET in combination with CT (0.87 - 0.92). Thus, the models performed significantly better when PET images, either alone or in combination with CT, were included as the input images.

6.2 The effect of model parameters

All the imaging modalities were tested to determine the effect of including spatial information (zero or eight sorted neighbors), sorting of the eight neighbors, slicing of the locally cropped image stacks and the choice of classifier on model performance.

Inclusion of the PET images in the modeling significantly improved the performance measures and appears to be the significant factor. Since PET is a functional imaging modality it provides, for the most part, clear differences between cancerous and non-cancerous regions, as previously discussed.

The model performance also improved significantly when neighborhood information was included. In this case, the model had more information that could be used to distinguish the two classes. However, sorting of the neighbors according to descending voxel intensity did not, in general, significantly effect model performance. Thus, this sorting provided no or little additional information.

The choice of classifier had little effect on performance, except for delineation in CT images where QDA performed significantly better. This indicates that cancerous and non-cancerous voxels in CT images were better classified using a non-linear boundary than a linear boundary. In contrast, for PET alone or PET in combination with CT, the model performance was either higher for LDA than QDA or there was insignificant effect of the choice of classifier. As a consequence, a classification based on a linear boundary for PET images performs well and thus a simple thresholding of PET images should also perform well. Applying a simple, absolute SUV threshold of 2.5 to the PET images resulted in an AUC of 0.85, DSC of 0.66 and κ of 0.58, indicating good model performance and moderate agreement between the segmented binary mask, resulting from testing the voxels (in the PET image stack) against the absolute threshold of 2.5, and the ground truth binary mask. The performed thresholding method for volume delineation is the simplest of the thresholding methods, and had a simple interpretation and high efficiency [43]. However, this thresholding method was sensitive to particularly the partial volume effect, an effect due to the combination of the limiting resolution of PET and image sampling, tumor heterogeneity, lack of consideration of the background and so forth [43]. Thus, more advanced methods, based on a relative thresholding or a combination of an absolute and a relative threshold, should be performed to achieve higher degree of overlap between the segmented binary masks, from the thresholding, and the ground truth binary masks [43, 44].

The model performance was not significantly affected by slicing when imaging modality were included as a factor, but was significant affected when the imaging modalities were tested separately, as PET otherwise had a dominantly effect in these models. The slicing of the locally cropped images increased the fraction of cancerous voxels, in the total image stack, from approximately 9% to 13% and thus improved the balance between the two classes. Therefore, slicing is recommended in order to increase the fraction of cancerous voxels in the image stack and enhance performance measures.

6.3 Dependencies of κ and DSC on class balance

There were large variations between patients regarding the performance measures, especially for κ and DSC, also within the same model combination. These two performance measures were tested for their relation with the fraction of cancerous regions and the number of cancerous voxels, as these two parameters also displayed variation between patients.

There was no clear relationship between the two performance measures κ and DSC and the fraction of cancerous regions or the number of cancerous voxels for models based on PET images alone and PET images in combination with CT images (Figures 5.7 and 5.8). In contrast, models based on CT images showed a near linear relation between κ and *DSC*, and the fraction of cancerous regions (Figure 5.9). Higher κ and *DSC* were achieved for CT image stacks with higher fraction of cancerous regions. Thus, ways to further improve the class balance could potentially improve CT classification models.

6.4 Assessment of the autodelineation program

An autodelineation program to identify cervical cancer tumors was developed by former PhD student at NMBU, Turid Torheim. It used different types of MR images (T2w, T1w abd DCE), not PET and CT images as in this study [4]. Torheim found improved model performance when spatial neighborhood information was included, in accordance with this study.

Inclusion of functional imaging modalities (such as PET and DCE-MRI) gave higher model performance than anatomical imaging modalities (such as CT, Tw1 and Tw2 MRI). The inclusion of the DCE-MRI features in Torheim's study had an significant effect on all performance measures except sensitivity, increasing *DSC* (from 0.18-0.20 to 0.41-0.44), κ (from 0.14-0.17 to 0.38-0.42), *AUC* (from 0.65-0.76 to 0.84-0.87) and specificity (from 0.50-0.54 to 0.92-0.94) [4]. In this study, the inclusion of the PET images led to a significant improvement of all performance measures, increasing *DSC* (from 0.27-0.4 to 0.65-0.68), κ (from 0.12-0.27 to 0.46-0.62), *AUC* (from 0.58-0.76 to 0.87-0.92) and specificity (from 0.26-0.50 to 0.84-0.96). However, sensitivity was lower, from 0.72-0.97 to 0.58-0.84, when either PET alone or PET in combination with CT was used compared to models solely based on CT. Thus, in both these studies, the model combination that gave the highest model performance was based on inclusion of functional imaging, indicating that cancerous regions are better segmented on the basis of glucose metabolism, given in the PET images, compared to the anatomy of cancerous tumor tissues, given in the CT images.

