Norwegian University
of Life Sciences

Master's Thesis 2019   30 ECT
Faculty of Science and Technology
Cecilia Marie Futsæther

Image processing, radiomics and model selection for prediction of treatment outcome of anal cancer using CT-, PET- and MR-sequences

Maria Cabrol

Data Science

Preface

This master thesis is conducted at the Faculty of Science and Technology at NMBU and concludes five years of studies at Ås.

Ås, 15.05.2019

_____

Maria Cabrol

Abstract

The goal of this master thesis was to predict treatment outcome of anal cancer using features extracted from medical image sequences. The medical images had beforehand been provided from and registered at the Norwegian Radium Hospital. They consisted of CT,- PET- and MR-sequences, in which MR was the modality of interest as former projects at NMBU has involved prediction of treatment outcome using PET- and CT-sequences. The treatment outcome of interest was PFS, Progression-Free Survival.

The MR-sequences were cropped to remove missing information and new tumour masks were created for the CT-sequences. This was to exclude areas with air and abnormally high intensity values within the tumour region. Features were then extracted and the impacts of choosing different extraction dimension, tumour mask or discretization level were examined by calculating the correlation values between the features that had been extracted using different image setting.

The examination showed that most CT-, first order and NGTDM features extracted from image sequences with different discretization levels calculated with fixed bin widths, where highly correlated. The tumour mask that was created did not adapt well to the image sequences when they were resampled. Also, 2D- extraction seemed like an acceptable choice, as features extracted from 2D and 3D were mostly highly correlated. This was especially the case for shape, CT- and ADC-features extracted with the original tumour mask.

Two data sets were created from the extracted features. Data set 1 contained PET-, CT- and ADC(MR)-features from 36 patients. Data set 2 contained PET- and CT-features from 81 patients. These data sets had 35 patients in common.

The data sets were split into training and test sets and model selection with 10 classifiers and five feature selectors using the performance metric AUC was performed using the training sets. No validation AUCs were higher than the no information rates of 85.19% and 83.33% for the training sets of data sets 1 and 2, respectively. The highest validation AUCs were equal to 71.24% and 76.93% for the training sets, respectively for data sets 1 and 2.

Most features among the 50 most selected features during model selection were PET-features. Also, PET- and shape features were the only features that were among the most selected features for both data training sets.

There was no base for concluding that the MR-features were important predictors. For this, MR-features should have been examined separately. They did neither particularly stand out from features extracted from other modalities during model selection using data set 1.

All results were preliminary and indicative, as they were based on small data sets. The examination of impact of choice of extraction parameters were also done for features extracted from the data set of 36 patients. The results show, however that the prediction performance of these

List of symbols and abbreviations

Symbols

MRI:

$\boldsymbol{B}$: Magnetic field

B: Magnitude of magnetic field

μ_m: Magnetic moment of proton

ω: precession frequency of proton

γ: gyromagnetic ratio, $\gamma = 2.7 \times 10^8 \, s^{-1} T^{-1}$

$\boldsymbol{M}$: Magnetization

CT:

I: Intensity of photon beam

μ: Attenuation coefficient of tissue $[cm^{-1}]$

PET:

β^+: Positron

Abbreviations

GTV, Gross Tumour Volume

CT, Computed Tomography

PET, Positron Emission Tomography

LOR, line of response

MRI, Magnetic resonance imaging

RF, Radio frequency

TR, repetition time

TE, echo time

DWI, Diffusion weighted Image

ADC, Apparent Diffusion Coefficient

PFS, Progression-Free Survival

GLCM, Gray Level Co-occurrence Matrix

GLDM, Gray Level Dependence Matrix

GLRLM, Gray Level Run Length Matrix

GLSZM, Gray Level Size Zone Matrix

NGTDM, Neighbour Gray Level Difference Matrix

LDA, Linear Discriminant Analysis

QDA, Quadratic Discriminant Analysis

KNN, K-nearest neighbour

SVC, Support Vector Machine

Table of contents

1. Introduction

This master thesis is based on a master thesis conducted at NMBU the spring of 2018 [1], where a software was developed for feature extraction, feature selection and classification of treatment outcome of head and neck cancer using image CT- and PET-sequences. This software has since then been developed further and is still under development [2]. It will be used in this master thesis to predict treatment outcome of patients with anal cancer and examine whether features extracted from MR-sequences could be useful for predicting treatment outcome. During the fall of 2018, image sequences of CT-, PET- and MR-sequences were aligned and prepared to create a data set [3] that could be used in this master project.

These projects are conducted as it is believed that medical images might be useful for describing patient disease and predicting treatment outcome. This is done using radiomics, that is defined as the extraction of a large number of image features, and machine learning, for feature selection and prediction of treatment outcome using these extracted features. The goal is to find a few predictors, features that alone describe the treatment outcome or patient stage well, to obtain a better understanding of the disease and to be able to offer more personalized treatment [4].

Section 2 describes anal cancer and medical image sequences, how the sequences are acquired and how they describe the human body. Discretization, one of the processes that the image sequences underwent during feature extraction, and linear correlation, the measure that was used for examining the effect of choosing different extraction settings, are also described in this section. The settings that were changed were extraction dimension, tumour mask for CT-sequences and discretization level. In section 3, the image data and outcome data and how they were processed were described. This section also includes the chosen feature extraction settings, how the final data sets were created and the model selection process. Section 4 presents the results, while section 5 discusses them. The last sections include the conclusions from this project and appendices.

2. Theory

2.1. Cancer and anal cancer

Cancer is the definition of diseases where cells divide abnormally in the body. These cells form a tumour that interferes with neighbouring tissue and might spread to other parts of the body. Tumours that form due to spread are called metastases. [5]

Anal cancer is cancer located in the anal canal. It a rather rare form of cancer that makes up approximately 2 % of new cancer incidents [6]. It is associated with HPV, affects more women than men and is more common among elder age-groups [6]. The survival rate is approximately 80 % [7].

2.2. Medical images

2.2.1. CT

2.2.1.1. Introduction
Computed Tomography (CT) uses the attenuation of X-ray beams that interacts with atoms to create a representation of the body. CT is often used to give an anatomical representation of the body as the imaging method distinguishes well between bone and soft tissue. CT is useful for locating cancer tumours but exposes the patient to ionizing radiation [8]. The CT-examination itself can last 10-30 minutes, but it might be required that patients that have been examined with contrast agents must wait 30 minutes before leaving the hospital [9].

2.2.1.2 Principles
X-ray beams consist of highly energetic photons. During a CT-examination, X-ray beams are sent through an intersection of the body. The photons are entirely or partly absorbed while interacting with the atoms in the intersection before the beams are detected after having passed through the body. The amount of photons absorbed is used to create the CT-images.

The absorption of an X-ray beam is dependent on the tissue and the tissue thickness in the X-ray beams trajectory. This can be found using Lambert-Beer's law [10],

$$I = I_0 e^{-\mu x},$$

where I_0 is the intensity of the beam before the interaction between the photon and the tissue and I is the intensity of the X-ray beam after interaction. μ $[cm^{-1}]$ is the attenuation coefficient of the tissue and x $[cm]$ is the distance the photon has travelled in the tissue. The attenuation coefficient μ is dependent on the tissue density [10] [11].

Attenuation is caused by different types of photon-atom interactions. Which interaction that occurs is dependent on the initial photon energy [10] and the atom it interacts with. The three most common types of photon-atom interactions during CT are:

- **Photoelectric effect**: A photon is entirely absorbed resulting in the release of an electron from the atom. The photon energy must be larger than the binding energy of the electron, which is the energy that is required to release the electron from the atom. The difference of the absorbed energy and the binding energy will become the kinetic energy of the electron. [11]
- **Compton scattering**: A photon is not entirely absorbed and results in an electron being released from the outer shell of the atom. The electrons in the outer shell are easier to release, as they have less binding energy than electrons in inner shells. The photon will change direction to conserve momentum [11], in addition to having transferred some of its energy.
- **Pair production**: A photon with high energy is converted into an electron and positron near an atom. The positron might later react with another electron and annihilate, converting into two γ-photons that travel in opposite directions. These photons might also interact with other atoms via Compton scattering and photoelectric effect. [11] [12]

During both photoelectric effect and Compton scattering, atoms are ionized, as an electron is released [10].

2.2.1.3. Image reconstruction

The purpose of CT is to find the attenuation coefficients along the trajectories of the photons to find the attenuation coefficients of every point in the intersection [8]. A detected signal only indicates the total amount of energy the photon has lost, and not the energy lost in every point in its trajectory. Therefore, photons are sent through the intersection from all directions, so that the detected signals can be combined to estimate attenuation coefficients along the trajectories. These calculations are computationally expensive and must be performed by a computer.

The initial CT-images after detection and computation are called sinograms and do not give a comprehensive representation of the body slices. The sinograms must be further transformed using reconstruction algorithms. These are often specifically designed depending on what part of the body that has been examined, what the patient has been examined for and what artefacts that need to be corrected for. Example of factors that might cause artefacts in CT-images are metal in the body and the presence of bone [13].

The slices are 2D-representations that together create a 3D representation of the body. The intensity values in the images are given in Hounsfield units (HU), dimensionless units that are linearly derived from the estimated attenuation coefficients. In the Hounsfield scale, water should have value 0 and air should have value -1000. A CT-image representing a slice in the body can often have HU-values from -1000 (air) to 2000 (dense bone) [14]. High HU-values and tissue with high attenuation coefficients, appear bright in a CT-image. A CT-image is shown in Figure 2.1.

13
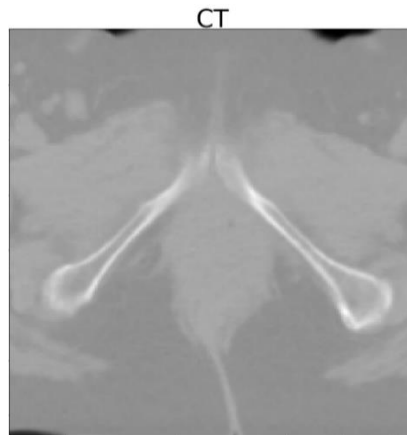
Figure 2.1: *Slice 16 in CT-sequence from patient 3.*

2.2.2. PET

2.2.2.1. Introduction

PET, Positron emission tomography, is an imaging method that can be used to quantify metabolic and molecular processes in the body [15] [16]. PET exposes the patients to ionizing radiation like CT and is a useful tool for detecting and estimating the aggressiveness of cancerous cells [11]. The PET-examination itself lasts 30-50 minutes but require that the patient meet up at the hospital several hours beforehand [17].

2.2.2.2. Principles

Approximately one hour before examination [17], a radioactive compound that can emit positrons is injected into the patient. A compound that is often used for examination of cancer is ^{18}F-deoxyglucose, glucose where one of the original components of glucose has been replaced with the unstable radionuclide ^{18}F [13]. After injection, the concentration of the injected compound will become higher in areas with high metabolic activity, including areas with cancerous cells [11].

^{18}F is unstable with a half-life of 110 minutes [13] and will eventually decay by positron emission (β^+ decay). A positron at rest annihilates with an electron such that two 0,511 MeV photons are emitted in opposite directions [11]. These photons are detected by the PET-machine and used to create the PET-images.

2.2.2.3. Image reconstruction

The first step of image reconstruction is to determine the location of the annihilations. This is done by pairing two signals that are detected within a short time interval by detectors that are placed in a ring around the patient. This ring is detects signals from an intersection or slice in the body.

Ideally, the location of an annihilation is along the direct line between the detectors that have detected the signals. This direct line is called the line of response (LOR) and the exact location along this line is determined by the time interval between the detected signals. A ring of detectors that detect the signal from an annihilation is illustrated in Figure 2.2.
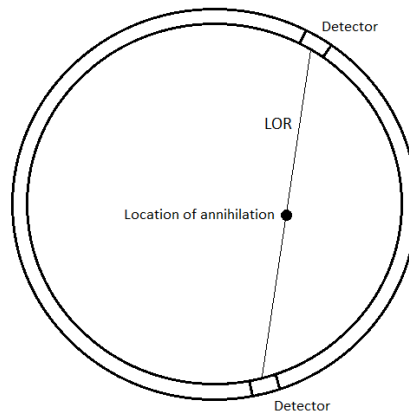
Figure 2.2: *Simple illustration of the detection of PET-signals. The ring of detectors surrounds the object that is examined and detects γ-photons from the annihilations. This illustrates the ideal case, where the location of the annihilation lies on the line of response (LOR). The detectors that do not detect a signal or the object of examination is not included in this illustration. This figure is a simplification of Figure 4.5 in [16].*

Factors that must be corrected for in the resulting PET-images are photon-atom interactions and the occurrence of more than two PET-signals within a short time interval. The photon-atom interactions attenuate and might deflect, change the direction of, the photons from the annihilations. The deflection of γ-photons and occurrence several PET-signals makes the location of the annihilation more uncertain [18]. Therefore, a PET-examination is often done in combination with CT or MR, to help correction, as these can give knowledge about the anatomical structure in the body [16].

The intensity values in the final PET-image indicate the number of signals detected from the specific area the pixel or voxel represents [18] and the radioactivity concentration in the specific area [13]. Thus, areas with high metabolic activity will appear brighter in the PET-images. A PET-image from the data set for this project is shown in Figure 2.3.
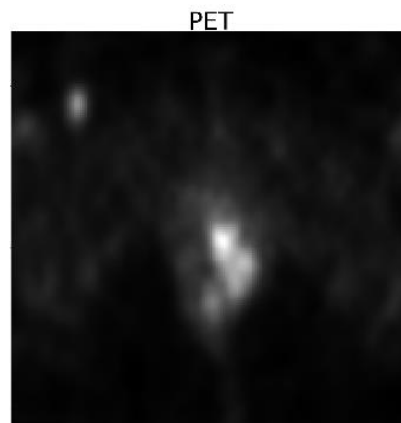
Figure 2.3: *Slice 16 in PET-sequence from patient 3.*

2.2.2.4. Standardized uptake value (SUV) *[19]*

A common method for making PET-images between different patients more comparable, is by correcting the intensity values in the images for patient weight and injected dose. This is done by replacing the PET-values in the PET-images by the standardized uptake values (SUV). SUV is given by the formula

$$SUV = \frac{C}{D/w},$$

where C is the radioactivity concentration [kBq/ml], D is the injected dose [kBq] and w is patient weight [g]. SUV becomes unitless because it is assumed that 1g = 1ml [13].

SUV does not correct for all factors that affect variation of glucose uptake between patients. Factors that affect the glucose uptake other than weight and injected dose are the patients initial glucose level and amount of body fat [13].

2.2.3. MRI

2.2.3.1. Introduction

MRI, magnetic resonance imaging, is a medical imaging method that exploits the magnetic properties of the hydrogen nuclei and the abundance of hydrogen to create a representation of the human body. An MR-image can give both different representations of the body, such as T1-weighted, T2-weighted and diffusion weighted MR-images, by changing the parameters within the imaging process. Unlike PET and CT, MRI does not expose the patient to ionizing radiation [3]. An MR-examination can last between 20-60 minutes, depending on the imaging parameters and the number of slices [20] [21].

2.2.3.2. Principles

A simple model of the hydrogen nucleus is that of a single proton spinning on its own axis, thereby creating a magnetic moment μ_m as the hydrogen nucleus is a charged particle in motion [22]. The proton has two possible energy states, called spin up and spin down, giving two possible spin directions. These spin directions decide the direction of the magnetic moment of the protons [23]. See Figure 2.4 below.
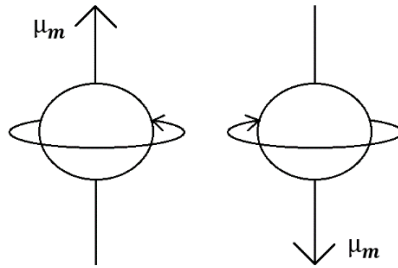
Figure 2.4: Two protons with different spin directions have magnetic moments that point in different directions. The vertical arrows show the direction of their magnetic moment $\overrightarrow{\mu_p}$, while the circular arrows show the spin direction. This figure is inspired by Figure 2-6 in [23].

The magnetic moments of protons that are not placed in an external magnetic field will point in different directions and not create net magnetization as the magnetic moments of the protons cancel each other out. This is illustrated in Figure 2.5.
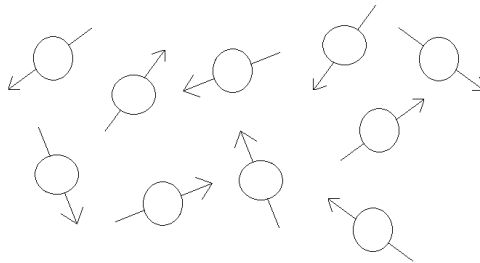
Figure 2.5: Protons not affected by an external magnetic field. The proton magnetic moments are pointing in arbitrary directions and cancelling each other out.

Protons placed in an external magnetic field $\boldsymbol{B_0}$, will attempt to align the direction of their magnetic moment μ_m along the direction of $\boldsymbol{B_0}$. This is because the protons experience a torque from $\boldsymbol{B_0}$ [22]. This alignment is not entirely successful, resulting in that the magnetic moment μ_m of a proton will be slightly tilted compared to $\boldsymbol{B_0}$. As the proton constantly experiences a torque, it starts to precess, or rotate in $\boldsymbol{B_0}$ with frequency

$$\omega_0 = \gamma B_0.$$

This frequency, ω_0, is called the Larmor frequency and is proportional to the magnitude of the external magnetic field B_0 with factor γ, called the gyromagnetic ratio with value $2.7 \times 10^8 \, s^{-1}T^{-1}$ [22]. The magnetic moment μ_m of protons in different energy states will align to $\boldsymbol{B_0}$, pointing in opposite directions. This is illustrated in Figure 2.6. Protons that are in the same energy state create net magnetization in the same direction along $\boldsymbol{B_0}$, as illustrated in Figure 2.7. A slight majority of the protons are in the lower energy state, giving that the overall net magnetization points in the same direction as $\boldsymbol{B_0}$. This magnetization is called $\boldsymbol{M_0}$ and is not easy to detect as it is much smaller than $\boldsymbol{B_0}$.
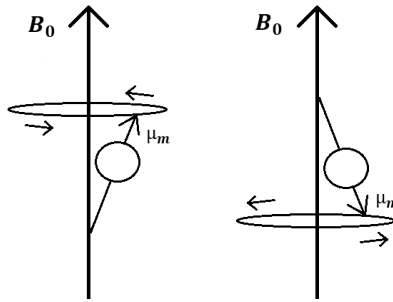
Figure 2.6: *Protons in different energy states precessing in a magnetic field B_0. B_0 is represented by the thick vertical arrows. The proton to the left is in the lower energy state, thus its magnetic moment μ_m is approximately pointing in the same direction as B_0. The proton to the right is in the higher energy state with its magnetic moment μ_m pointing in the opposite direction. The spin directions of the protons are not included in this illustration. This figure is inspired by figure 2-19 (p.26) and figure 4-8 (p.44) in* [23].
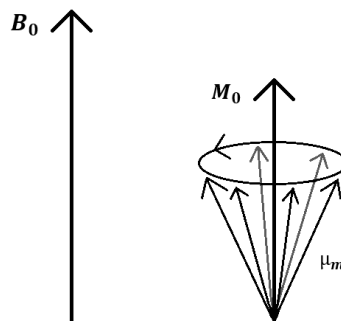
Figure 2.7: *Net magnetization M_0 created by several protons in the lower energy state that are precessing in B_0. M_0 is the vector sum of the proton magnetic moments and has the same direction as B_0, as the vector component of the magnetic moments in the orthogonal plane of B_0 cancel each other out. The arrows that are tilted compared to the vertical arrows represent the direction of the magnetic moments μ_m of several protons. The circular arrow represents the direction of precession. This figure is inspired by Figure 8.4 p. 140 in* [22].

A gradient field is a magnetic field where the magnetic strength decreases or increases linearly along the direction of the field [23]. A proton that is placed in a gradient magnetic field, will experience a torque that depends on its location in the field. Protons in different locations in the gradient magnetic field will thus precess with different frequencies.

During an MR-examination, the patient is placed in a MR-machine that creates the external magnetic field B_0 along the length of the patient body. The direction of this field is usually denoted as the z-direction in the MR-system. The x- and y-directions are orthogonal to this direction and each other, where the x-direction is usually along the horizontal axis (from patient side to side), and the y-direction is usually along the vertical axis (from patient back to front) in the body [22]. This is illustrated in Figure 2.8.

Axial images represent slices in the the *xy*-plane of the body, sagittal images represent slices in the *xz*-plane of the body and coronal images represent slices in the *yz*-plane of the body [22]. Other necessary equipment for MRI are RF- and gradient coils. RF-coils are used to both induce RF-pulses and detect change in net magnetization, while gradient coils are used to create gradient fields.
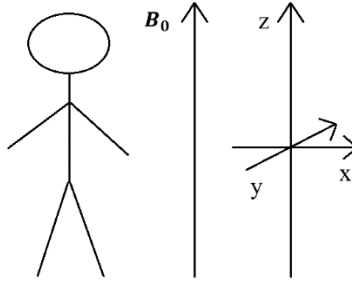
Figure 2.8: *Coordinate system in the MRI-machine and in this thesis.*

The direction of B_0 will from now be assumed to be along an axis called z, while the *xy*-plane will represent the orthogonal plane of B_0.

2.2.3.3. Excitation and relaxation

Protons in the lower energy state can be excited into the higher energy state by an external supply of energy. During MRI, excitation is done by applying radio frequency (RF) pulses with a frequency that is close to the Larmor frequency, creating resonance [23]. Protons in the lower energy state will start to precess in phase and attempt to align their magnetic moment μ_m with B_0 in the opposite direction. This means that the magnetic moments of the protons are now pointing in the same direction and that they will flip their rotation angle towards the orthogonal plane of B_0, here called the *xy*-plane, while precessing.

In the *xy*-plane, the protons will also attempt to precess along the magnetic field that is induced by the RF-pulse. This precession frequency, ω_1, is proportional to the magnetic field, B_1, induced by the RF-pulse and thus much smaller than the precession frequency, ω_0, in B_0. That the protons precess in phase creates a detectable net magnetization, M_1, in the *xy*-plane [23].

When the RF-pulse is switched off, the protons stop precessing in phase and return to equilibrium, the state they were in before the RF-pulse was applied. This process is called relaxation and is what is used to create det MR-images [22].

Two important relaxation types in MRI are T1- and T2-relaxation. T1-relaxation is also called longitudinal relaxation or spin-lattice relaxation and involves energy loss due to interaction with surrounding tissue [22]. This causes the protons that were affected by the RF-pulse to return to their original energy state and once again align their magnetic moment $\vec{\mu_p}$ with the external magnetic field B_0 [23].

T2-relaxation is also called the transverse relaxation or spin-spin relaxation [23]. The protons start to dephase after the RF-pulse is switched off. This means that the protons start to precess with different frequencies. Dephasing has two main reasons; differences in strength of magnetic moment between nearby protons, called spin-spin interactions, and differences in strength of magnetic moments between nearby protons due to inhomogeneities in the external magnetic field [22]. T2*-relaxation involves dephasing due to both spin-spin interactions and magnetic field inhomogeneity while T2-relaxation mostly involves dephasing due to spin-spin interactions. T2*-relaxation happens faster than T2-relaxation [23].

Dephasing is caused by that nearby protons are precessing with slightly different frequencies [22] and is enhanced by that the magnetic moments of nearby protons affect each other and inhomogeneities in the magnetic field. This cause an overall magnetization loss in the *xy*-plane, as M_1 weakens. This change in magnetization can be detected by receiver RF-coils in the MR-machine.

T1- and T2-relaxation are separate processes that occur simultaneously [23]. Different tissue types have different T1- and T2- relaxation times that are used to distinguish between the tissues. The T1 value of a tissue can be understood to be the time rate that magnetization M_0 along the external magnetic field, recovers, while the T2 value of a tissue can be understood as the time rate the magnetization along the *xy*-plane, M_1, weakens. T2-relaxation happens faster than T1-relaxation.

2.2.3.4. Spatial encoding
It was previously mentioned how protons are affected in a gradient magnetic field. During MRI, three gradient fields are applied to help determine the position of the source of the detected signals [22].

A gradient applied along the external field B_0, suitably called the slice select gradient, makes it possible to differentiate between different slices in the body. A specific slice in the body is examined by applying RF-pulses with frequencies that are close to the Larmor frequency of the magnetic field strength in the slice.

Two gradients are applied along the x- and y- axis of a slice to help determine the position of the detected signal within the slice. The gradients "encode" the protons, such that the protons in a given *xy*-position in the slice precess with a unique combination of frequency and phase [23]. It is necessary to encode all frequencies with all phases to obtain the entire frequency distribution in the area that is examined.

In axial images the slice select gradient is applied along the *z*-axis, the frequency encoding gradient is usually applied along the *x*-axis and the phase-encoding gradient is applied along the *y*-axis [23].

2.2.3.5. Pulse sequences
A pulse sequence is a series of pulses that is applied during MRI [23]. Basic components of a pulse sequence are RF-pulses and gradient pulses, where each pulse sequence is defined by the type, order and timing of its components to obtain the desired MR-images [22]. A pulse sequence is not enough during MRI, as the spatial encoding gradients must be varied to obtain the entire frequency distribution in the body [23], and that this must be repeated for the desired number of slices.

Two important timing properties of pulse sequences are TR and TE. TR, the repetition time, is the time interval between the application of two RF-pulses. The second RF-pulse flips the protons back into the *xy*-plane, often before the protons have managed to return to equilibrium. TE, the echo time, is the time interval between the application of a RF-pulse to acquisition of a MR-signal.

2.2.3.6. Contrast enhancement

A T1-weighted image requires short TR and short TE, while a T2-weighted image requires long TR and short TE [22]. Another means of obtaining images with better contrast is to apply contrast agents, like gadolinium. These reduce the relaxation times of the tissues they encounter after injection, such that differences between the signals obtained from different tissues are more apparent [22].

2.2.3.7. Image types

Using different pulse sequences, different image types highlighting specific properties in the body can be obtained [22]. In this project, the following image types where available:

2.2.3.7.1. T2-weighted images

A T2-weighted MR-image enhances the T2-relaxation of the tissues. Bright areas in the images represent areas with longer T2-constants (fluids) and grey and dark areas represent areas with shorter T2-constants (water-based tissue and fat-based tissue) [22]. The right image in Figure 2.10 is a slice from one of the T2-weighted sequences that is used in this project.

2.2.3.7.2. Diffusion weighted images

A diffusion weighted image shows the diffusion of water molecules. These images are obtained by applying a dephasing gradient pulse followed by a rephasing gradient pulse [24]. The molecules that have not moved between the application of the gradient pulses will be completely rephased. The molecules that have moved considerably will not be completely rephased and a lower signal will be obtained from the area they originally belonged to.

The degree of diffusion that the image represents, is described by its b-value. The b-value is determined by the time interval between the application of the dephasing gradient and rephasing gradient and the magnitude and duration of the gradients [25]. A pulse sequence with a larger b-value detects molecules that are moving slower than the molecules that a pulse sequence with a lower b-value detects. Molecules that are "trapped" during the application of the two gradients will thus give a stronger signal and appear bright in the resulting images [24] [26]. The images in Figure 2.9 are diffusion weighted images of different b-values.

2.2.3.7.3. ADC-mapping

An ADC-mapping is created by combining diffusion weighted images of different b-values. An ADC (Apparent Diffusion Coefficient) value is computed for all corresponding pixels or voxels in the diffusion weighted images or image sequences [27] and approximates the diffusion coefficient, or the change in signal intensity for a change in the b-value [28]. The computation of the ADC-mapping gives a quantitative representation of the images while it also corrects for unwanted T2-effects in diffusion weighted images with large b-values [24] [26]. An ADC-map is shown in the left slice in Figure 2.10.
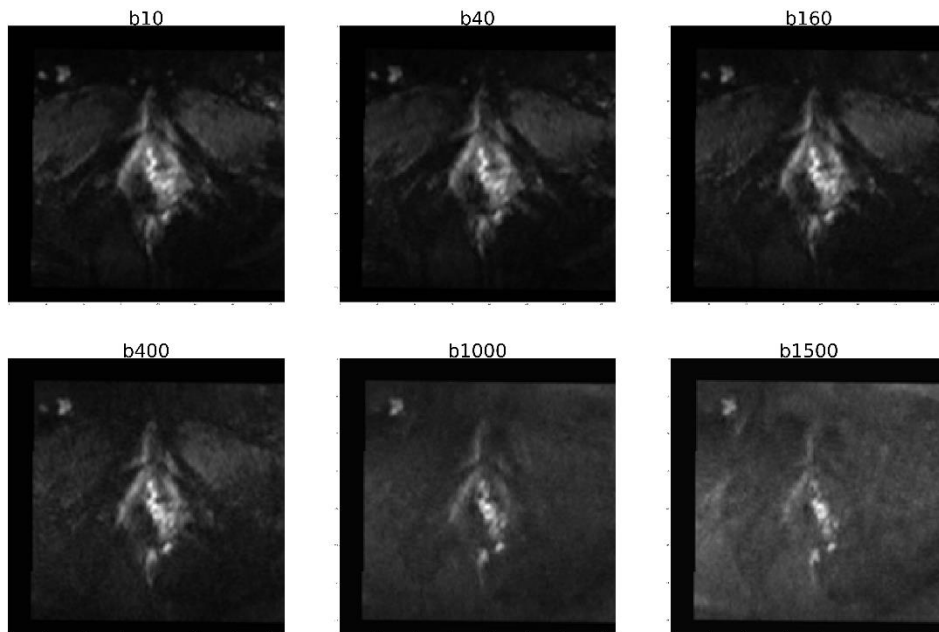
Figure 2.9: *Diffusion weighted images from patient nr. 3. The images are all slice nr. 16 in the sequence they belong to, thus representing the same slice in the patient body. The diffusion weighted images have the following b-values: b10, b40, b160, b400, b1000, b1500.*
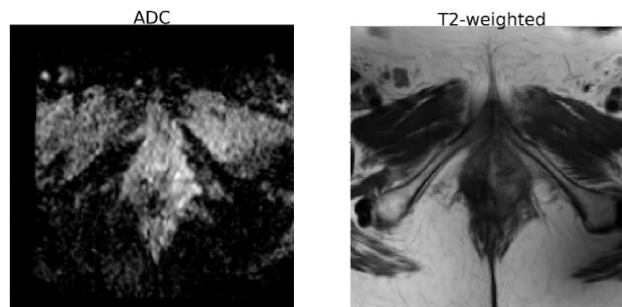
Figure 2.10: *Slice in ADC-map (right) calculated from the diffusion weighted images of patient 3 and the corresponding T2-weighted MR-slice (left). The slice these images represent correspond to the slice in the diffusion weighted images in Figure 2.9.*

2.3. Image discretization

Image discretization, also called image quantification, is the process of reducing the number of unique intensity values in an image or image sequence. This is done by dividing the image intensity range into equally sized intervals, often called bins, and assign all intensity values belonging to the same interval to the same intensity value. The size of these intervals, the bin width, is dependent of the discretization level, or the desired number of intensity values. If an image sequence should be discretized to only contain eight intensity values, the intensity range should be divided into eight equally sized intervals, where all intensity values belonging to one interval will be set to the same intensity value, giving eight unique intensity values in the new image.

Figure 2.11 illustrates histograms of the same image sequence, but with different discretization levels. The upper histogram gives a more detailed representation of the intensity value distribution than the lower, as it has a greater number of bins and columns in the histogram. In the second histogram, there are fewer bins, and more intensity values are assigned to each bin. The columns in this histogram are wider and taller than in the upper histogram. Figure 2.12 shows a slice from the same sequence, also with different discretization levels. The image with lower discretization level is less detailed and has a more defined transition between areas with intensity values that are distinctly different.
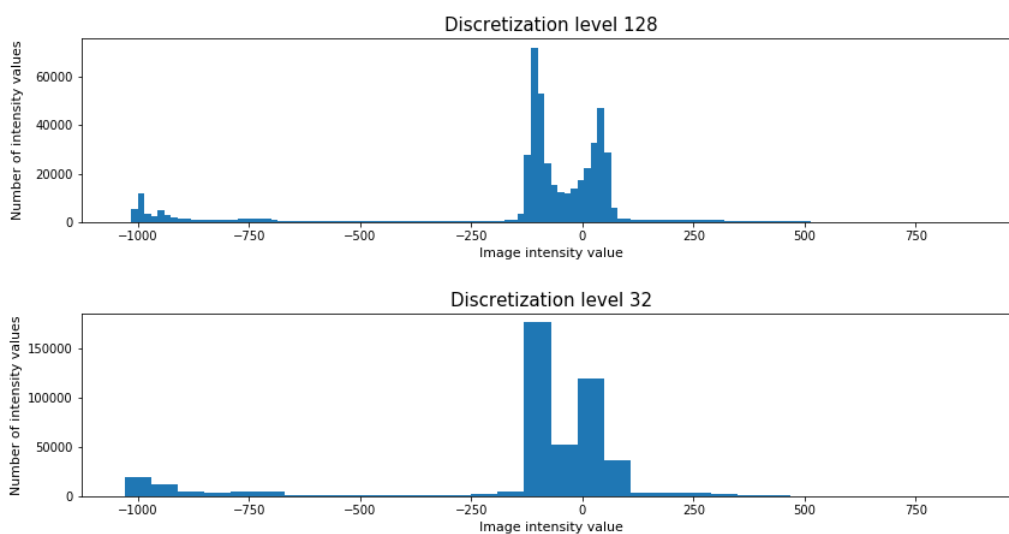
Figure 2.11: *Histogram of the HU-values in the CT-sequence of patient 45. The upper histogram shows the sequence intensity distribution with 128 bins. The bottom histogram shows the sequence intensity distribution with 32 bins.*
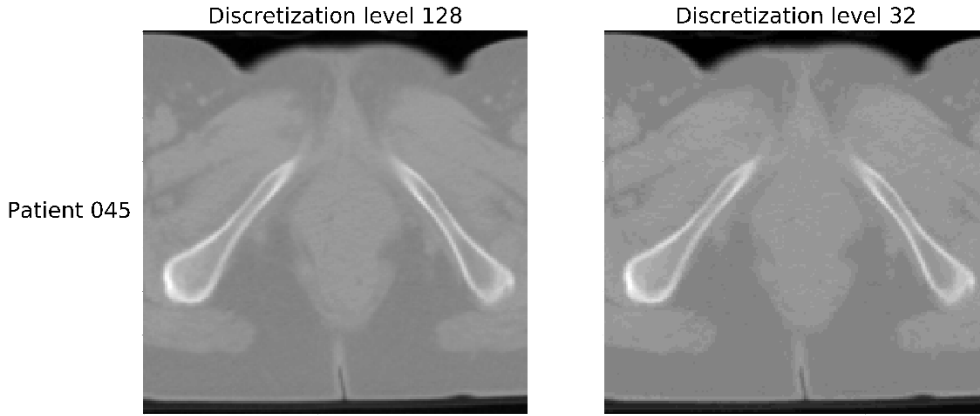
Figure 2.12: Image slice in CT-sequence from patient 45. The slice to the left has discretization level 128, while the slice to the right has discretization level 32.

2.3.1. Binning

An image or sequence can be binned by applying the formula

$$x_{binned} = \left\lfloor \frac{x}{W} \right\rfloor - \left\lfloor \frac{x_{min}}{W} \right\rfloor + 1 \tag{1}$$

on all intensity values x. Here, x_{min} is the smallest intensity value in the original image or sequence and W is the bin width. x_{binned} is the new intensity value of x [29].

2.3.2. Bin width

The bin width, W, for one single image or image sequence can be computed by the formula

$$W = \frac{x_{max} - x_{min}}{n_{bins}}. \tag{2}$$

Here, n_{bins} is the discretization level and the desired number of bins. $x_{max} - x_{min}$ is the range of the intensity values in the image or sequence. The intensity range is thus n_{bins} times larger than the bin width.

In this thesis, a fixed bin width for all sequences of one modality is calculated by a modification of formula (2). This includes intensity values from all sequences of the same modality,

$$W_{global} = \frac{\overline{\lceil x_{max} \rceil} - \overline{\lfloor x_{min} \rfloor}}{n_{bins}}. \tag{3}$$

x_{max} in formula (2) is replaced by the mean of the ceiled ($\lceil \ \rceil$) maximum intensity values of each sequence and x_{min} in formula (2) is replaced by the mean of the floored ($\lfloor \ \rfloor$) minimum intensity values of each sequence.

2.4. Linear relationship between variables

2.4.1 Pearson correlation

A Pearson correlation value expresses the direction and strength of the linear relationship between two variables. It is found by the formula

$$r_{xy} = \frac{cov(x, y)}{\sqrt{var(x)} \times \sqrt{var(y)}}. \tag{4}$$

Here $cov(x, y)$ is the covariance between x and y and $var(x)$ and $var(y)$ are the variances of the variables x and y [30].

3. Methods

3.1 Programming

Programming in this master project has been done in Python 3.6.6 *in Windows 10*. Python scripts were developed and used for image conversion, image processing, image computations, feature extraction, preprocessing of data and classification of treatment outcome. All necessary packages were installed in a Python environment created specifically for this master project to avoid difficulties concerning instalment of Python packages that were not compatible to the Python packages already installed in the Python platform Anaconda3. All Python packages can be found in Appendix A.1. It was also necessary to install a C-compiler to use the package *PyRadiomics* [31]. Also, the instalment of Ubuntu 16.04, a Linux subsystem for Windows, was necessary for running model selection with a Python package called *biorad* [2]. *biorad* is developed by Severin Langberg and is inspired from the master thesis conducted at NMBU in 2018 [1]. This package is a wrapper for, thus dependent of *PyRadiomics*. *PyRadiomics* is developed for feature extraction from medical images.

Jupyter Notebook was used for exploration, visualization and splitting the data sets into training and test data sets. Jupyter notebook is a Python interface that makes it possible to visualize the data while programming.

3.2 The data set

3.2.1 Image data

The image data set was provided by OUS, the Norwegian Radium Hospital and consisted of medical image sequences from 86 anal cancer patients who received treatment in the period December 2013 – November 2016. The sequences represent the pelvis and abdomen [32] of the patients in three dimensions. The axial medical images were equally spaced along the patient body and each represent a slice in the body.

3.2.1.1. Modalities and GTV
Every patient had at most been examined by three different modalities, CT, PET and MR, in two stages; the beginning of treatment and approximately two weeks into treatment. There were three types of MR-sequences; a T2-weighted sequence, 12 diffusion weighted MR-sequences of different b-values, and an ADC-maps that was derived from the diffusion weighted MR-sequences. This means that there were at most 16 different types of medical images per patient available for treatment classification in each stage; CT, PET, T2-weighted MR, 12 diffusion weighted images with b-values b0, b10, b20, b40, b80, b160, b200, b400, b800, b1000, b1200 and b1500, and ADC.

All patients had in addition been examined by CT before treatment for dose planning. The tumour delineation of interest, the gross tumour volume (GTV), had been delineated in these sequences by oncologists and converted into a binary mask. The GTV defines the border of the tumour volume [33] and was used to extract features from the tumour volume in the image sequences.

3.2.1.2. Early processing

As all image sequences were not taken simultaneously, the GTV-mask would not necessarily define the tumour area in image sequences that belong to the same patient but was taken at a different time. Therefore, all image sequences had been registered beforehand, so that all image sequences belonging to one patient should be aligned to the CT-sequence for dose planning and the GTV should define the tumour area in all sequences. The process of alignment is described in more detail in [3]. Figure 3.1 shows the available image modalities and the GTV mask after the images had been aligned.
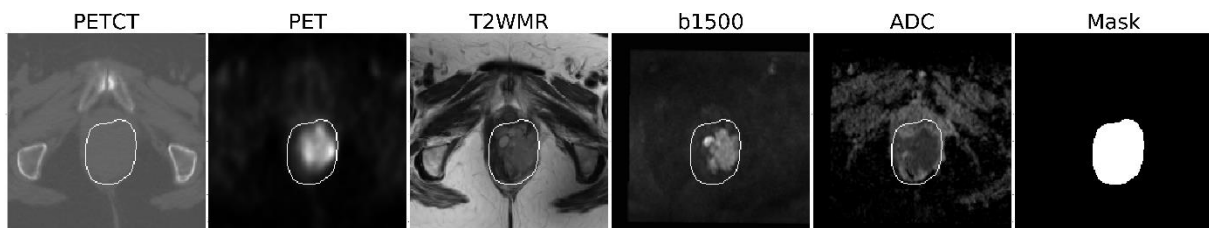
Figure 3.1: Corresponding slices in medical image sequences of different modalities. Modalities from the left are CT, PET, T2-weighted MR, diffusion weighted MR with b-value 1500 and ADC. The last image is the mask of the GTV in the corresponding slice. The tumour area in the first five images were delineated using the mask. These slices belong to patient 3.