On an overall basis, these two studies are based on both different imaging modalities and cancer type, and thus the comparison of the performance measures must consider this difference. For example, the achieved *DSC* and κ is higher for the best model in this study compared to the best model in Torheim's study. However, the achieved performance measures should rather be compared to similar studies as the autodelineation model is clearly significantly affected by the imaging modality used as the input images. Another difference between the studies is the number of patients, 206 patients in this study compared to 78 patients in Torheim's study, which can influence the significance of the factors and also results in fewer patients to train the model on yielding potentially higher degree of overfitting. Torheim achieved better balance between the classes, twenty-seven percent compared to nine or thirteen percent, and thus might perform better in the segmentation of the two classes. If only primary tumors were considered for the dataset used in this study, there would be an increase in the balance between classes, as the images stacks could be cropped smaller than what was the case when the cropping was based on all cancerous regions (both tumors and cancer-infiltrated lymph nodes).

A former Master student at NMBU, Elise Mühlbradt, further developed Torheim's autodelineation program using the same dataset, testing the effect of different pre-processing methods and classification algorithms on model performance [45]. Mühlbradt found that other pre-processing methods, such as Median filtering, Savitzky-Golay filtering or Contrast Limited Adaptive Histogram Equalization (CLAHE), did not significantly improve performance measures relative to autoscaling [45]. Thus, in this Master's thesis, autoscaling was used for image pre-processing. Moreover, as voxel intensities in the PET and CT images are on different scales and were used together for the combined PET and CT images, autoscaling or similar pre-processing was needed in order to adjust for the differences in voxel intensity between these two imaging modalities.

None of the classification methods implemented by Mühlbradt, random forest, k nearest neighbors, SVM or AdaBoost, resulted in significantly different performance from LDA [45]. These methods were therefore not implemented in this Master's thesis, which focused on the methods Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). LDA and QDA draw a border, linear or quadratic, respectively, between the two classification groups. These models would thus be more robust for different types of cancerous regions and would therefore classify the cancerous voxels more equally for the cancerous regions, of each patient, compared to more advanced classification methods with higher tendencies of overfitting [40].

In this study, the same dataset was divided into different training and validation sets, but it would have been beneficial to test the classification on an independent dataset to reduce the bias resulting from developing and testing the model on the same patients [40]. This is also the aim for the autodelineation program, to train it on a dataset, and then being able to identity all types of cancerous regions that are not part of the training dataset. For the cervical cancer dataset, the leave-ten-out and leave-one-out cross validation resulted in no significant difference for the five performance measures [45]. Although a leave-ten-out cross validation is more time-efficient, due to the reduction in the number of times the training has to be performed, it was decided to use the leave-one-out cross validation in this study as this would be the situation in a clinic.

Although the models performed well for certain model combinations, fully automatic delineations are still not accurate enough for all patients, as shown in this study. The autodelineation can assist the physician in the delineation of the cancerous regions, and in particular the probability maps could be beneficial, as these (probability) maps showed the probability of the voxels belonging to a tumor.

Several studies have shown that there are variations in the contouring performed by different physicians (inter-variance) and also variations in contouring performed by the same physician on the same medical images (intra-variation) [46–48]. Considering the oncologist's and nuclear medicine physicist's contouring as the ground truth, is naturally not accurate. This assumption was made in order to train the model and get an estimate of the performance measures from the program, based on the supervised learning classification. The accuracy of the physician, although considered the ground truth, would affect the accuracy of the autodelineation program for the performance measures. The voxel classification cannot be more accurate than the contouring performed by the physicians, as this is used in the training of the model.