After registration, all images had also undergone resampling, so that all image sequences had the same resolution ($1 \times 1 \times 3 \ mm^3$).

3.2.1.3. Available modalities

Not all patients had all image modalities available. It was found that only 36 of the 86 patients had MR-sequences available. Of these, all MR-sequences were only available from the base stage, at the beginning of treatment. As MR was the initial modality of interest, it was chosen to use images from the base stage to ensure the highest possible number of patients in a data set containing MR-sequences.

3.2.1.4. Missing image data

The MR-sequences generally contained fewer slices than sequences of other modalities, giving some MR-slices without information after image registration. This was assumed to have been caused by the image registration, correcting for different patient position during acquisition of images of different image modalities. For some patients, this meant that some slices that should contain tumour delineation did not contain information pertaining to the tumour due to discontinuities, or "cuts". The area within the tumour delineation in some slices were thus

partially or entirely blank. This was especially the case for several diffusion weighted MR-sequences, and hence also several ADC-maps and some T2-weighted MR-sequences.

The slices containing tumour delineation that were not entirely present in the MR-sequences were usually located in the extremities of the body normally examined for anal cancer, in the sequence beginnings and/ or ends. Figure 3.2 shows slices in an ADC-map that should all contain information about the tumour. This is not the case for the first two slices due to a "cut". None of the CT- and PET-sequences contained blank or incomplete tumour slices.
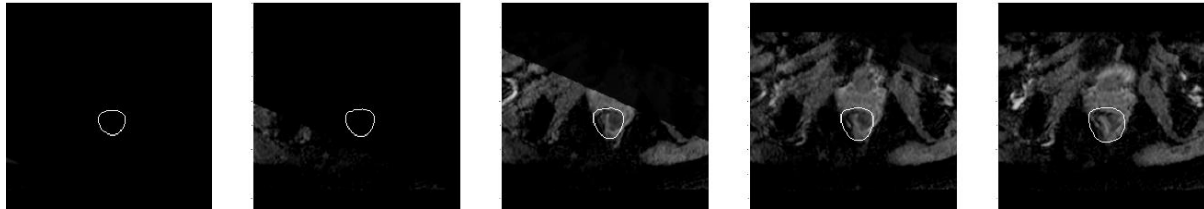
Figure 3.2: *Slices in ADC-map from patient 7 where two slices are missing information within the tumour delineation.*

3.3 Processing

3.3.1. Conversion and organization

The image files for each patient were stored in separate folders, with subfolders that contained all images from the different stages and a folder for the tumour mask. To adapt to the Python program for feature extraction and classification, *biorad*, it was necessary to create new folders that contained all patient sequences of the same modality and stage. This was done using the Python package *os*.

The image sequences were provided in the MATLAB file format and converted to the NRRD file format using the Python packages *pynrrd* and *NumPy*. *NumPy* was used for transposing the images and *pynrrd* was used for image conversion. Furthermore, it was necessary to include the image metadata in the NRRD files, as this was not done automatically in the conversion. This was mainly so that *PyRadiomics* would register the resolution of the images and use it as base if operations like resampling should be enabled. The converted image sequences were named with modality and patient ID.

3.3.2. Cropping

All image sequences were firstly cropped to only contain slices with information pertaining to the tumour. This was not necessary considering that the program for feature extraction alone applies the masks on the images, such that features are extracted from the region of interest, the tumour in this case. However, this made it easier to further remove the MR-slices with missing

or only partially present information within the tumour delineation in the MR-images. The tumour slices that were not intact at each end of the MR-sequences were manually found by inspection of the sequences. The number of slices to remove from each end in each sequence were stored in a *Microsoft Excel* file, read in a *Pandas* data frame and removed from the sequences. This was done to avoid further reduction of the number of patients. In Figure 3.2 the first two slices would be removed as there is lacking information within the tumour delineation.

All cropping was done using functions written in Python called *tumour_crop* and *slice_crop.* These are found in Appendix B.1.

3.3.3. Removal of air and abnormally high HU-values in CT-sequences

Most CT-sequences contained air within the tumour delineation. This can be seen in Figure 3.3. The slices in this figure belongs to two different patients where air is included in the tumour delineation. The upper slices show air in the entry of the anal canal and the lower slices show air packages located further in the anal canal.

As air was not considered as information of interest for classification, new tumour masks were created for the CT-sequences. Voxels with HU-values outside the range [-150, 200] within the tumour delineation were set to one. This created a mask for air and darker areas within the tumour delineation, HU-values lower than -150, and HU-values higher than 200. The removal of HU-values higher than 200 was suggested by Professor Eirik Malinen from UiO [34], as HU-values higher than 200 are not considered to give clinical information. The voxels with HU-values in the range [-150, 200] belonging to the two patients in Figure 3.3 are delineated in the second column of the figure. The mask created in this step, would thus cover the area that is included in the first column but not included in the second column of the figure.

Further, the areas set to one in this mask were dilated two times with a square structuring element with shape (3,3) in the *xy*-plane. This was done so that the resulting mask also should include the voxels nearby the voxels with HU-values not included in the range [-150, 200], as these might be partly representing or affected by air. This was suggested by Espen Rusten, a medical physician and former PhD at UiO [35]. The modified tumour mask was only made for the CT-images as the occurrence of air could vary between the acquisition of different images. The dilated mask was combined with the original mask for the patient, by setting the intensity values that were equal to one in the dilated mask to zero in the original mask. The last column in Figure 3.3 shows examples of these resulting masks. These masks would only be applied on CT-sequences.

The scripts and functions for creating the modified tumour masks are found in Appendix B.2.
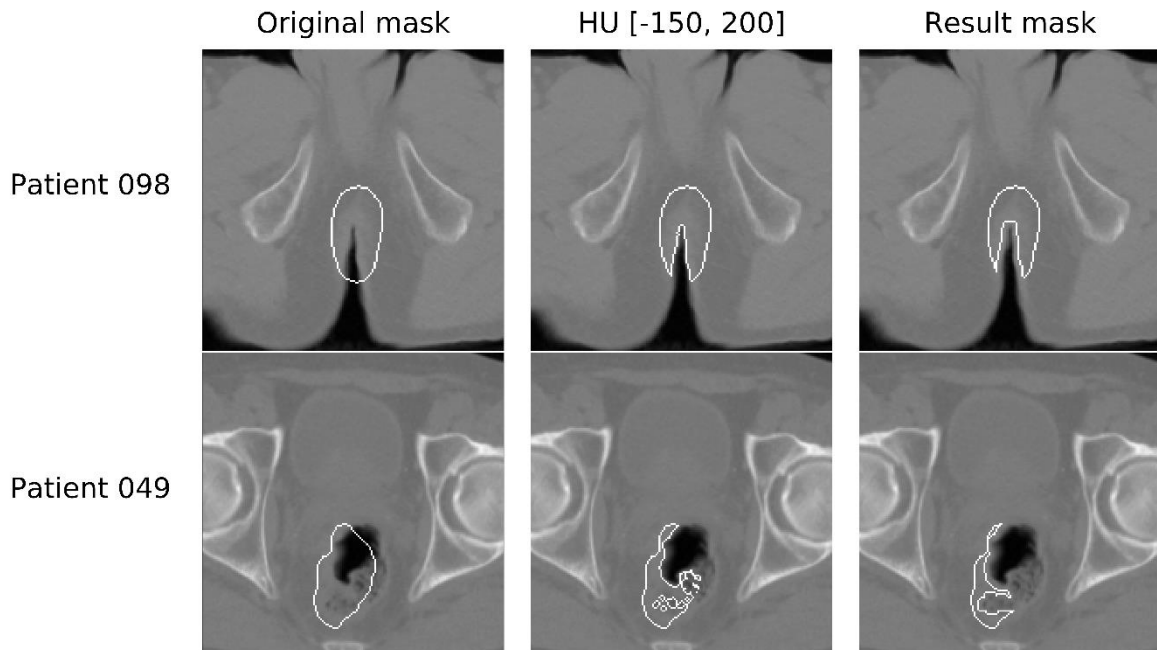
Figure 3.3: *Contours of different masks applied on a CT-slice belonging to patients 98 (top) and 49 (bottom). The left images show the contour of the original tumour delineation, the middle images show the contours of a mask where HU-values outside the range of [-150 - 200] within the original tumour are not included. The right images show the contour of the masks where the areas that were excluded in the second column are expanded (dilated) by the square structuring element (3,3) and two dilations.*

3.4. Feature extraction

The Python package *biorad* was used for feature extraction. One way of giving the parameters for feature extraction with *PyRadiomics* is by providing a parameter file for input. Parameter files were created for each combination of modality, discretization level and extraction dimension used in this project. The parameters that were used in this project are described below. An example of a parameter file is provided in Appendix C.1.

3.4.1. Feature extraction parameters

3.4.1.1. Image discretization
Discretization is necessary for simplifying the extraction of texture features. The documentation in *PyRadiomics* recommends using a fixed bin width for discretization of the intensity values in all images of one modality. This is for making the intensity ranges comparable between patients [36]. In this project, bin widths for discretization levels 8, 16, 32, 64, 128 and 256 were calculated from the tumour area.

The bin widths used in this thesis were calculated with function *bin_widths_tumour* in Appendix B.3. It is based on Equation (3). All bin widths are listed in Table 3.1.

Modality/Discretization level	8	16	32	64	128	256
CT with modified mask	29.5382	14.7691	7.3845	3.6923	1.8461	0.9231
CT	90.6528	45.3264	22.6632	11.3316	5.6658	2.8329
PET	2.3090	1.1545	0.5773	0.2886	0.1443	0.0722
ADC	302.6736	151.3368	75.6684	37.8342	18.9171	9.4586

Table 3.1: *Table of bin widths at different discretization levels for CT with the modified tumour mask, and CT, PET, and ADC with the original mask. The bin widths are rounded to four decimals and were calculated based on the intensity values within the tumour in the sequences.*

3.4.1.2. Voxel array shift

The first order features *Energy, Total Energy* and *RMS*, from *PyRadiomics*, are particularly sensitive to negative values [29]. As the CT-sequences contained both positive and negative intensity values within the tumour delineation, a parameter called voxel array shift was defined in the parameter files for extracting features from the CT-sequences. This was to ensure that all HU-values within the tumour delineation were shifted to belong to a range only containing positive values when extracting these features.

The voxel array shift values were set to the floor, $\lfloor x \rfloor$, of the lowest intensity value in tumour tissue in all sequences of each modality. These values are listed below in Table 3.2.

Modality and mask	Voxel array shift
CT with original mask	1024
CT with modified mask	150

Table 3.2: *Voxel array shifts for the CT-sequences where the features will be extracted with the original mask (upper) and with the modified mask that does not include HU-values outside of the range [-150, 200] (lower).*

It was not necessary to set a voxel array shift for the ADC- and PET-sequences, as they only contained positive intensity values.

3.4.1.3. Distance between neighbours

All textures features belonging to the NGTDM and GLCM texture classes were extracted by considering voxels located with distance 1 voxel from each other as neighbours [29]. The features that were extracted are briefly described in section 3.1.2.6.

3.4.1.4. Removal of additional features
All features concerning

- Image of mask file location
- Program versions during extraction (for example *pyradiomics*, *numpy, Python*)
- Lists of settings and filters for extraction
- Hash, resolution and size of mask and image files
- Minimum, mean and maximum intensity value in entire sequences
- Number of voxels within mask
- The centre of mass of the mask and its location in the sequences

were not extracted by setting the parameter *additionInfo* to False in the parameter files.

3.4.1.5. Extraction dimension
It is recommended to compute texture features from images with isotropic voxels [37]. As the images had been resampled to resolution $1 \times 1 \times 3 \text{ mm}^3$, two options in *PyRadiomics* were considered; to include a parameter called Force2Ddimension or define that the images should be resampled in the parameter files. Both options were performed, also to examine the impact of the choice.

The images had already been resampled during registration in earlier processing, first for registration and then to ensure that all image sequences had the same resolution [3]. If the images should be resampled again for extraction, it was considered that the best option would be to resample the sequences to resolution $1 \times 1 \times 1 \text{ mm}^3$, thereby keeping the higher resolution in the *xy*-plane. In this project, resampling was performed with the Bspline interpolator from the *Sitk* package.

The parameter Force2Ddimension ensured that features were extracted from the images in two given dimensions. In this project, this parameter was set so that features would be extracted from the slices, or the *xy*-plane, as $x = y = 1$ and $z = 3$.

3.4.1.6. Extracted features
In this project, mainly three types of features were extracted: Shape features, first order features and texture features.

Shape features describe the shape and size of an object and are independent of the sequence intensity values, unlike first order and texture features. First order features are often more common statistical measures that describes the intensity value distribution of the sequences, while textural features describe the spatial distribution of the intensity values in the sequences.

Five texture feature classes were used in this project: GLCM, GLDM, GLRLM, GLSZM and NGTDM. These features were extracted from matrices describing texture in the sequences. The definitions of these matrices are given below.

- GLCM, Gray Level Co-Occurrence Matrix, describes the co-occurrences of pairs of intensity values. The image components, pixels or voxels, that these intensity values belong to, have a given spatial relationship, meaning that the first component of given intensity value will be located within a certain distance and direction relative to the second component of a given intensity value [38] [29].
- GLDM, Gray Level Dependence Matrix, finds the occurrences of neighbouring voxels that satisfy a condition concerning the center voxel [29].
- GLRLM, Gray Level Run Length Matrix, finds the runs of equal intensity values in a given direction in the sequences [29].
- GLSZM, Gray Level Size Zone Matrix, finds the number of connected voxels with the same intensity value. Two voxels are connected if they have the same intensity value and are considered as neighbours [29].
- NGTDM, Neighbour Gray Tone Difference Matrix, contains the occurrence of an intensity values and the fraction of occurrence of an intensity value, the sum of the differences between all intensity values of one value and the mean intensity values of its neighbours [29].

Most available features from *PyRadiomics* were extracted from the image sequences.

There were

- 14 shape features,
- 18 first order features,
- 23 GLCM features,
- 16 GLRLM features,
- 16 GLSZM features,
- 14 GLDM features
- 5 NGTDM features.

Shape features were extracted separately from first order and texture features. Shape features were extracted once from the original tumour mask, while the 92 first order and texture features were extracted for every combination of modality, mask, discretization level and extraction dimension.

The texture feature *SumAverage* in the GLCM class was not extracted due to a deprecation warning informing that this feature was identical to another feature called *JointAverage* [29]. A complete list of the extracted features can be found in Appendix C.2.

At the end of the extraction with *biorad*, features with

- missing values
- zero variance (features that had the same value for each patient)

were removed. The missing values were features related to the extraction process, called reader and label. Features that were removed due to zero variance, was the ADC-feature Minimum, as the minimum intensity value within the tumour in all ADC-maps were equal to zero.

4432 features were extracted in total ((92 first order and texture features × 4 modalities × 6 discretization levels) + 14 shape features – 6 ADC-feature Minimum removed due to zero variance) × 2 extraction dimensions from the 36 patients with CT-sequences, PET-sequences and ADC-maps. Here, the CT-sequences with the modified tumour masks are considered as a forth modality as the modified mask was only applied for the CT-sequences.

3.4.2 Feature files

The features were stored in separate CSV files each containing shape features or first order and texture features extracted from a specified modality with a specified discretization level. Feature files containing features extracted in different dimensions were separated in folders.

The files containing first order and texture features were named with the modality and the discretization level of the image sequences the features were extracted from. The column names in these files were also changed to not only contain feature class and feature name, but also modality and discretization level so that it would be possible to differentiate between features from different discretization levels when the files later would be merged to form data sets.

3.5. Examination of features extracted with different parameter combinations

The effects of extracting features with different extraction dimensions, tumour masks and image discretization levels were explored by examining the Pearson correlation values between features extracted with different image settings. Pearson correlation is given by formula (4) and were possible to compute with the Python package *Pandas*.

Pearson correlation values were computed between:

- features extracted using 2D extraction and 3D extraction,
- features extracted from CT-sequences using either the original tumour mask or the modified tumour mask excluding HU-values in the range [-150, 200],
- features extracted from image sequences of different discretization levels.

These correlation values were computed between corresponding rows and columns in the feature files consisting of the features extracted with the same image settings except for the

setting of interest (tumour mask or extraction dimension). A row in these feature files contained all features extracted from the sequence of a patient, here called patient features, while a column contained the same feature extracted from all the patients in the data set.

The correlation values computed between features extracted from image sequences at different discretization levels were between corresponding columns of features extracted from the same modality but at different discretization levels.

3.6. Processing and feature extraction for a PET/CT-data set

It was decided to also create a data set of PET- and CT-features to obtain a data set with more observations than 36. 81 of 86 patients had CT-sequences did not have artefacts that were visible for an untrained eye and were acceptably registered. Modified tumour masks were created for these CT-sequences as described in section 3.3.3 and features were extracted with the parameters described in section 3.4.1.

3.7. Data sets

3.7.1 Choice of features for data set

It was chosen to create data sets with first order and texture features extracted with one discretization level and shape features. Discretization levels 128 was chosen for the CT-sequences and discretization level 16 was chosen for the PET-sequences and the ADC-maps. For the CT-sequences, it was chosen to use features extracted with the modified tumour mask instead of the original tumour mask. All chosen features had been extracted in 2D.

3.7.2. Final data sets

Two data sets were thus created. One data set containing 289 shape, PET-, CT- and ADC-features from 36 patients and one data set containing 198 shape, PET- and CT-features from 81 patients. The overview of these data sets is given in Table 3.3. These data sets will from now be described as data set 1 and data set 2.

Data set	Modalities	Number of patients	Number of features
1	PET, CT, ADC	36	289
2	PET, CT	81	198

Table 3.3: *Overview of the two data sets created in this project*

3.8 Endpoints

Three endpoints were available for this project: Complete response, recurrence and metastasis. Complete response describes whether the patient shows signs of cancer after treatment (1) or not (0). Cancer can still reoccur for patients that initially have positive complete response after treatment [39]. This variable was inversed to follow standard notation, such that positive complete response was set to 0 and negative complete response was set to 1. The variable recurrence described whether the cancer recurred locally (1) or not (0), while the variable metastasis described whether metastases were found after treatment (1) or not (0).

The endpoints were combined to a variable called Progression-Free Survival (PFS), to create a more balanced data set. PFS describes whether the patient gets better after treatment as suggested by Eirik Malinen [40], meaning that all variables, complete response, recurrence and metastasis, should be positive (equal to 0) for PFS to be positive. Positive PFS was denoted as 0 and negative PFS was denoted as 1 in this data set.

The endpoint data set included 93 patients, meaning that image sequences were not available for all of them. 11 of these patients had negative complete response, 4 had recurrence and 8 had metastasis of the initial 86 patients. 3 patients had all negative outcomes, 1 patient had negative complete response and recurrence, and 7 patients had only negative complete response. 5 patients had metastasis without other negative occurrences. This gave that 16 of the patients in the outcome data set had negative PFS.

All patients in data set 1 and data set 2 were included in the outcome data set. Matching the data sets with the outcomes gave that

- 5 of the 36 patients in data set 1 had negative PFS,
- 13 of the 81 patients in data set 2 had negative PFS.

These operations were performed with the Python package *Pandas* in a Jupyter Notebook with Python 3.
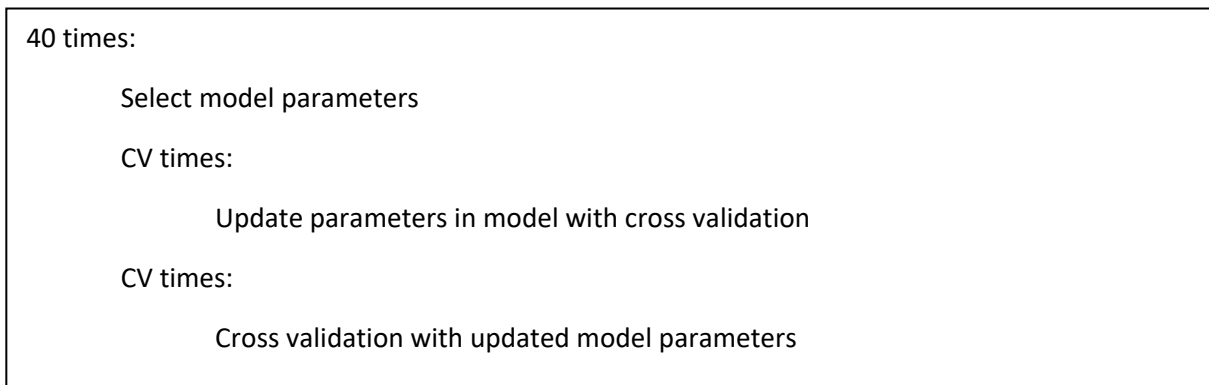
3.9. Classification

Both data sets were split into a training set and a test set, where only the training set was run through a model selection process in *biorad*. This was done using the function *train_test_split* from the Python package *scikit-learn* in a Jupyter Notebook and a test split of 0.25, meaning that a fourth of the data from each data set belonged to the test set. This gave

- 27 observations with 4 patients with negative PFS in the training data set of data set 1 and
- 60 observations with 10 patients with negative PFS in the training data set of data set 2.