The autodelineation program needs to be further developed before it can be implemented at hospitals. The autodelineation program was originally developed to analyze different types of MR images of cervical cancer. During this thesis, the autodelineation program has been adapted and tested on CT, PET and CT + PET images of head and neck cancer (HNC). Thus, the program can be used to detect different cancer types in different imaging modalities, demonstrating its flexibility. The autodelineation program shows promising results when tested on both different types of MR images of cervical cancer and CT/PET images of head and neck cancer.

6.5 Proposals for further research

To further test the developed autodelineation model it should be tested on new and independent datasets. Oslo University Hospital (OUS) has access to ^{18}F-FDG PET and CT images together with clinical data for about 200 HNC patients at three collaborating institutions on medical imaging and radiotherapy of HNCs [10]. In addition, the autodelineation program should be adapted and tested for other cancer types than cervical cancer and HNC.

The optimal pre-processing method of the input images should be investigated in order to find the best method for specific imaging modalities and datasets, as image properties can be enhanced by the use of appropriate filters and other pre-processing methods [36, 40].

During this thesis, only two classification methods, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), were used to classify the voxels, due to time limitations. There are a large number of other classification methods that should be tested, in order to find the optimal classification method for the given model specifications [35, 38, 40, 41].

The influence of spatial information on the model performance can be tested using more than eight neighbors (in 2D unfolding) and 3D unfolding of various number of neighbors. In addition, other methods to achieve spatial information on the voxels should be tested, such as a Local Binary Pattern code per voxel [38, 40].

Optimization of the post-processing of the voxel classification using different post-processing methods should be performed for different imaging modalities and datasets, in order to improve the output binary image compared to the contoured mask (ground truth). Further research should implement and test the use of supervoxels, to find out if this implementation makes the cancerous regions more coherent. Active contour modeling should be tested for the effect of object outlining from a possibly noisy 2D image [40]. Also, different classification methods might perform better depending on the fraction of voxels of cancerous regions, especially for the CT images. Deep learning techniques can be applied to further test the autodelineation program [36].

The oncologist's/nuclear medicine physicist's contouring of the cancerous regions were considered the ground truth during this thesis. This assumption is not accurate, as there can be both large intra- and inter-variations in tumor contouring. A method to assess the robustness of the autodelineation program given the uncertainties in the contouring should be developed, to be able to display the cancerous regions with an uncertainty margin.

Chapter 7

Conclusion

An autodelineation program, developed on the basis of different kinds of MR images of cervical cancer, was further developed using PET/CT images of head and neck cancer. The program's performance was assessed for different model factors for the performance measures Dice Similarity Coefficient (DSC), area under the ROC curve (AUC), κ-statistics, sensitivity and specificity. There were large variations between the patients in the performance measures also within the same model combination, especially for the *DSC* and κ values.

PET had a significant effect on all performance measures. Of the twenty-four models, a total of thirteen models based on different combinations of either PET or CT + PET, gave an *AUC* larger than 0.90, with *DSC* of 0.64-0.68 and κ of 0.56-0.62, indicating a very good model performance and substantial agreement between the ground truth and the binary mask produced by the model. The model combinations based on the LDA classifier, inclusion of both sliced and unsliced PET images (PET and PET + CT), resulted the in same *AUC* of 0.92 indicating excellent performance of these four models. For these four models the *DSC* (0.68) and κ (0.61-0.62) were high, indicating a substantial agreement between the ground truth and the mask produced by the autodelineation program.

Although PET had a significant effect on the performance measures, it was also prone to false positives and false negatives as the tracer ^{18}F-FDG provides information about the glucose metabolism of different tissues. Non-cancerous tissue can have high glucose metabolism and cancerous tissue is not alway more metabolically active than normal tissue. A simple thresholding of PET was promising, yielding similar performance measures as the total of the twenty-four model combinations.

There was a tendency for higher sensitivity for models based on CT (0.72-0.97) than for both PET alone (0.62-0.76) and PET in combination with CT (0.58-0.84).

On the other hand, CT had significantly lower specificity (0.26-0.50) than both PET (0.88-0.96) and CT + PET (0.84-0.96).

Apart from the effect of imaging modality, the factors spatial information, neighbors, sorting of the eight neighbors, classifier choice and the slicing of the image stacks were tested for significance. The *DSC*, κ and sensitivity all indicated a positive significant effect when spatial information, unfolding with eight neighbors, were introduced. However, whether these neighbors were sorted had little effect. The use of a linear (LDA) compared to a non-linear (QDA) classifier had significant effect on the performance measures. Slicing of the locally cropped image stacks is recommended as a balance between the classes was desirable, in the classification, and there were indications of a near linear relation for CT between both κ and *DSC* and the fraction of voxels of cancerous regions.