The data were split with *train_test_split* parameter *random states* set to 5426 for data set 1 and 3603 for data set 2.

Model selection was further performed on the training sets of the data sets in the *biorad* package. Figure 3.4 describes the process of examining the performance of one classifier-feature selector combination.

```
40 times:

        Select model parameters

        CV times:

                Update parameters in model with cross validation

        CV times:

                Cross validation with updated model parameters
```

Figure 3.4: Simplified step of model selection process where the performance of one combination of a feature selector and classifier is calculated. This is repeated for every combination of feature selector and classifier. The model parameters that are mentioned in line 3 include feature selection parameters classifier parameters.

The model selection process combined chosen classifiers and feature selectors to models consisting of one classifier and one feature selector. The models were then updated to find optimal models for classification of the observations in given data and find how well these models performs. In this project this was used to get an indication of how well the extracted and chosen features could predict PFS for the patients in data sets 1 and 2.

CV was the number of folds the data set was split into during cross validation. A fold in a data set contains a fraction of the observations in the data set, where each fold in the data set contains an approximately equal number of observations. During cross validation, the model would be trained on CV-1 of the folds and applied on the remaining fold, the validation fold, to obtain a validation score. A training score would also be obtained by applying the trained model on the CV-1 folds. This was repeated CV times where all folds would have been used as validation fold once during the process.

After each experiment, some of the following properties was be stored:
- Number of features selected by the feature selector,
- Classification parameters,
- How many times each feature had been chosen during cross validation in line 6 in Figure 3.4,
- Mean validation score and mean training score.

During model selection in this project, the performance of the combinations of classifiers

- LDA
- QDA
- Decision tree
- Logistic regression
- Ridge regression
- KNN
- XGBoosting
- LightGBM
- SVC

and feature selection methods

- ReliefF
- Wilcoxon
- Chi Square
- Fisher Score
- Mutual Information

were explored. This gave 50 possible models as there were 50 combinations of classifiers and feature selection methods. AUC was the metric for updating the model parameters and evaluating the performance of the models.

biorad came with several default hyper spaces for choosing the optimal parameters, only some changes were made to suit the data sets used in this project. The maximum number of neighbours with the KNN classifier (*n_neighbours*) were set to the maximal number of patient observations in the folds used for training during cross validation, 17 for data set 1 and 49 for data set 2. As the data set were very unbalanced, the number of folds during cross validation was set to 3 for data set 1 and 10 for data set 2. This was so that every fold would at least include one patient with negative PFS and the program would be able to run

4. Results

4.1. Image discretization levels

It is recommended practice in *PyRadiomics* to use fixed bin widths for image discretization. As the fixed bin widths were computed over all images of one modality, most discretized images ended up having discretization levels that were different from the intended discretization level. The actual discretization levels were centered around the intended discretization level but had a range that spanned the two closest levels, as shown in Figure 4.1. For example, a sequence with the intended discretization level of 32, could have an actual discretization level in the range 16-64. Further, the range of actual discretization levels increased as its intended discretization level increased.
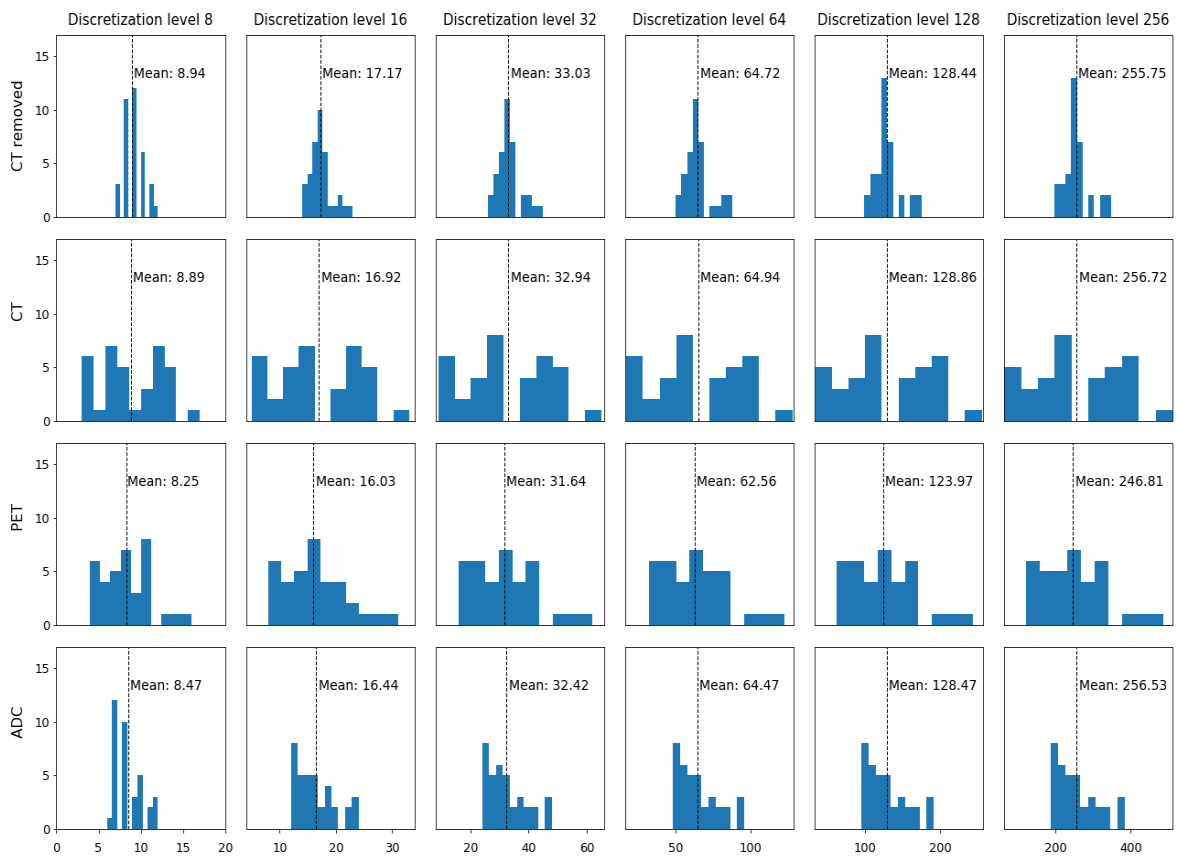
Figure 4.1: *Histogram of actual discretization levels after discretization with the fixed bin widths used in this project. The intended discretization level is denoted over the columns of the histograms. The histograms in each row belong to the same modality, where the modality is given to the left of the row. The mean actual discretization level is highlighted with a dotted line and denoted in each histogram. The actual discretization levels in this figure were found by subtracting the sequence maximum with the sequence minimum after discretization with the specified bin widths. The rows and columns share x and y-axis. The x-axis represents the actual discretization levels and the y-axis represents the number of sequences with the discretization levels in x.*

4.2. Feature extraction with different parameters

The impact of extracting features with different image settings were examined by calculating the Pearson correlation coefficient between the resulting features. The settings that were explored were dimension of extraction, tumour mask and discretization levels.

4.2.1. Features extracted from different dimensions

4.2.1.1. Correlation values between patient features
Pearson correlations between patient features calculated using 2D and 3D extraction are shown in the heatmap in Figure 4.2.

The figure shows that the shape features extracted at different dimensions had correlation values close to one for all patients. The correlation values between corresponding CT-features and ADC-features were quite high and similar, with correlation values just below 0.9. The correlation values between the PET-features had the most variation, with the lowest correlations close to 0.4. This is especially apparent between discretization levels 16, 32, 64 and 128 and shows that the PET-features were most sensitive to extraction dimension.

Figure 4.2. *Heatmap of Pearson correlation values between patient features found using 2D and 3D extraction. The correlation values are between corresponding rows of matrices consisting of patient features extracted with the same extraction settings except for extraction dimension. The x-axis of the figure denotes the patient ID and the y-axis denotes the feature matrices. The first feature matrix denoted on the y-axis contains shape features, while the other matrices contain first order and texture features extracted from the combinations of modality and specified discretization levels. Feature files named by "ct_nr", "ct_rem_nr", "pet_nr" and "adc_intact_nr", contain features extracted from CT-sequences, PET-sequences or ADC-maps after removal of slices not containing information within the tumour delineation. The tumour area has been discretized to discretization level nr. "ct_nr" contains tumour features extracted with the original tumour mask, while "ct_rem_nr" contains tumour features extracted with the modified tumour mask correcting for air and HU-values higher than 200. The colour bar to the right indicates the correlation values in the heatmap.*

4.2.1.2. Correlations between individual features

Which *features* that were affected by different extraction dimension is shown in Figure 4.3. These correlations were found by computing the Pearson correlations between the columns in the feature matrices.

First order features and CT-features extracted with the original masks were the least affected by extraction dimension. Compared to the other modalities, CT-features extracted with modified masks were more heavily affected by extraction dimension. Often, this applied for some features from every feature classes of every discretization level. Shape features are not included in Figure 4.3, as they were only extracted once and not for every combination of modality and discretization levels. The shape features extracted from different extraction dimensions were hardly affected with correlation values over 0.99.
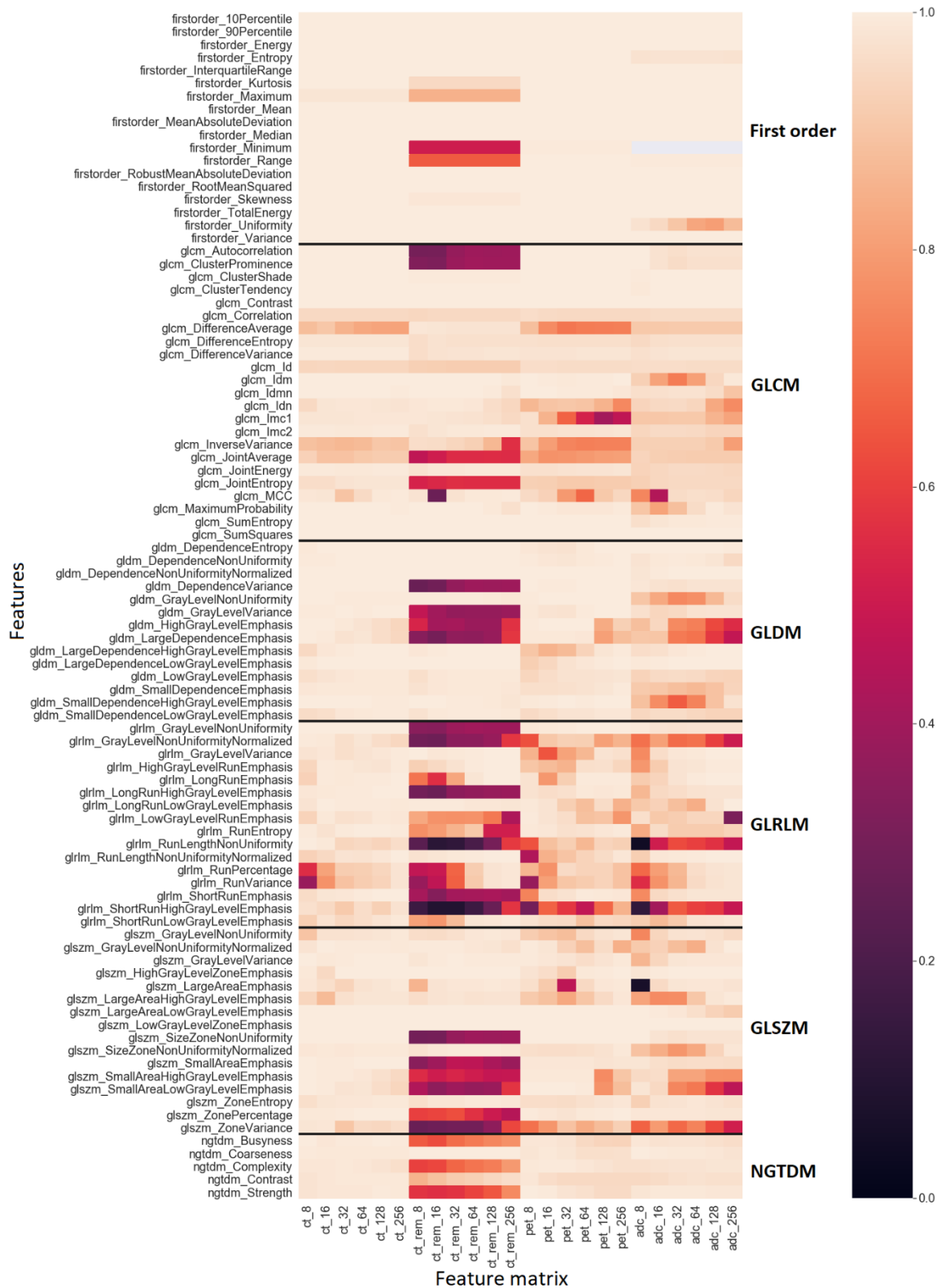
Figure 4.3: *Heatmap of Pearson correlation values between first order and texture features found using 2D and 3D extraction. The feature matrices are denoted in the x-axis, while the features are denoted in the y-axis. The feature matrices are named with the modality and the image discretization level while the features are denoted with feature class and feature name. Note that the light grey cells belonging to the first order ADC-feature Minimum does not contain any information for all discretization levels. This feature was removed due to zero*

variance, as the minimum value within the tumour delineation was zero for all ADC-maps. Feature classes are separated by a bold line and the colour bar to the right indicates the correlation values in the heatmap. The feature class names are denoted to the right of the heatmap and are as follows: First order, GLCM (Gray Level Co-occurrence Matrix), GLDM (Gray Level Dependence Matrix), GLRLM (Gray Level Run Length Matrix), GLSZM (Gray Level Size Zone Matrix) and NGTDM (Neighbouring Gray Tone Difference Matrix).

4.2.2. Correlation values between CT-features extracted with different tumour masks

4.2.2.1. Correlation values between patient features

Figure 4.4 shows the correlation values between patient features extracted from the CT-sequences using different tumour masks. The two heatmaps show the correlation values between features found using 2D or 3D extraction. The figure shows that all patient features were highly correlated, regardless of extraction dimension.
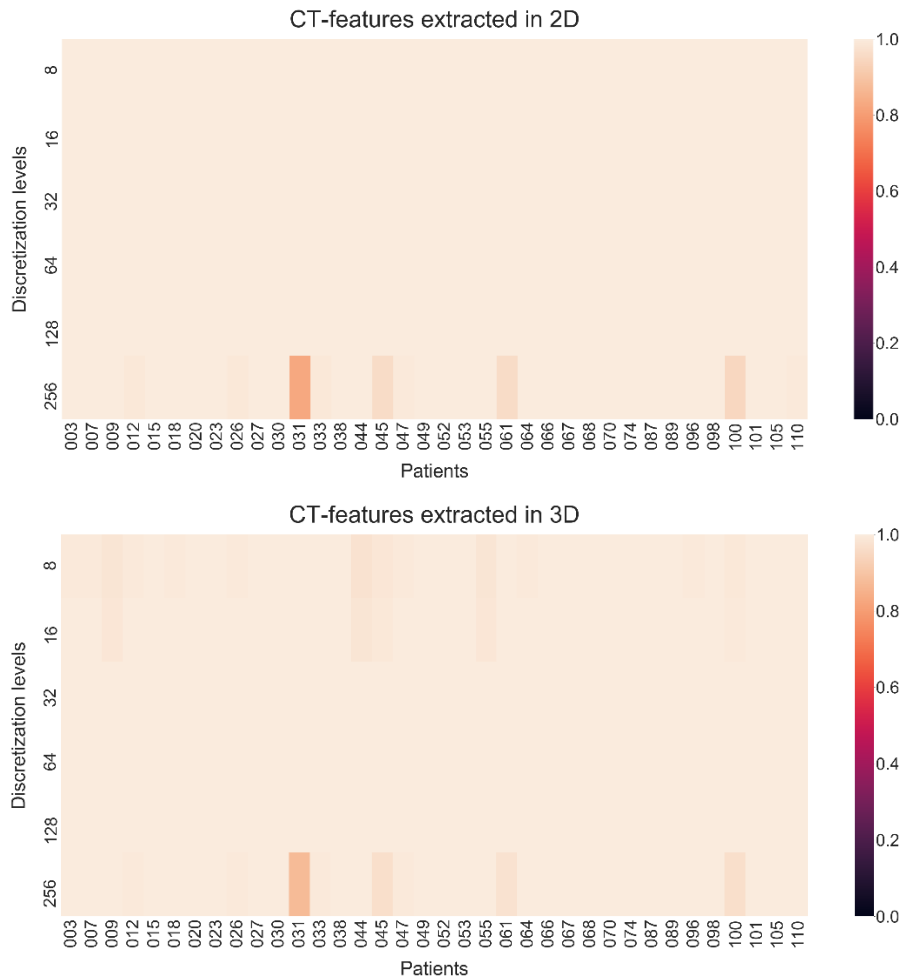
Figure 4.4: *Heatmaps with absolute correlations between patient features extracted from CT-sequences using the original tumour masks and the modified tumour masks. The upper heatmap shows the correlations of the features extracted in 2D while the lower heatmap shows the correlations of the features extracted in 3D. The patient ID is denoted along the x-axis, while*

the y-axis denotes the discretization level of the CT-sequences. The colour bar indicates the correlation values in the heatmaps.