In conclusion, the autodelineation program shows promising results when tested on both different types of MR images of cervical cancer and CT/PET images of head and neck cancer. The classification should be tested on new, independent datasets to better assess the model's performance. In addition, different methods for pre-processing, extraction of spatial information, classification algorithms, post-processing and so forth, should be further tested in order to achieve optimization. This autodelineation model has the potential of becoming a useful tool for physicians in contouring and radiotherapy planning assessment of different types of cancer based on a variety of different imaging modalities.

Bibliography

[1] Cancer today. International Agency for Research on Cancer. [Online]. Available: http://gco.iarc.fr/today/fact-sheets-cancers?cancer=29&type=0&sex=0

[2] (2017) Cancer in norway 2016: Cancer incidence, mortality, survival and prevalence in norway. Cancer registry of Norway. [Online]. Available: https://www.kreftregisteret.no/globalassets/cancer-in-norway/2016/cin-2106.pdf

[3] How cancer starts. Visited the 2017-11-20. [Online]. Available: http://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts

[4] T. Torheim, E. Malinen, K. H. Hole, K. V. Lund, U. G. Indahl, H. Lyng, K. Kvaal, and C. M. Futsaether, "Autodelineation of cervical cancers using multiparametric magnetic resonance imaging and machine learning," *Acta Oncologica*, vol. 56, no. 6, pp. 806–812, 2017.

[5] Kreft i hode-halsregionen. Oncolex, by oncologist Jan Folkvard Evensen. Visited the 2017-10-14. [Online]. Available: http://oncolex.no/Hodehals

[6] Head and neck cancers. National Cancer Institute. Visited the 2017-10-14. [Online]. Available: https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet#q1

[7] Alkohol og tobakk hovedårsak til hode og halskreft. Visited the 2017-11-06. [Online]. Available: https://www.kreftregisteret.no/Generelt/Nyheter/Alkohol-og-tobakk-og-ikke-HPV-smitte-er-hovedarsak-til-hode-og-halskreft/

[8] Human papillomavirus (hpv). Visited the 2017-11-06. [Online]. Available: https://www.cdc.gov/hpv/parents/whatishpv.html

[9] S. Marur, G. D'Souza, W. H. Westra, and A. A. Forastiere, "Hpv-associated head and neck cancer: a virus-related cancer epidemic," *The lancet oncology*, vol. 11, no. 8, pp. 781–789, 2010.

[10] Bioradiance - biologic imaging for outcome prediction and radiotherapy dose painting for head and neck cancers. Oslo University Hospital (OUS). 2017.

[11] C. Fakhry, W. H. Westra, S. Li, A. Cmelak, J. A. Ridge, H. Pinto, A. Forastiere, and M. L. Gillison, “Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial,” *Journal of the National Cancer Institute*, vol. 100, no. 4, pp. 261–269, 2008.

[12] Årsaker til munnhulekreft. Oncolex, by oncologist Jan Folkvard Evensen. Visited the 2017-11-08. [Online]. Available: http://oncolex.no/Hodehals/Diagnoser/Munnhule/Bakgrunn/Arsaker

[13] Årsaker til kreft i nese og bihuler. Oncolex, by oncologist Jan Folkvard Evensen. Visited the 2017-11-17. [Online]. Available: http://oncolex.no/Hodehals/Diagnoser/Nese-bihuler/Bakgrunn/Arsaker

[14] Årsaker til strupekreft. Oncolex, by oncologist Jan Folkvard Evensen. Visited the 2017-11-17. [Online]. Available: http://oncolex.no/Hodehals/Diagnoser/Strupe/Bakgrunn/Arsaker

[15] Årsaker til kreft i spyttkjertler. Oncolex, by oncologist Jan Folkvard Evensen. Visited the 2017-11-17. [Online]. Available: http://oncolex.no/Hodehals/Diagnoser/Spyttkjertler/Bakgrunn/Arsaker

[16] Tnm. Union for International Cancer Control. 11/29/2017. [Online]. Available: http://www.uicc.org/resources/tnm

[17] Types of cancer treatment. National Cancer Institute (NCI). 12/1/2017. [Online]. Available: https://www.cancer.gov/about-cancer/treatment/types

[18] “4. definition of volumes,” *Journal of the International Commission on Radiation Units and Measurements*, vol. 10, no. 1, pp. 41–53, 2010. [Online]. Available: http://dx.doi.org/10.1093/jicru/ndq009

[19] J. Lilley, *Nuclear physics: principles and applications*. John Wiley & Sons, 2013.