4.2.2.2. Correlations between individual features
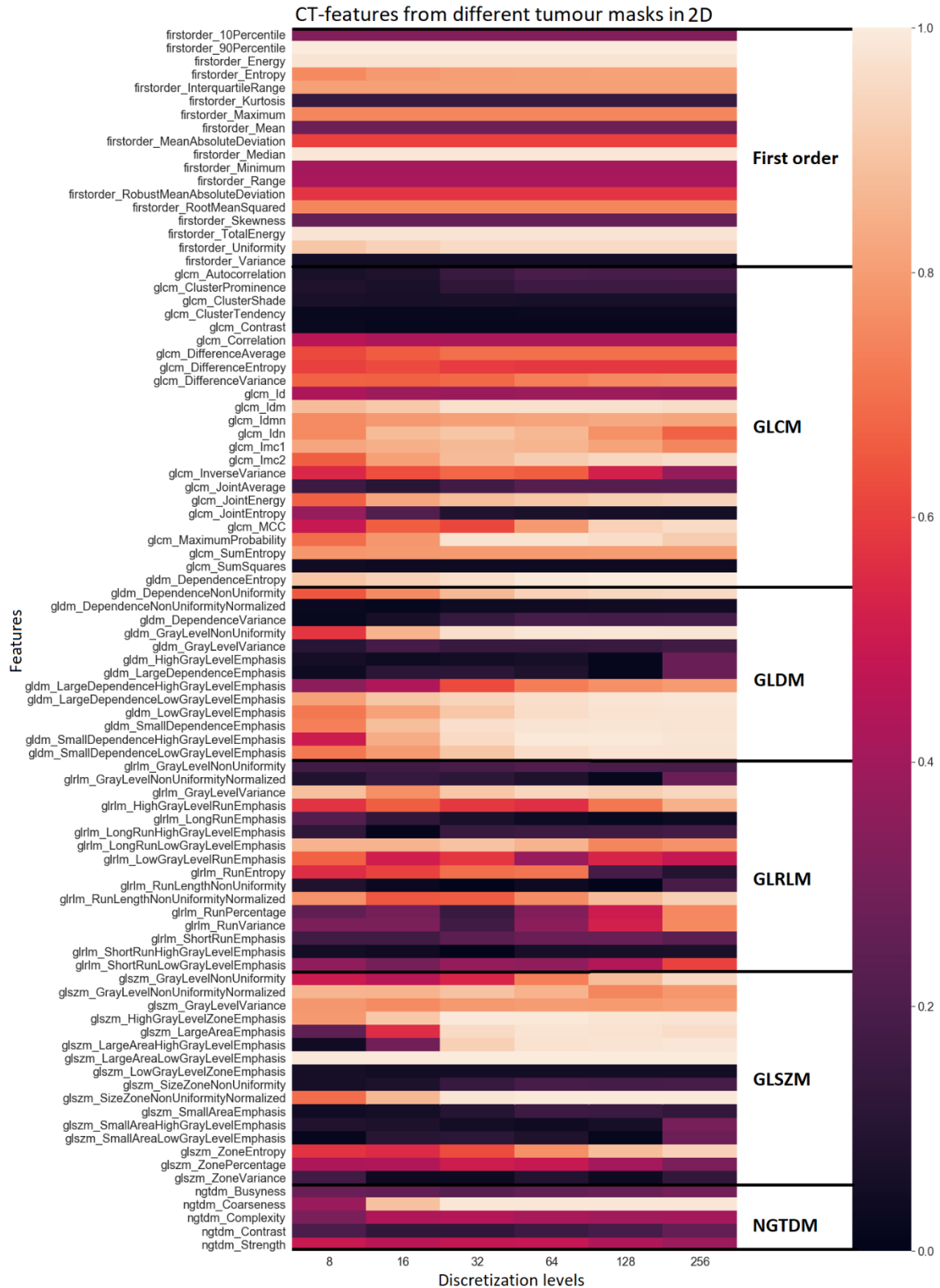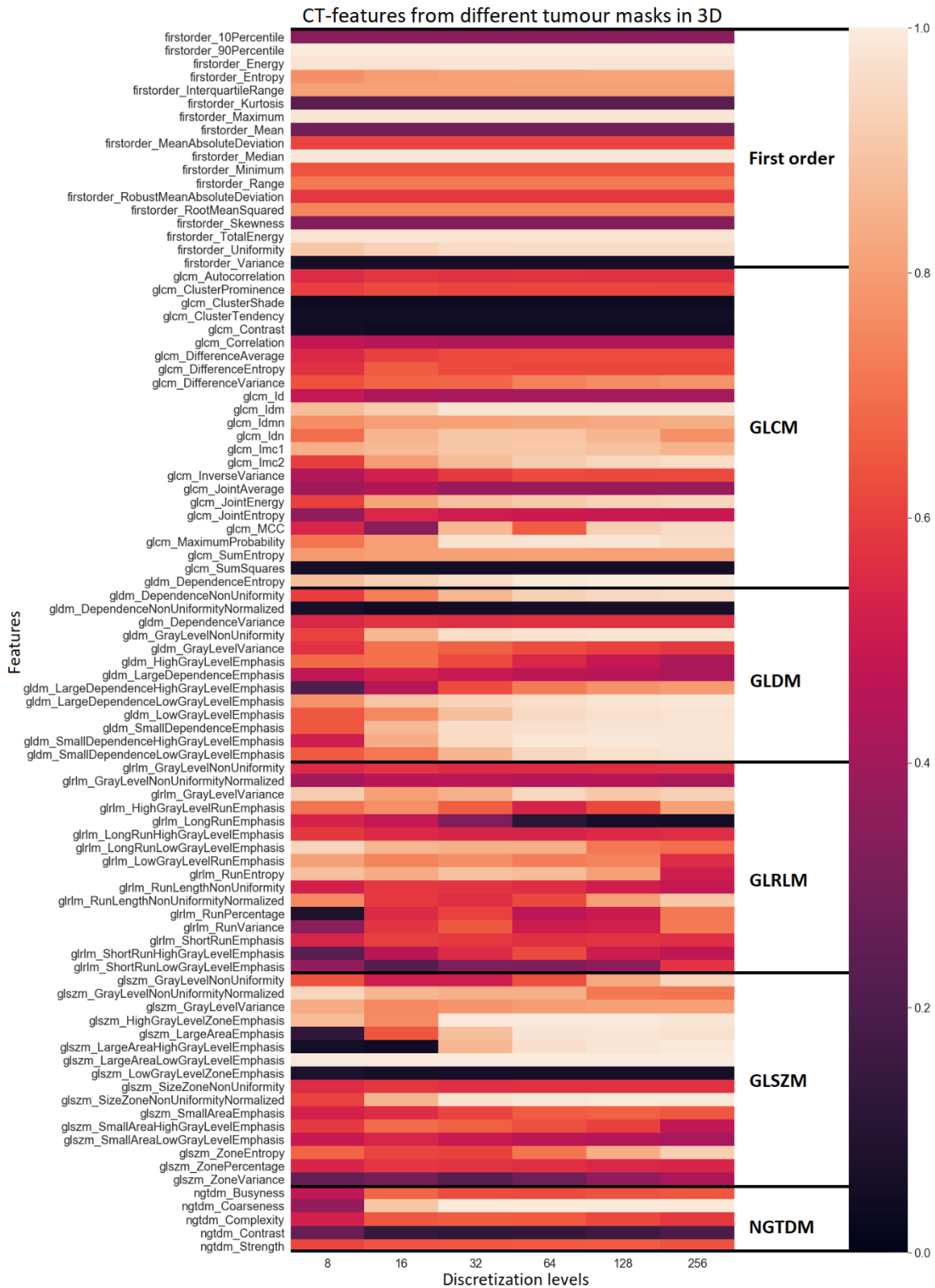
Figure 4.5: *Heatmaps with absolute Pearson correlation s between CT-features extracted using original and modified tumour masks. The upper and lower heatmap show the correlation values between features extracted in 2D and 3D. The discretization levels are denoted in the x-axis and the features are denoted with feature class (First order, GLCM, GLDM, GLRLM, GLSZM and NGTDM) and feature name in the y-axis. The feature classes are separated by a*

bold line and the feature class names are denoted to the right of the heatmap. The colour bars to the right indicate the correlation values.

Each cell in the heatmaps in Figure 4.5 represent the correlation between a CT-feature extracted with different tumour masks. A dark cell indicates low correlation and that the feature is greatly affected by being extracted with tumour masks that includes different ranges of HU-values. A light cell indicates high correlation and that the feature is not much affected or simply scaled linearly when using the different masks. The heatmaps shows that the features in general were less impacted by being found with 3D extraction than with 2D extraction.

4.2.3. Correlation values between features extracted at different discretization levels

Exploration of correlations between features extracted from the same modality but at different discretization levels showed that most were highly correlated. Generally, the lowest correlation values were between features at discretization levels 8 and 256, the lowest and largest discretization levels used in this project. The correlation values between features from images with similar discretization levels were also mostly high.

Most CT-features and features extracted from the first order and NGTDM feature classes were correlated with correlation values close to 1 for all discretization levels. The correlation values were not always this high for features belonging to the other feature classes.

The correlation values between the first order feature Uniformity extracted from the CT-sequences using modified tumour masks and PET-sequences and ADC-maps using original tumour masks at different discretization levels is shown in Figure 4.6. The heatmap shows that this feature was highly correlated for all discretization levels in the CT- and PET-sequences. The figure shows that the Uniformity feature extracted from the ADC-sequences at similar discretization levels was highly correlated, but not for discretization levels that were quite different. Uniformity was the first order feature that had the lowest correlation values between different discretization levels.
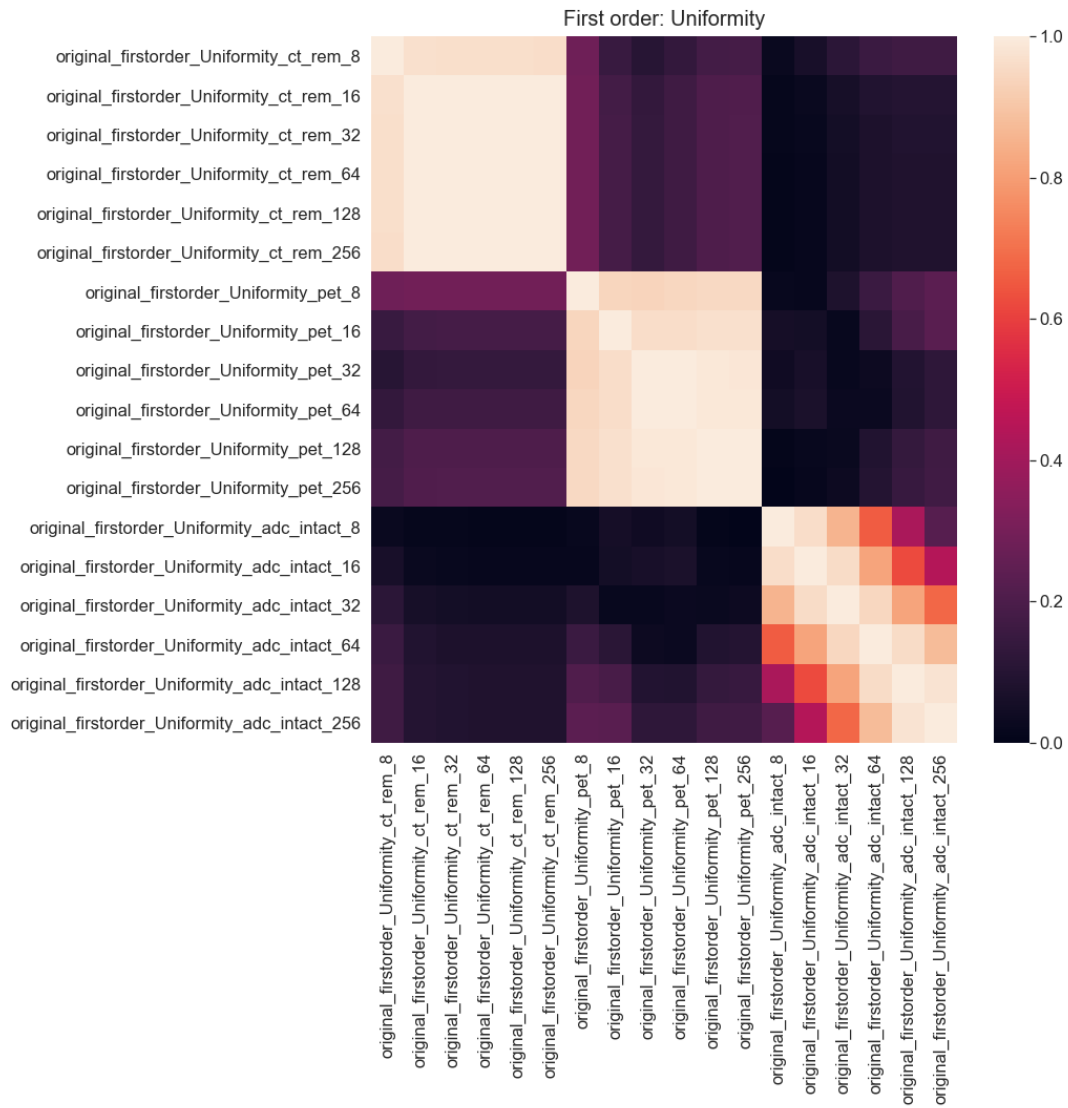
Figure 4.6: *Heatmaps of correlation values between CT-, PET- and ADC-features extracted in 2D with different discretization levels. The correlation values belong to the first order feature Uniformity. The first six rows and columns (respectively from top and left) belong to the CT-features extracted with the modified mask (ct_rem). The next six rows and columns belong to the PET-features and the last six rows and columns belong to the features extracted from the ADC-maps cropped to only contain slices that did not lack information within the tumour delineation (adc_intact). The discretization levels are sorted from top to bottom: 8, 16, 32, 64, 128 and 256 for each modality. The colour bar to the right indicate the correlation values between the features.*

4.3. Model selection

The Python package *biorad* was also used to evaluate which classification models that could work well to predict whether the patients in the data set had negative or positive PFS. Different combinations of classifiers and feature selection methods were examined. The data sets had beforehand been split into training and test sets, where the training sets were evaluated with *biorad*. Each training set contained 75 % of the observations in the initial data sets and were evaluated 40 times with the process described in Figure 3.4. in Methods for each combination of classifier and feature selector.

For each combination of classifier and selector

- 40 mean train AUCs,
- 40 measures of the variance of the train AUCs,
- 40 mean validation AUCs,
- 40 measures of variance of the validation AUCs.

Were computed using cross validation. The mean AUCs that are presented below are the mean of the 40 AUCs computed using the different models.

4.3.1. Data set 1

Data set 1 consisted of 36 patients and 289 features, consisting of shape features extracted from the original tumour mask and first order and texture features extracted from PET-sequences, CT-sequences and ADC-maps. The CT-features were extracted using the modified tumour mask, while the PET- and ADC-features were extracted using the original tumour mask. All features were extracted in 2D. The training set consisted of 27 patients, of which 4 had negative PFS. The test set included the remaining 9 patients of which 1 had negative PFS. The mean validation AUCs and training AUCs of model selection are shown in Figure 4.7.

The validation AUCs from the model selection on data set 1 were in the range $0.4586 - 0.7124$. The corresponding training AUCs were in the range $0.5 - 1$ as shown in Figure 4.7. The classifier LDA with Fisher score feature selector gave the highest mean validation AUC of 0.7124, while the classifier QDA with the ReliefF feature selector gave the lowest mean validation AUC of 0.4586.

For this data set, all models with feature selectors ReliefF and Mutual Information gave lower validation AUCs. Also, the classifiers Random Forest and XGBoost seemed not to perform as well as the other classifiers. The classifier that seemed to give the highest model performances was LDA using Fisher score feature selection, Chi Square feature selection and Wilcoxon feature selection.
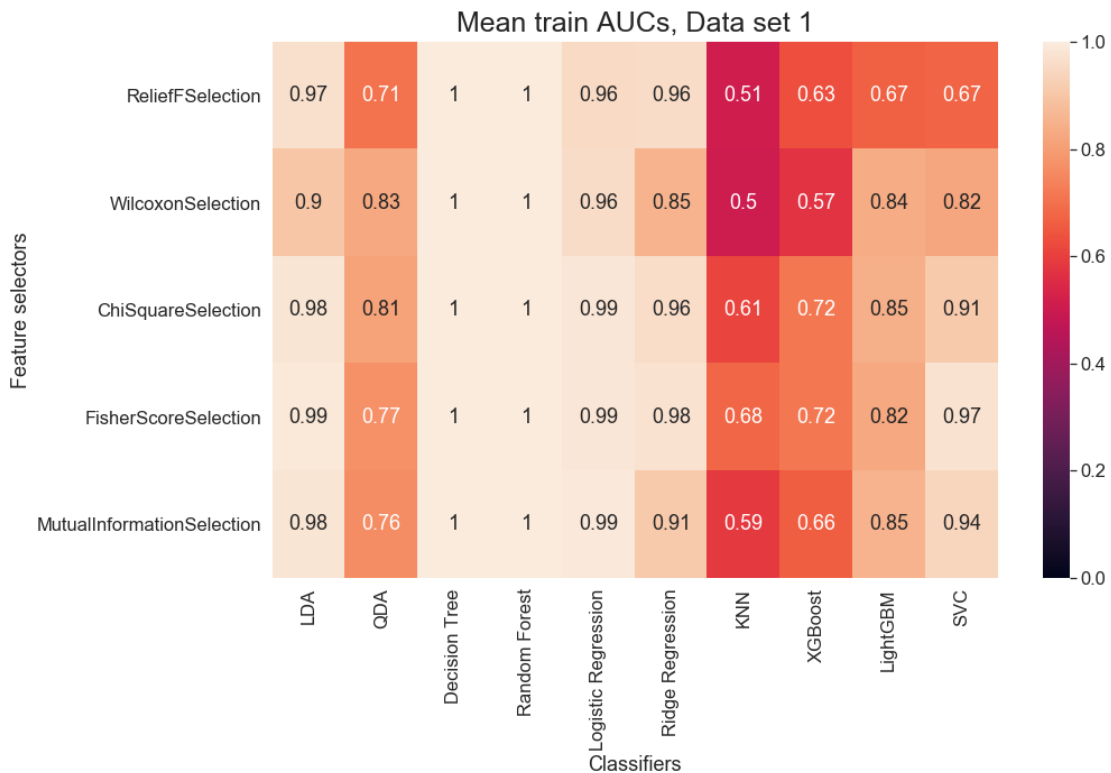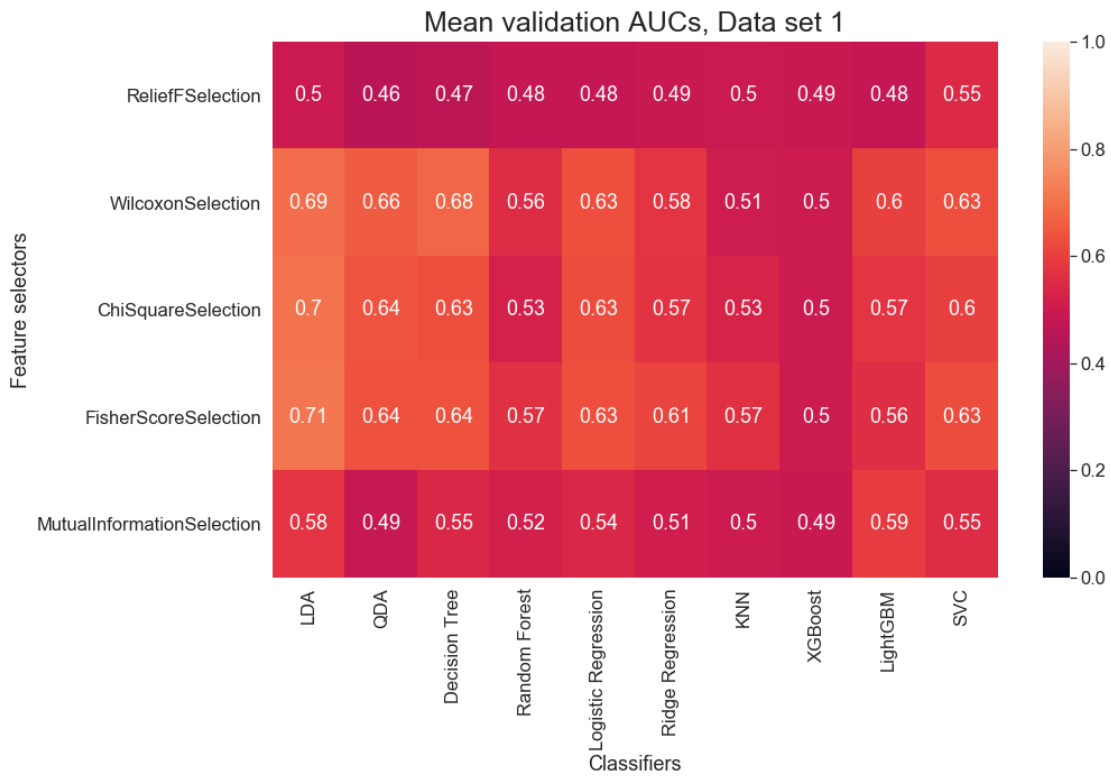
Figure 4.7: *Mean validation AUCs (upper heatmap) and mean training AUCs (lower heatmap) of 40 experiments for each combination of 10 classifiers and five selection methods on the training set of data set 1 with* biorad. *The classifiers are denoted along the x-axes and the feature selectors are denoted along the y-axes. The colour bars indicate the range of the AUCs. The AUCs are given with two decimals.*

The differences between train and validation AUCs for this data set were generally high, as seen in Figure 4.8. Only the classifier KNN gave approximately the same validation and train AUCs. The variances of the validation and train AUCs were also computed during model selection with *biorad*. The variance of the train and validation AUCs were computed for each experiment and are given in Figure 4.9.
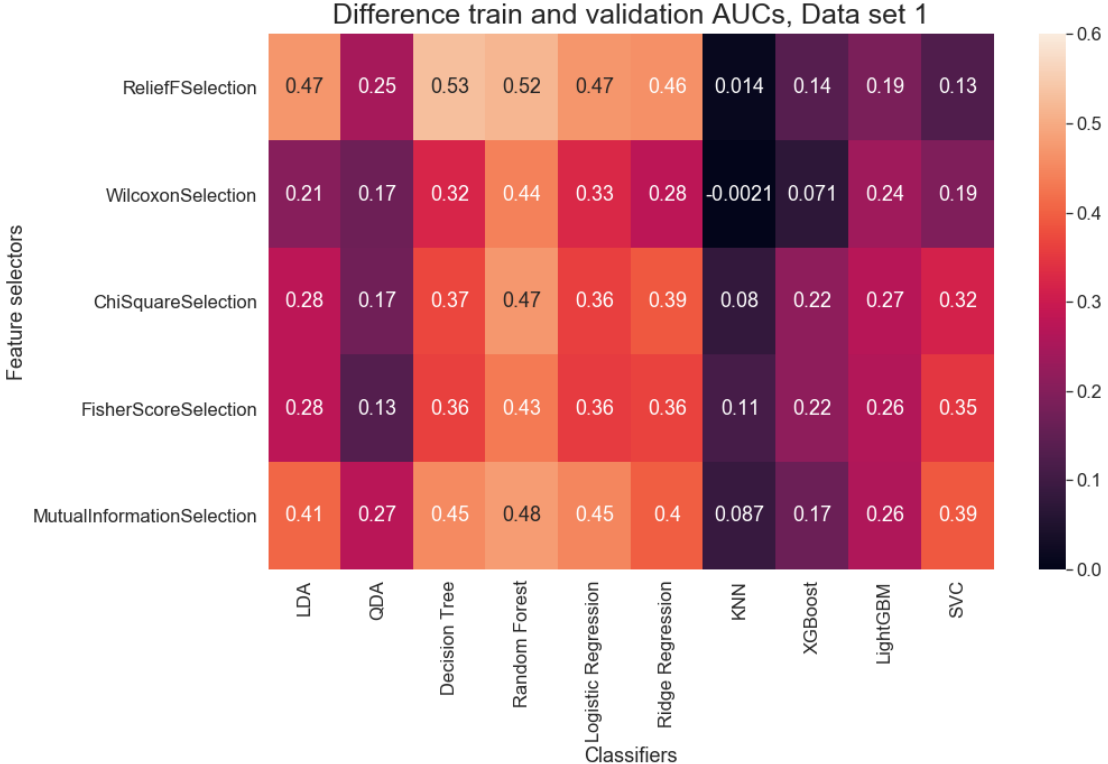
Figure 4.8: *Difference between training and validation AUCs in Figure 4.7 for 10 classifiers and five feature selection methods on the training set of data set 1. The classifiers are denoted along the x-axis and the feature selectors are denoted along the y-axis. The colour bar to the right shows the range of differences between training and validation AUCs from this experiment.*

Figure 4.9: *Mean variance of validation AUCs (upper heatmap) and mean variance of train AUCs (lower heatmap) of model selection on the training set of data set 1. The classifiers are denoted along the x-axes and the feature selectors are denoted along the y-axes. The colour bar to the right shows the range of the variances of the AUC values from this experiment.*

4.3.2. Data set 2
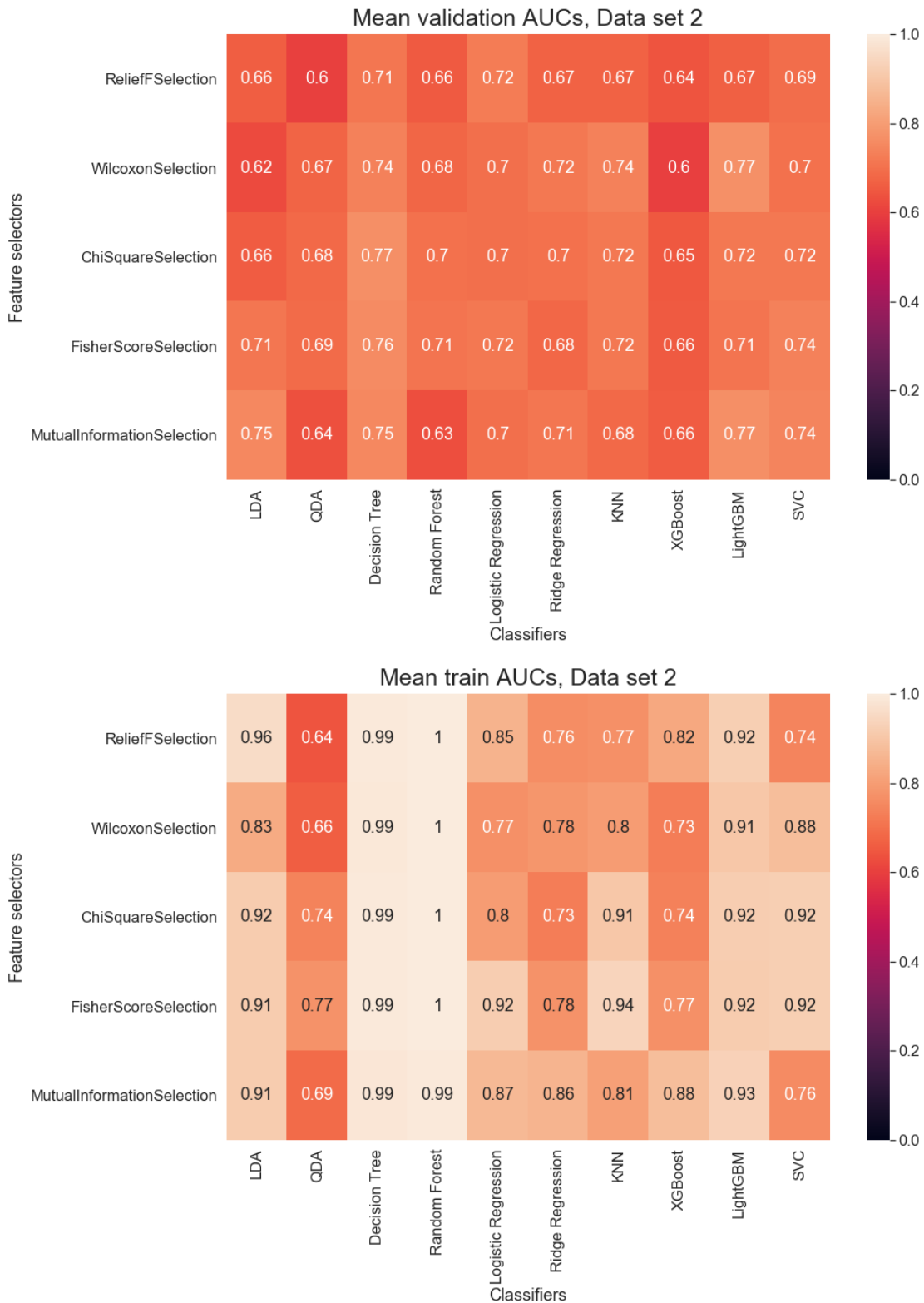
Figure 4.10: *Mean validation AUCs (upper heatmap) and mean training AUCs (lower heatmap) from 40 experiments performed for the combinations of 10 classifiers and five feature selection methods on the training set of the data set 2. The classifiers are denoted in the x-axes,*
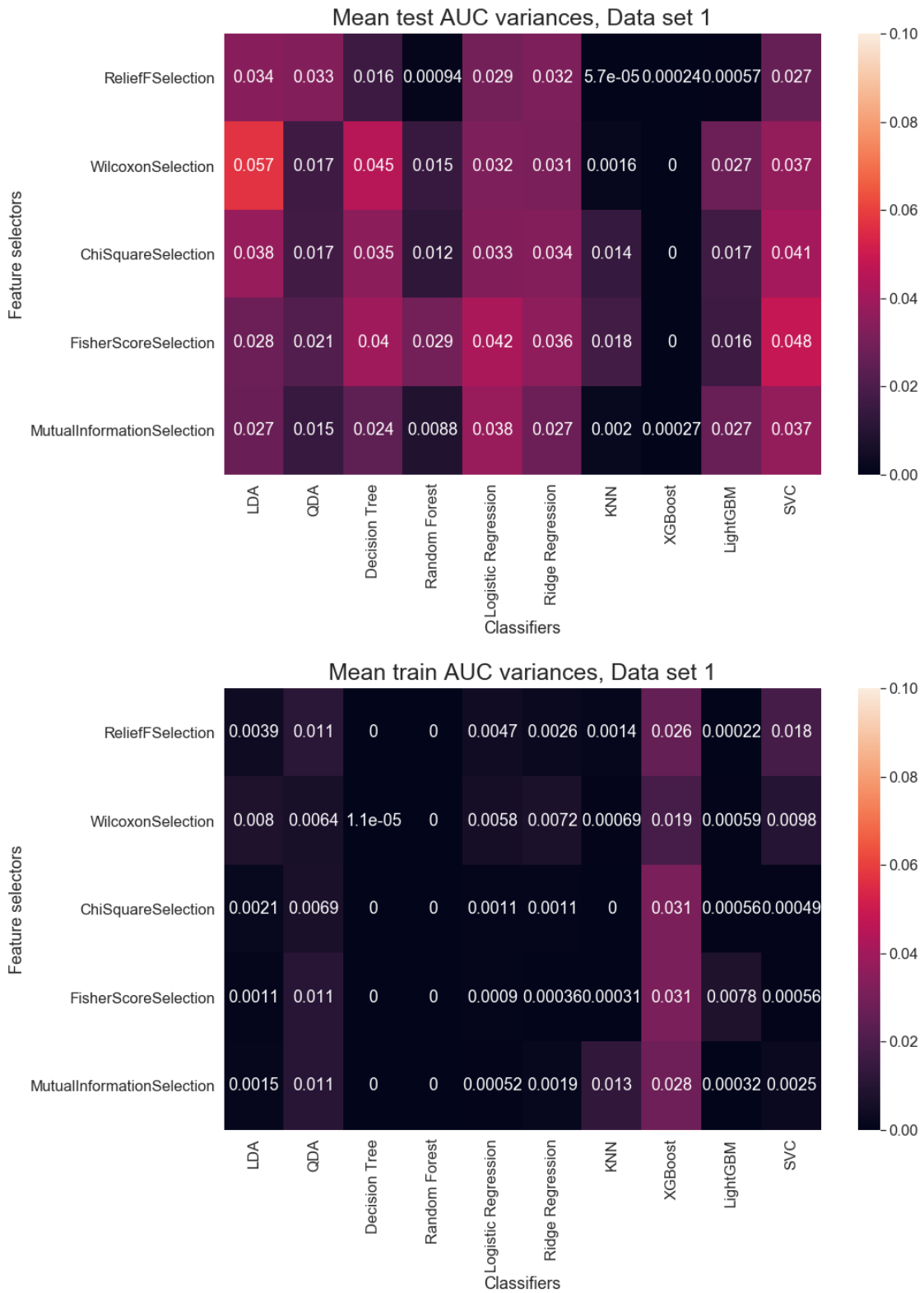
while the feature selectors are denoted in the y-*axes. The colour bars indicate the range of the AUCs. The AUCs are shown with two decimals.*

Figure 4.10 shows the mean validation and mean train AUCs from model selection of the training set of data set 2. The training set consisted of 60 patients with 198 shape, first order and texture features extracted from the tumour regions in PET- and CT-sequences. 10 of these patients had negative PFS.

The mean validation AUCs for the training set of data set 2 were in the range 0.5983 – 0.7693 while the training AUCs were in the range 0.64 – 1. The classifier-selector combination of LightGBM with the Wilcoxon feature selector gave the highest mean validation AUC of 0.7693. The two classifier-selector combinations Decision tree with Chi square feature selection and LightGBM with Mutual information feature selection gave the AUCs of 0.7688 and 0.7680, that were quite close to the highest mean AUC.

The model-selector combination of XGBoost with Wilcoxon feature selection gave the lowest mean validation AUC of 0.5983.
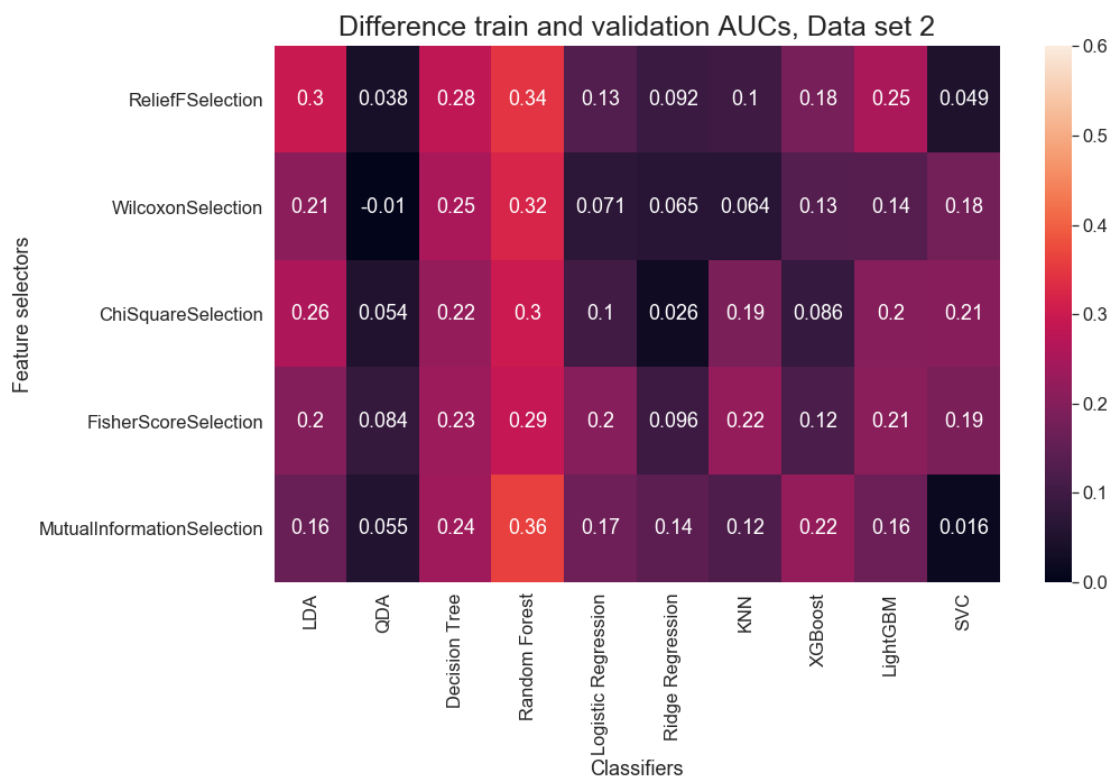
Figure 4.11: *Differences between training and validation AUCs in Figure 4.10 for the combinations of 10 classifiers and five feature selection methods performed on data set 2. The classifiers are denoted along the* x-*axis and the feature selectors are denoted along the* y-*axis. The colour bar indicates the range of the differences between the train and validation AUC values for this experiment.*

Figure 4.11 shows that the classifier-selector combinations

- including the classifier QDA,
- SVC with selectors ReliefF and Mutual Information,
- Ridge Regression with selector ChiSquare,

gave the train AUC and validation AUC values that were most similar, while the classifiers LDA, Random Forest, Decision Tree and LightGBM in general had the largest differences between training and validation AUCs. For this data set the variance of validation AUCs was visibly larger than the train AUCs as shown in Figure 4.12.
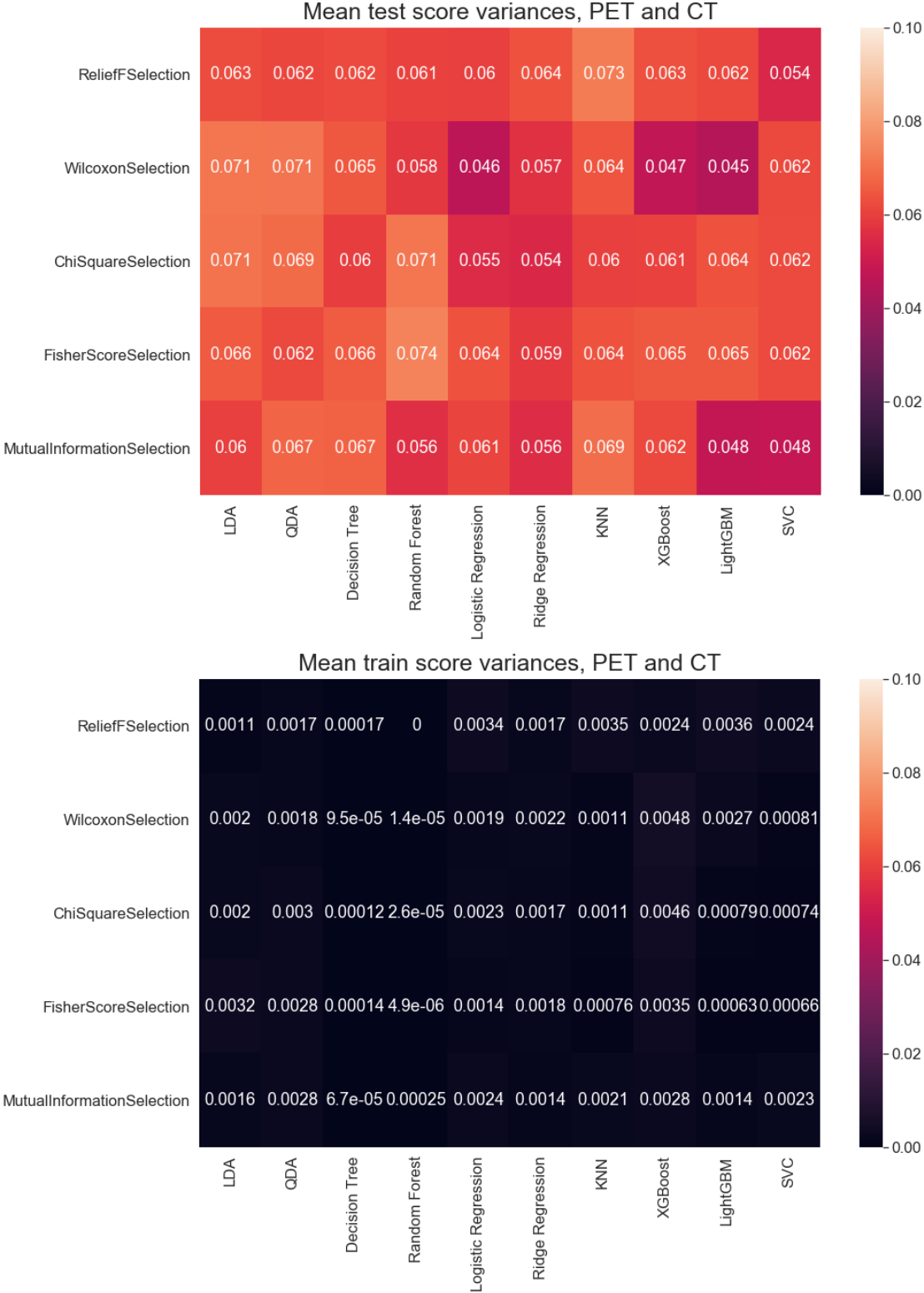
Figure 4.12: *Mean of variance of validation AUCs (upper heatmap) and train AUCs (lower heatmap) from the model selection on the training set of data set 2. The classifiers are denoted*

along the x-axes and the feature selectors are denoted along the y-axes of the heatmaps. The colour bars show the values of the variances in the heatmaps.