[20] About icru - mission statement. The International Commission on Radiation Units and Measurements. 11/29/2017. [Online]. Available: https://icru.org/mission-statement/uncategorised/mission-statement

[21] Pet/ct. Kreftforeningen. Visited the 2017-10-14. [Online]. Available: https://kreftforeningen.no/om-kreft/undersokelse-ved-kreft/pet-scan/

[22] S. A. Kane, *Introduction to physics in modern medicine*, second edition ed. CRC Press by Taylor & Francis Group, 2009.

[23] R. Wahl, *Principles and practice of PET and PET/CT*, second edition ed. Lippincott Williams and Wilkins, a Wolters Kluwer business, 2009.

[24] J. C. Miller, H. H. Pien, D. Sahani, A. G. Sorensen, and J. H. Thrall, "Imaging angiogenesis: applications and potential for drug development," *Journal of the National Cancer Institute*, vol. 97, no. 3, pp. 172–187, 2005.

[25] Y. J. Choi, "Perfusion imaging of the head and neck."

[26] E. Malinen, *Personal communication*. Professor i Biophysics and Medical Physics, 2017, department of Physics, University of Oslo (UiO).

[27] C. Cuenod, L. Fournier, D. Balvay, and J.-M. Guinebretiere, "Tumor angiogenesis: pathophysiology and implications for contrast-enhanced mri and ct assessment," *Abdominal imaging*, vol. 31, no. 2, pp. 188–193, 2006.

[28] K. T. Bae, "Intravenous contrast medium administration and scan timing at ct: considerations and approaches," *Radiology*, vol. 256, no. 1, pp. 32–61, 2010.

[29] Basics of fdg. Hardvard Medical School. 11/30/2017. [Online]. Available: http://www.med.harvard.edu/jpnm/chetan/basics/basics.html

[30] E. Pauwels, M. Ribeiro, J. Stoot, V. McCready, M. Bourguignon, and B. Maziere, "Fdg accumulation and tumor biology," *Nuclear medicine and biology*, vol. 25, no. 4, pp. 317–322, 1998.

[31] S. R. Cherry and M. Dahlbom, "Pet: physics, instrumentation, and scanners," in *PET*. Springer, 2006, pp. 1–117.

[32] B. H. Brown, R. Smallwood, D. Barber, P. Lawford, and D. Hose, *Medical physics and biomedical engineering*. IOP Publishing Ltd, 1999.

[33] S. Vandenberghe, E. Mikhaylova, E. D'Hoe, P. Mollet, and J. Karp, "Recent developments in time-of-flight pet," *EJNMMI physics*, vol. 3, no. 1, p. 3, 2016.

[34] N. B. Smith and A. Webb, *Introduction to medical imaging: physics, engineering and clinical applications*. Cambridge university press, 2011.

[35] Supervised learning workflow and algorithms. MathWorks. Visited the 2017-10-23. [Online]. Available: https://se.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html#bswluhd

[36] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[37] Biograph 16 pet ct. Direct diagnostic services. Visited the 2017-10-15. [Online]. Available: http://www.directdiagnosticservices.com/Biograph-16-PET-CT.html

[38] M. G. Jameson, L. C. Holloway, P. J. Vial, S. K. Vinod, and P. E. Metcalfe, “A review of methods of analysis in contouring studies for radiation oncology,” *Journal of medical imaging and radiation oncology*, vol. 54, no. 5, pp. 401–410, 2010.

[39] S. A. Stacker, M. G. Achen, L. Jussila, M. E. Baldwin, and K. Alitalo, “Metastasis: lymphangiogenesis and cancer metastasis,” *Nature Reviews Cancer*, vol. 2, no. 8, pp. 573–583, 2002.

[40] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 810.

[41] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning: Data mining, inference, and prediction.”

[42] J. Marques de Sá, “Estimating data parameters,” *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*, pp. 141–161, 2007.

[43] H. Zaidi and I. El Naqa, “Pet-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques,” *European journal of nuclear medicine and molecular imaging*, vol. 37, no. 11, pp. 2165–2187, 2010.