4.3.3. Most selected features

4.3.3.1. Data set 1
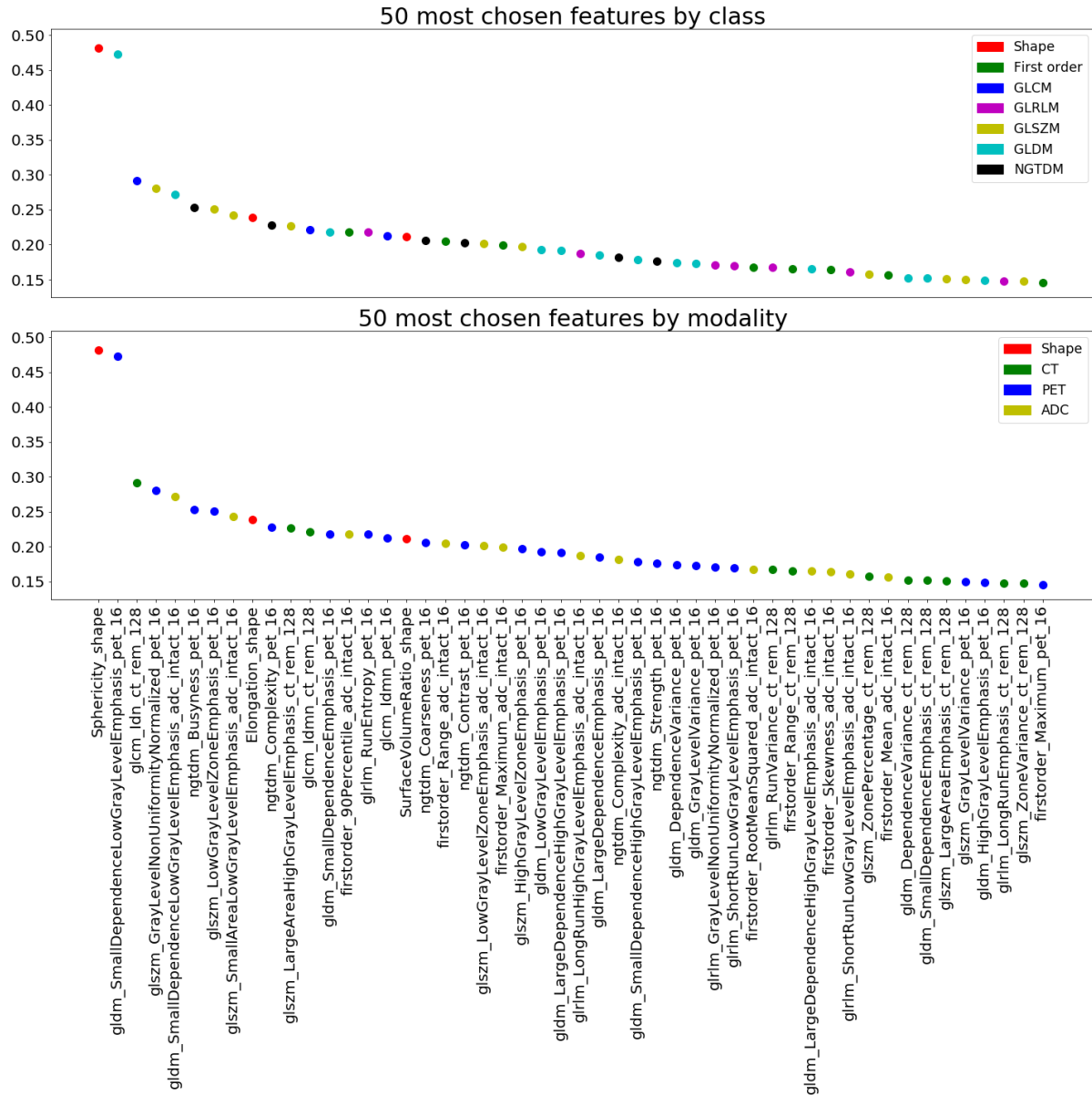
***Figure 4.13:** Plots of how many times the 50 most chosen features in data set 1 were selected. The features are sorted after how many times they were chosen in descending order along the x-axis, while the y-axis represent the percentage of times they were chosen. The upper plot distinguishes between the feature classes, while the lower plot distinguishes between modality. The features are denoted along the x-axis with feature class, feature name, modality and discretization level.*

A feature could at most be chosen 6000 times during model selection on the training set of data set 1. No features were chosen more than half of the time. For all selectors except the Wilcoxon feature selection method, 0-3 features were chosen in more than half of the experiments involving the specific feature selectors. The features that were chosen the most were not necessarily the same for all selection methods. The Wilcoxon selection method chose approximately 17 features in more than half of the experiments. Independent of selector method, two features were chosen most, the shape feature Sphericity and the GLDM feature Small Dependence Gray Level Emphasis as shown in Figure 4.13.

4.3.3.2. Data set 2
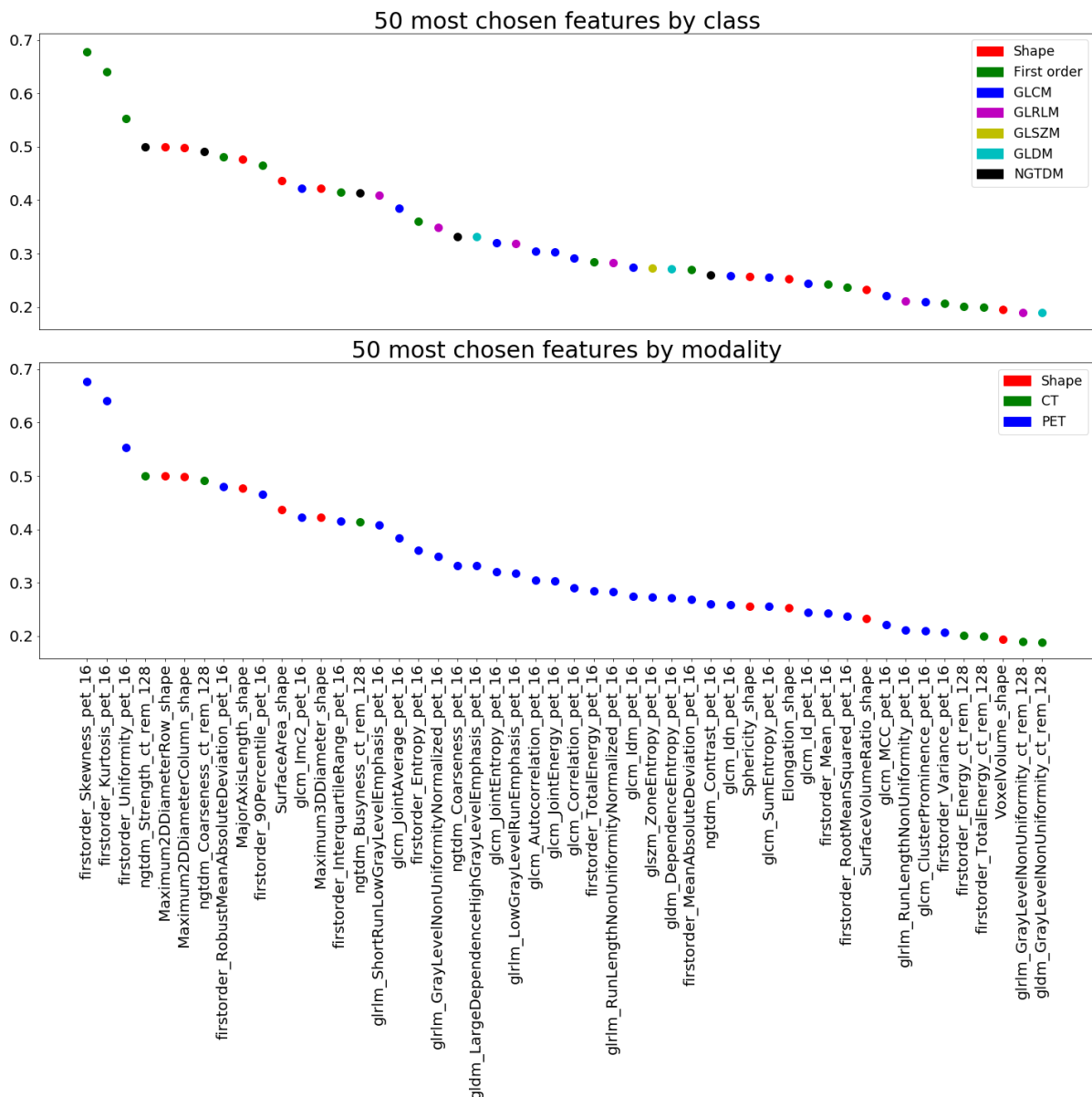
Figure 4.14: *Plots of how many times the features in data set 2 were selected. The features are sorted after how many times they were chosen in descending order along the* x*-axis, while the* y*-axis represent the percentage of times they were chosen. The upper plot distinguishes between the feature classes, while the lower plot distinguishes between modality. The features are denoted along the* x*-axis with feature class and name.*

57

During model experiments with the training set of data set 2, a feature could at most be chosen 20 000 times. Only three features were chosen more than half of the time; These were the first order PET-features Kurtosis, Skewness and Uniformity. PET-features were in general found among many of the most chosen features. The feature selectors each chose 25 or less features in more than half of the experiments. The features chosen the most for different feature selectors were not necessarily the same. For the different classifiers, 15 or less features were chosen in more than half of the experiments.

Shape was a separate category in the plots in Figures 4.13 and 4.14 as the shape features were independent of image modality.

5. Discussion

The purpose of this project was to prepare image sequences of CT-, PET- and MR-images of patients with anal cancer and predict treatment outcome of the patients.

This section starts by describing the image sequences, how they were processed and issues that were encountered during these processes. Further, the effects of extracting features using different image settings and the preliminary results from model selection with the software *biorad* will be discussed.

5.1. Image data set

In this section, the initial image acquisition, registration and processing used on the image sequences in this project will be evaluated.

5.1.1. Earlier acquisition and processing

The image sequences must be acquired and processed similarly in order to make a standardized data set. That patient images are as standardized as possible is difficult to obtain, as patients are examined separately and processed in several steps, where these steps often are conducted by different people as mentioned in [3]. This means that the images might have been acquired by different imaging parameters, or that different oncologists might have delineated the tumours differently.

If the registration was successful is also a point of uncertainty. The images had been registered beforehand with help from the program MICE as described in [3]. For some sequences, it could seem that the tumour delineation did not entirely include the entire areas with high metabolic activity in the PET-sequences, or areas with prominent diffusion coefficients in the ADC-sequences. The ADC-slice in Figure 5.5 is an example of a slice where it is uncertain whether the tumour delineation properly delineates the tumour. It seems like the delineation should be located more to the right where the slice has an area of brighter intensity values. The purpose of registration was to ensure that features were extracted from the tumour area in all sequences belonging to a patient, regardless of modality. A poor registration would thus cause the extracted features to not represent the tumour, which would not be ideal.

Acquisition, tumour delineation and image registration are all steps that had been done before the start of this project. The sequences were acquired in a specific stage of treatment and can therefore not be acquired again. It is more likely that tumour delineation and image registration could be redone to possibly ensure more standardized image sequences.

5.1.2. Choice of patients in data sets

The image sequences in this project had to be examined, excluded if necessary and processed further before features could be extracted.

Data set 1 included at least one CT-sequence with artefacts caused by metal in the patient body. This sequence belonged to patient 70. That the sequence contained artefacts was found manually, as the histogram of the area within the tumour delineation of this sequence did not visually stand out from the histograms of the other tumours in data set 1. This is seen in Figure 5.5. The artefacts were visible as "rays" in the sequences and were in this case also present within the tumour delineation. This could certainly affect extracted texture features and should possibly not have been included in data set 1. Unlike data set 1, data set 2 did not include patients with artefacts as these were found and removed. It was justified to remove patients from data set 2 and not from data set 1 as data set 1 contained fewer patients than data set 2.

Data set 2 also excluded a patient with a tumour delineation that did not correspond well with the CT- and PET-sequences. This tumour included bone in the CT-sequences and did not contain high SUV-values in the PET-sequences. This is an example of an observation that should not be included and was the only patient excluded due to issues with image registration or delineation.

5.2. Image examination and processing

5.2.1 CT-sequences

5.2.1.1. HU-distribution
Assuming all CT-sequences were taken by the same parameters, HU-values should be absolute and representative across patients [14]. Except from the histogram of patient 38, the histograms of the patients in Figure 5.1 seemed to be quite comparable. The approximately same HU-ranges were also found in the tumour intensity histograms from different patients. Therefore, it could be justified that no image operations were performed on the HU-values of the CT-sequences.

The histogram of the area within the tumour delineation of patient 38 was slightly shifted towards higher intensity values than the histogram of the tumours of the other patients as seen in Figure 5.1. This observation was included in both data set 1 and data set 2 but should be included with caution in possible later projects as its HU-distribution differs from the others. This resulted in that the modified tumour mask correcting for air for the CT-sequence of patient 38, did not fully remove the areas that appeared darker.

It was not investigated whether features extracted from the CT-sequences of patients with artefacts or slightly shifted histogram stood out from features extracted from the other patients. The PET- and CT-sequences from the patients that were only included in data set 2 were not

examined as properly as data set 1 but had overall similar HU-distributions as the sequences belonging to data set 1.
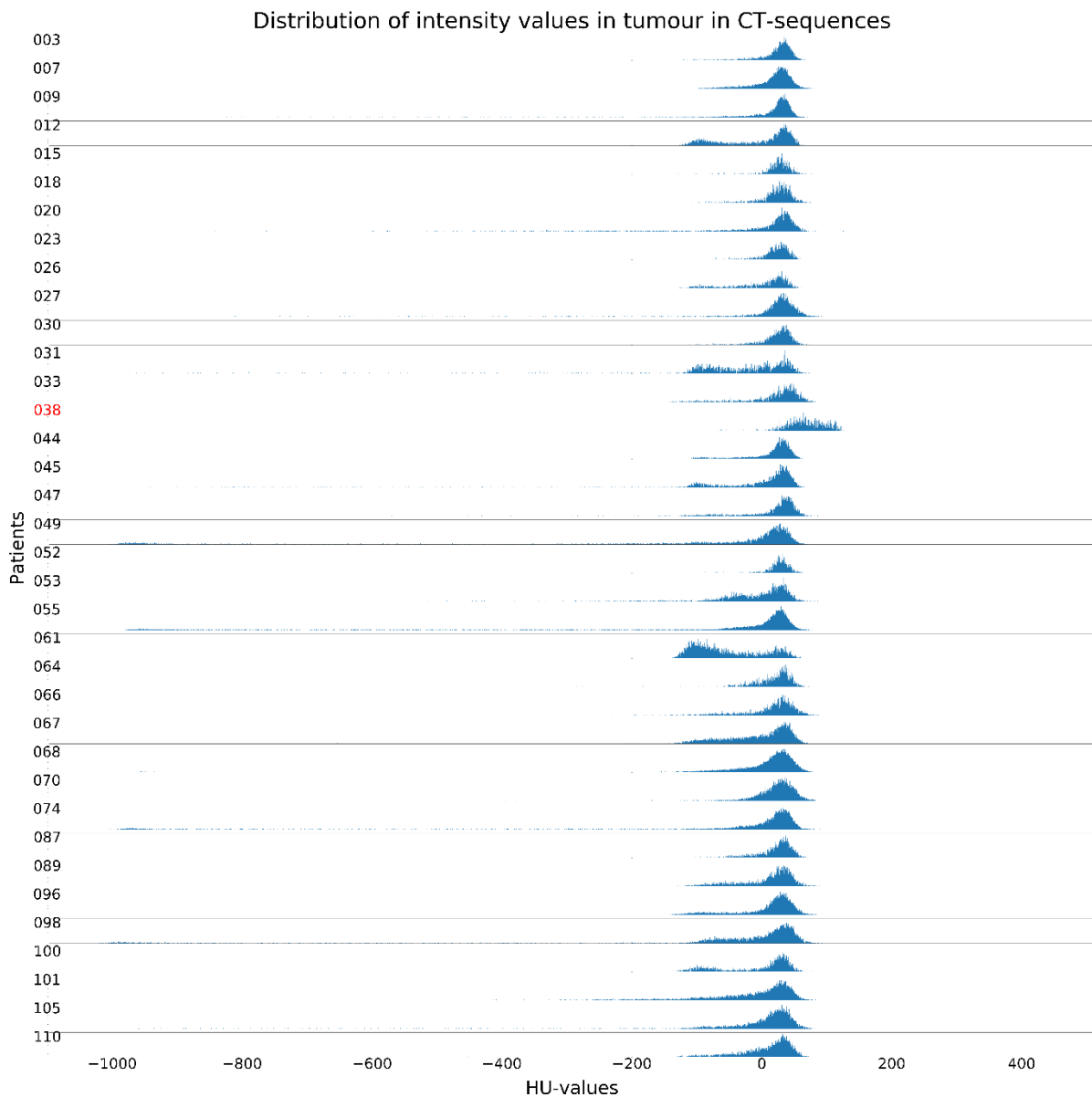
Figure 5.1: *Histogram of intensity value distribution in the tumours in the CT-sequences in data set 1. The label of the patient with the shifted histogram is marked in red.*

5.2.1.2. Removal of air and high intensity values in CT-images

The number of dilations, size and dimensions of the structuring elements and dilation dimension was considered when creating masks excluding HU-values in the range [-150, 200]. Dilation in three dimensions seemed useful for removing air effects in neighbouring slices for some patients but also removed large areas in neighbouring slices for patients with large air packages, thereby removing information that by first glance did not seem necessary to remove. It was therefore chosen to dilate two times and in two dimensions within the slices. It seemed also reasonable to choose a lower number of dilations as the lower chosen threshold for considering

a HU-value to represent air was quite high (-150) and included lighter grey areas around the boarders of the darkest areas within the tumour delineation.

In addition to removing neighbouring voxels that might be partially affected by nearby voxels with HU-values outside the range of [-150, 200], dilation of the areas also seemed useful by making the borders of the masks less rough in some sequences. This is illustrated in the third column of Figure 3.3 in Methods. The final choice of structuring element, number of dilations, dimension of structuring element and dimension of dilation was influenced by sequences with air. These choices were made based on data set 1, that only contained one image sequence with HU-values higher than 200 within the tumour delineation. The same choices for dilation were later performed on data set 2 without much exploration and might not be the best choice for creating tumour masks for the entire data set of CT-sequences.

If removal of air and high intensity values should be performed again, it might be useful to create masks manually for each patient. Another option is to choose the lower threshold based on HU-values instead of by visual exploration. Air have HU-values near -1000 [14]. This is much lower than the lower threshold of -150 that was set in this project. The value of -150 was chosen after inspection of the HU-values in the sequences. They represented most of the darker areas within the tumour delineation in the CT-sequences. PET-features were among the most chosen features from both model selections on the training sets of data set 1 and 2. This was especially the case data set 2. It is a concern that the choice of lower threshold removed information from the CT-sequences that could have been useful during model selection.

The functions for structuring elements and dilation were written in Python instead of using packages that already exists in Python. This was done due to compatibility difficulties between already installed Python packages. The final functions might not work as well and fast as already existing functions. A final suggestion when making modified tumour masks for CT-sequences would be to find all voxels instead of just the voxels within the tumour with HU-values outside of the desired intensity range (in this project [-150, 200]). Then, voxels, located within the tumour and that might be affected by nearby voxels outside the tumour representing air, could also be excluded from the original mask.

5.2.2. PET-images

The intensity values in the PET-sequences had beforehand been converted to SUV-values during the registration process described in [3]. The sequences from different patients should therefore be comparable, even though there is a concern that SUV-values might not entirely correct for uptake variation as this might be affected by more factors then dose and patient weight [3]. Thus, the PET-sequences were not processed further in this project. The histograms of the SUV-values within the tumour are shown in Figure 5.2.
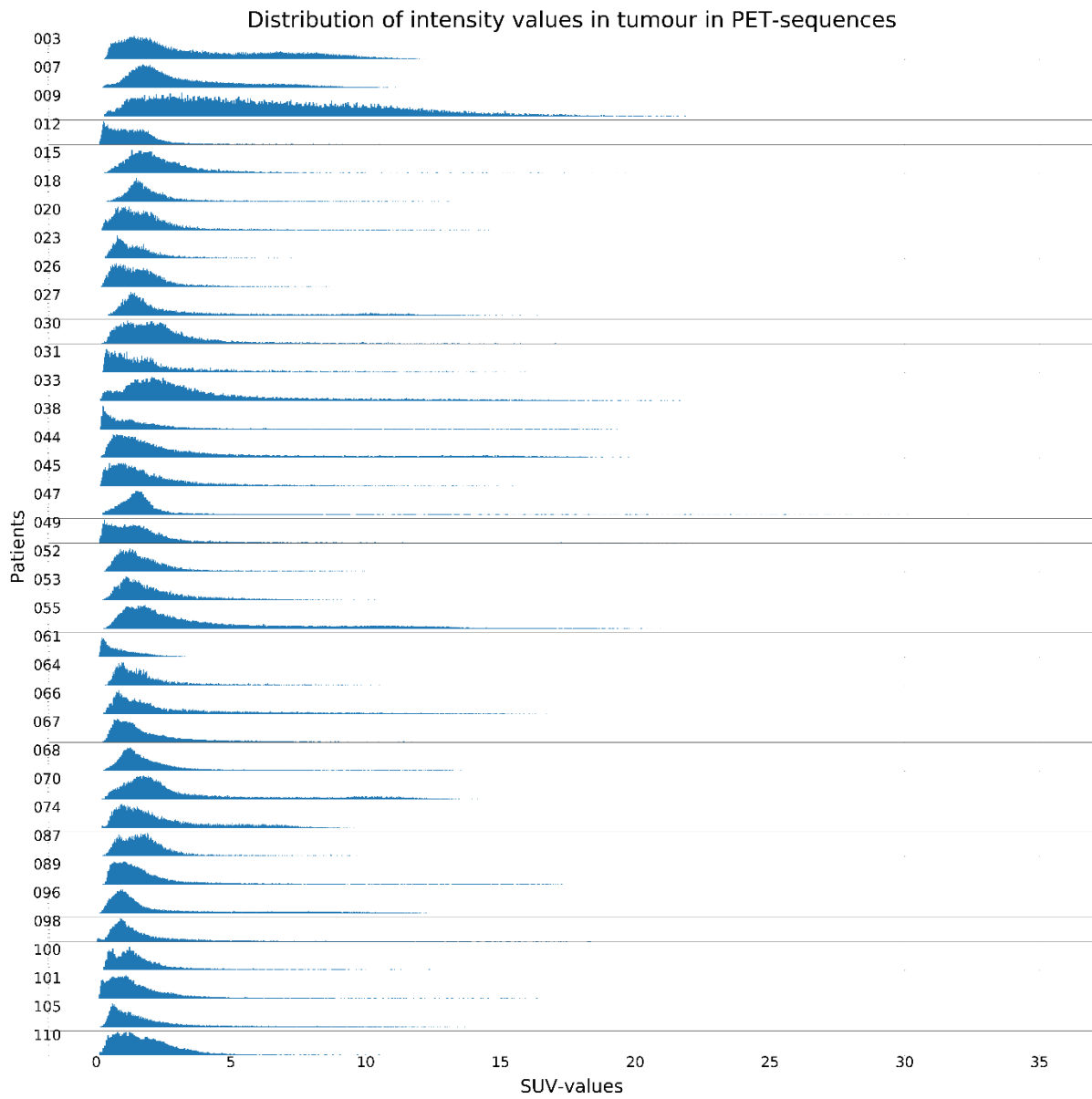
Figure 5.2: *Distribution of SUV-values within the tumour delineation in PET-sequences.*