[44] I. Støen, “Optimal thresholding of pet-based autocontouring of boost volume for radiotherapy of anal carcinoma,” Master’s thesis, Norwegian University of Science and Technology, Ås, 2017.

[45] E. Mühlbradt, “Videreutvikling av et diagnostisk verktøy for automatisk svulstinntegning av livmorhalskreft i mr-bilder,” Master’s thesis, Norwegian University of Life Sciences, Ås, 2016.

[46] A. C. Riegel, A. M. Berson, S. Destian, T. Ng, L. B. Tena, R. J. Mitnick, and P. S. Wong, “Variability of gross tumor volume delineation in head-and-neck cancer using ct and pet/ct fusion,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 65, no. 3, pp. 726–732, 2006.

[47] X. Geets, J.-F. Daisne, S. Arcangeli, E. Coche, M. De Poel, T. Duprez, G. Nardella, and V. Gregoire, “Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between ct-scan and mri,” *Radiotherapy and oncology*, vol. 77, no. 1, pp. 25–31, 2005.

[48] G. D. Hugo, D. Yan, and J. Liang, “Population and patient-specific target margins for 4d adaptive radiotherapy to account for intra-and inter-fraction variation in lung tumour position,” *Physics in medicine and biology*, vol. 52, no. 1, p. 257, 2006.

Appendices

Appendix A: Standardization of z scores

Standardization or auto-scaling is a method yielding $z\ scores$,

$$z_{scores,i} = \frac{x_i - E(x)}{\sigma(x)}, \tag{1}$$

where *E(x)* and $\sigma(x)$ is the expected value and standard deviation, respectively, of the total sample variable array x. x consists of the sample variables x_i, for i = 1, 2, ...,n-1, n. In this Thesis, the x would correspond to the image stack(s) containing all the voxel intensities, and x_i would be the voxel intensity of a specific voxel i. In the MATLAB computation of the $z\ scores$, potential `NaN` values are omitted using the option `'omitnan'` inside the $z\ scores$ function.

The expected value, *E(X)*, is equal to the sample mean $\bar{x}$. The mean without subtraction, $\bar{x}'$, is larger than the mean after the subtraction by the value 1024, $\bar{x}$, as seen in Equations 2 and 3. The mean before the subtraction is

$$\bar{x'} = \frac{1}{n}(\sum_{i=1}^{n} x'_i), \tag{2}$$

while the mean after the subtraction is given as

$$\begin{aligned} \bar{x} &= \frac{1}{n}(\sum_{i=1}^{n} x_i) \\ &= \frac{1}{n}(\sum_{i=1}^{n} x'_i - 1024) \\ &= \frac{1}{n}(\sum_{i=1}^{n} x'_i) - \frac{1}{n}(n \cdot 1024) \\ &= \frac{1}{n}(\sum_{i=1}^{n} x'_i) - 1024 \\ &= \bar{x'} - 1024. \end{aligned} \tag{3}$$

The marked variables, x'_i and $\bar{x'}$, are the sample variable and mean before the subtraction, while the unmarked variables, x_i and $\bar{x}$, are the sample variable and mean after the subtraction. There are n samples of x.

The numerator of Equation 1, *x - E(x)*, will be equal both with and without a subtraction because

$$\begin{aligned} x' - E(x') &= (x + 1024) - E(x + 1024) \\ &= x + 1024 - E(x) - E(1024) \\ &= x + 1024 - \bar{x} - 1024 \\ &= x - \bar{x} \\ &= x - E(x), \end{aligned} \tag{4}$$

where all the variables are defined above. The step between the first and the second line, in the equation above, is due to the linearity of the expected value

$$E(X + Y) = E(X) + E(Y). \tag{5}$$

The denominator of Equation 1, $\sigma(x)$, is equal with and without a subtraction because

$$\sigma(x + c) = \sigma(x), \tag{6}$$

where c is a constant. In this case, c = -1024, with a standard deviation of zero. Both the numerator and denominator in Equation 1 are exactly the same with and without the subtraction, as is evident from Equations 4 and 6, proving that the same subtraction performed to all the sample point yield no difference in z scores. Therefore, there is no need to subtract the value 1024 from the CT_{number} when auto-scaling is performed.

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

Postboks 5003
NO-1432 Ås
Norway