5.2.3. MR-images
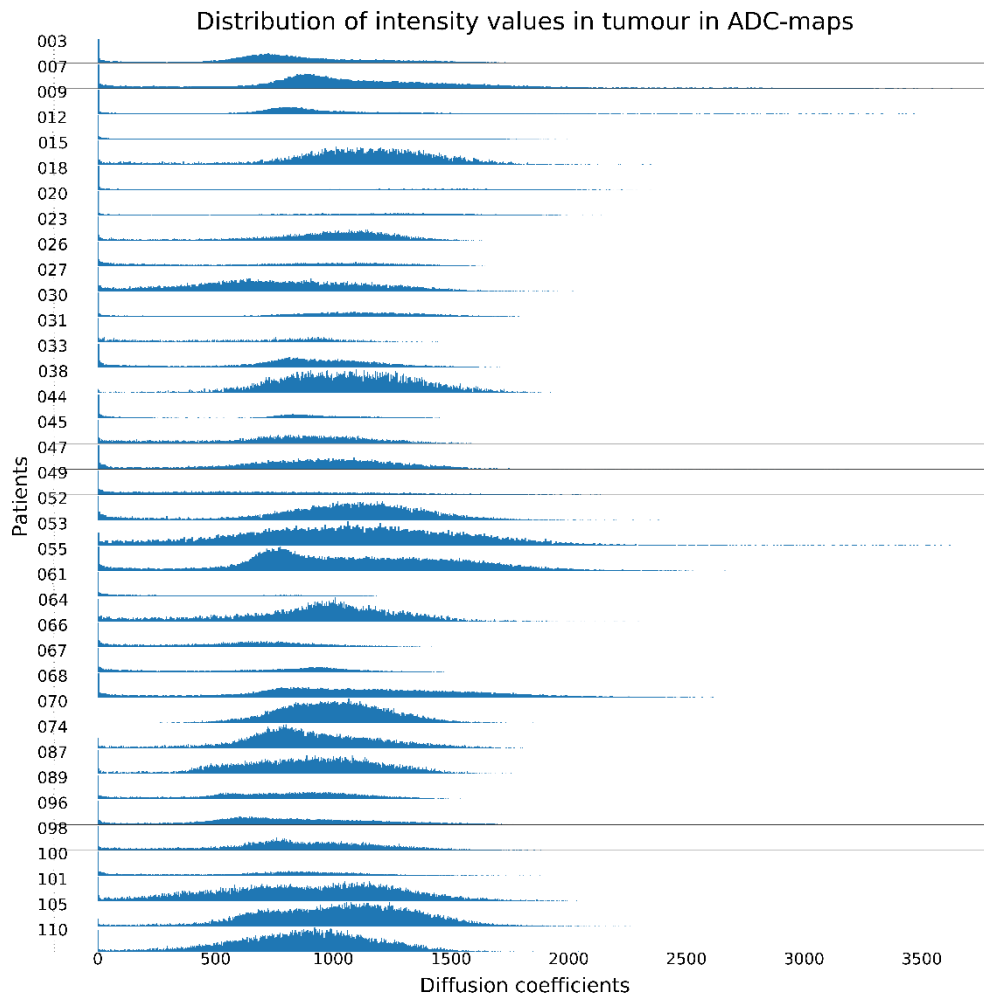
Figure 5.3: *Distribution of diffusion weighted coefficients within tumour delineation in ADC-maps.*

No operations were performed on the intensity values in the ADC-maps. The ADC-maps were cropped so that tumour slices that partially or entirely lacked information pertaining to the tumour were not included. This would also have been necessary for the diffusion weighted MR-sequences or the T2-weighted MR-sequences if they had been included in this project.

The main reason such operations were not necessary for the ADC-maps, was that diffusion coefficients should be absolute. Also, all features were extracted from the areas defined by the tumour delineation. If, however, operations depending on all intensity values in the sequences, like normalization, should be performed, only voxels with information pertaining to the body should be used. For some MR-images, this would mean cropping the edges of the images, excluding all slices including discontinuities, or "cuts", or creating masks excluding edges or discontinuities without information.

If all slices with discontinuities had been removed in this project, not only the first two slices from the left, but the four first slices from the left would have been removed in Figure 3.2 in Methods, reducing the number of sequence slices even more. A mask excluding edges and discontinuities might be a better solution for slices where the discontinuities are small. The slice in Figure 5.4 is example of an MR-slice with edges that does not contain any information.

Figure 5.4: *T2-weighted MR-slice with edges that do not contain information. This slice belongs to patient 110.*

5.2.3.1. Removal of discontinuities within the tumour region in ADC-maps

Later inspection of the DWI-sequences showed that even though the area within the tumour in an ADC-map no longer seemed to be affected by discontinuities after cropping, the edges of these cuts could still occur within the delineation in the corresponding DWI-sequence. These "edges" can be described as bands that are along the "cuts" in the images with darker intensity values than the rest of the slice. They should not be confused with the edges described in the previous section. They were not as evident in the ADC-sequences as in the DWI-sequences. This is shown in Figure 5.5. As the ADC-maps are derived from the DWI-sequences, these slices should therefore ideally not be included in the ADC-maps. These slices were not found until later in the project and were thus not removed. This applied to patients with ID 9, 47, 49 and 67.
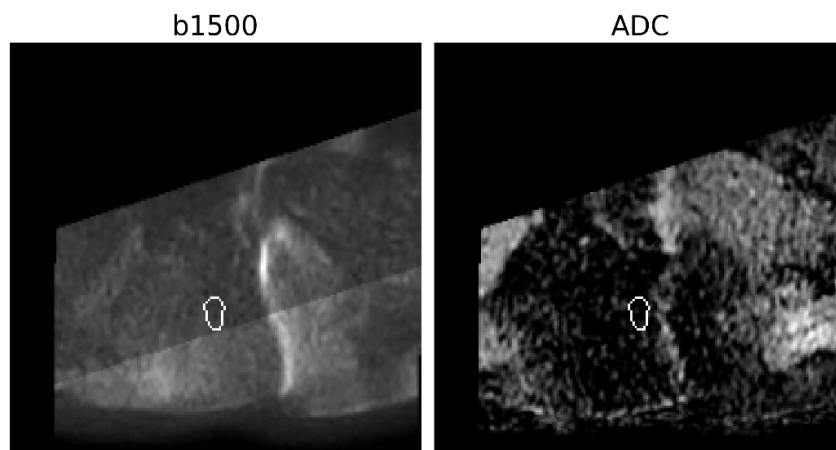
Figure 5.5: *Corresponding slice with a "cut" from the DWI-sequence with b-value 1500 (right) and the ADC-map (left) of patient 47. The edge of this "cut" can be seen in the diffusion weighted image within the tumour delineation.*

5.2.3.2. T2-weighted MR-sequences and diffusion weighted MR-sequences

Of the MR-images, diffusion weighted images were initially not processed or used for feature extraction. It was assumed that the ADC-maps captured the relevant information. The T2-weighted MR-sequences were cropped to remove slices with discontinuities affecting the area within the tumour delineation, as was done for the ADC-maps. It was initially intended to use the T2-weighted MR-sequences, but they were later excluded as it was found that the intensity ranges were not comparable between the patients as shown in Figure 5.6.
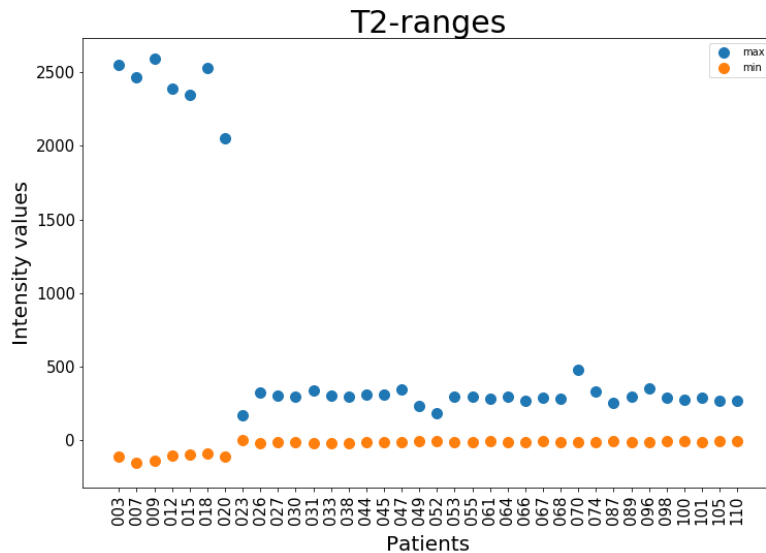
Figure 5.6: *Maximum and minimum intensity values within the tumour region of the cropped T2-weighted MR-sequences. Note that the ranges change at patient 23.*

Not including T2-weighted MR-sequences and diffusion weighted MR-sequences reduced the final number of features in data set 1. However, T2-weighted MR-sequences distinguish well between soft tissue. Extracted features from these sequences could prove useful if they were corrected in a satisfactory manner [35]. If the diffusion weighted images should be used, they should be normalized globally over all diffusion weighted sequences belonging to one patient [40].

5.2.4. Tumour mask

For patients such as the upper patient in Figure 3.3 in Methods, the tumour mask should have been modified to not include the air in the entry of the anal canal. Then, extracted shape features would not include areas outside of the body.

5.3. Feature extraction parameters

5.3.1. Bin widths and actual discretization levels

CT-sequences with modified tumour masks had the actual discretization levels within the tumour delineation that were closest to the intended discretization level after discretization. This is seen in Figure 4.1 in Results.

This was as the modified masks should only cover intensity values in the range [-150, 200], giving the patients more similar ranges. The range of actual discretization levels in the tumour was largest for CT-features extracted using the original mask. This can be explained by that the sequences that contained air had very low intensity values compared to the sequences that did not contain air. As the bin widths were computed from the average minimum and maximum intensity value of all CT-sequences belonging to a data set, the images with air affected the average minimum considerably. This gave sequences with a lower range of HU-values a lower actual discretization level, and sequences with higher range of HU-values a larger actual discretization level.

A fixed bin width gave therefore similar discretization levels for image sequences with similar intensity value ranges. The range of actual discretization levels after discretization were also less stable for the PET-sequences and ADC-maps. This as the maximum intensity value differed between patients, which is indicated in Figures 5.2 and 5.3.

It was not expected that the actual discretization levels could differ so much from the intended discretization levels when using fixed bin width. However, discretization by a fixed bin width ensures that the sequences are processed similarly and that they are not discretized based on intensity ranges that are specific for the patient the sequences belong to.

5.3.2. Correlation values between features extracted with different image settings

This section include discusses the effect of using different tumour masks, discretization levels using different bin widths and extraction dimension during feature extraction.

5.3.2.1. Correlation between observations with features extracted from different dimensions

The correlations between features extracted in different dimensions were examined. In general, first order features and NGTDM features were the least affected independently of modality. CT-features extracted with the original mask were the least affected by extraction in different dimensions for all features concerning the intensity values in the sequences.

Certain CT-features extracted using the modified tumour mask, were more affected by extraction dimension than most of the other features belonging to other modalities. The features that were affected for this mask were often affected for all specified discretization levels. The first order features Minimum and Range were affected, indicating that the tumour masks were

created such that they excluded voxels of unwanted intensity values (outside the range of [-150, 200]) more successfully in the original slices than after both the modified tumour mask and sequences had been resampled. This is a possibility, as the masks were created based on the original slices in the sequences and that the dilation of unwanted intensity values was performed in 2D and not 3D.

The heatmaps in Figure 4.2. in the Result section also showed that the correlation values between patient features extracted from different dimensions where quite high, and that CT- and ADC-features were affected similarly as the correlation values that belonged to them were quite similar for all patients. Shape features were barely affected with correlation values close to one. Only PET-features seemed to be sensitive to what dimension the features were extracted from, even though most of the correlation values between patient PET-features were higher than the highest correlation values between patient features from CT and ADC. The high variability between patient PET-features extracted at discretization levels 16, 32, 64 and 128 indicates that extraction of PET-features was more sensitive for being extracted using different extraction dimension, different discretization levels or both.

5.3.2.2. Correlation between CT-texture features extracted with different masks

Correlation between CT-features extracted with different tumour masks were examined, showing that the choice of mask affected the features extracted from two dimensions more than the features extracted from three dimensions.

The features that were the least affected from 2D extraction, were features describing high intensity values and homogeneity. High intensity values were not much affected for the CT-sequences with different tumour masks in data set 1. An explanation for this is that only one CT-sequence had intensity values higher than 200 in data set 1. That features describing homogeneity was not affected by different extraction dimension, can be explained by that the texture within the tumours in the CT-sequences in general seemed to be quite smooth.

Some of the features that were the most affected by different masks were features that described the lower intensity values and features describing asymmetry in the intensity distribution. The symmetry of the intensity value distribution would certainly change when applying the modified tumour mask, as it would remove the long tails of lower intensity values that can be seen in some of the histograms in Figure 5.1. Some examples are for patients 55 and 74. In addition, more GLRLM, GLSZM and NGTDM features were affected by being extracted in 2 dimensions than 3 dimensions using different tumour masks.

The correlation values between patients in Figure 4.4. in Results show that patient features were not greatly affected by using different tumour masks. This can be an indication of that the features were affected similarly for each patient, but that the individual features were affected differently for different patients. That the features were less affected by different masks in 3D than 2D, could also be an indication that the modified tumour mask did not fully remove all areas containing air.

5.3.2.3. Correlations between features from different discretization level

Exploration of features extracted from the same modality, but different discretization levels using fixed bin widths, showed that most were highly correlated. This indicates that it could be enough to only extract some features with fixed bin width at one discretization level. This was especially the case for first order and NGTDM features. The features from the other feature classes were also mostly highly correlated but could vary more between discretization levels, indicating that choices of discretization level for these features should be handled with care.

Only correlation values between features extracted at different discretization levels were examined. Choosing one discretization level was one way of reducing the data set, even though the resulting data sets still contained many features. It could be useful to look further into correlation values between all features, independently of their belonging to modality or feature class and remove features such that there would only be low correlation values between the remaining features. This would at least suit the classification models LDA and QDA, that reported high collinearity between the variables during model selection. However, for other models, the removal of highly correlated values might cause the loss of information [4].

5.3.2.4. Correlations

It should be noted that the correlation values that were examined were based on features extracted from the 36 patients in data set 1. This means that these results might not be representative for other features extracted from other patients with anal cancer.

5.3.3. Choices

5.3.3.1. Choice of dimension and tumour mask

It was considered whether the features should be extracted in 2D or 3D and argued if the features should either be extracted in 3D after resampling from resolution $1\times1\times3$ to resolution $1\times1\times1$ or be extracted in 2D to avoid resampling. Extraction in 2D would possibly restrict the information provided by the extracted features to the slices, while resampling for feature extraction in 3D would cause the image sequences to lose more information.

The heatmap in Figure 4.3. in the result section showed that few individual features extracted with the original tumour mask were much affected by different extraction dimension. This applied for some CT-features belonging to the GLRLM class, and some PET- and ADC-features belonging to the GLCM, GLDM, GLRLM and GLSZM classes.

For shape, CT- and ADC-features, 2D extraction with the original tumour mask could be a suitable option as CT- and ADC-features from between patients seemed to be affected similarly and shape was barely affected.

Tumour masks were created for each patient for removing air and abnormally high intensity values in the CT-sequences. These were created in 2D and forced the decision that the 2D was the best option for extraction dimension in this project. This was as the examinations of correlation values between features extracted using different masks and extraction dimension indicated that the masks did not remove unwanted intensity values as well after they had been resampled and applied to the resampled CT-sequences, as was mentioned in section 5.3.2.1.

5.3.3.2. Choice of features and processing

That the feature values seemed to be highly correlated between different discretization levels, motivated the choice of only choosing one discretization level per modality to the final data sets. This removed 5/6 of the first order and texture features belonging to the modalities that were included in each data set. The discretization levels that were chosen for each data set were partly based on the results from the initial analysis performed in [1]. These results suggested to use discretization level 128 for CT and discretization level 16 for PET The discretization levels used for ADC was not chosen based on results, but that ADC seemed to be more alike PET-sequences by that they both describe body functions. These discretization levels should be chosen more carefully in potential later projects. Especially considering that these levels were initially chosen for model selection of features extracted from sequences of patients with head and neck cancer, and not anal cancer.

The importance of robust features, features that are not dependent of factors like tumour size or discretization level, have gained more importance. This has given that corrections for such dependencies have been suggested [41] that could be interesting to apply to extracted features. Also, *PyRadiomics* provide more options for image processing before feature extraction, than was chosen for this project. Examples of parameters that were normalization of the images of the application of filters.

71

5.3.4. Model selection

The data sets had few observations compared to the number of features and were quite unbalanced. The fractions of negative PFS in the data sets and the training sets are given in Table 5.1.

	Negative PFS in entire data set	Negative PFS in training set
Data set 1	13.89% (5/36)	14.81 % (4/27)
Data set 2	16.16% (13/81)	16.17 % (10/60)

Table 5.1: *Overview of fractions of patients with negative PFS in the data sets created for this project. The fractions are given as percentages. The rows belong to the data sets, the columns belong to the entire data sets (left) and the training set of the data set (right).*

The cross validations required that each fold had at least one observation of each class, giving that the number of folds in cross validation could only be as high as the number of observations with negative outcome in the data set model selection was performed for. This number of folds was both used during optimization of feature selector and classifier parameters and cross validation of the resulting model from this optimization. Model selection was performed on the training sets of data sets 1 and 2. This gave that three folds were used for the training set of data set 1 and that 10 folds were used for the training set of data set 2. The number of folds could have been set to four for the training set of data set 1 to increase the number of folds. The low number of folds gave that the feature selector and classifier parameters were only updated a few times and that the resulting model from this optimization was cross validated a few times for each data set. This gives that the optimal combinations of parameters might not have been found for the few folds of cross validations, and that the mean train and validation AUCs were based on scores of high variability, as they were from few cross validations. It was thus expected that there would be high bias for the resulting validation and train AUCs. This seemed especially to be the case for the validation scores from the model selection on data set 2. The mean variances of the validation AUCs from model selection on the training set of data set 2 were in general higher than for model selection performed on the training set of data set 1.

	No information rate	Highest mean validation score	Model
Training set of data set 1	85.19 %	71.24 %	LDA with Fisher Score feature selection
Training set of data set 2	83.33 %	76.93 %	LightGBM with Wilcoxon feature selection

Table 5.2: *Overview of no information rate (left column) and highest mean validation AUC (right column) from model selection on the training sets of data sets 1 and 2. These are given in percentage and with two decimals. The rows belong to the data sets used for model selection.*

The no information rates and highest mean validation AUCs from each model selection are given in Table 5.2. No mean validation scores were higher than the no information rates, the scores that would be obtained if the models had predicted that all patients had positive PFS.

Possible reasons for this, are that the images had not been acquired or processed such that the features extracted did not fully represented the tumour or provided information distinguishing

between the patients with different PFS. That the features extracted could not distinguish between the groups of patients, and that features from the other MR-modalities might have been useful. The high number of features might have also contributed; The feature set could also have been reduced more beforehand by removing features such that no remaining features would be highly correlated. The data sets should ideally also have contained more patients to more fully optimize the models before model training and give AUCs with less variance in each experiment. Reconsideration of image processing and further reduction of the features could thus improve the validation AUCs.

Overall, the classifier LDA gave the highest validation scores, 0.69 – 0.71, with Wilcoxon feature selection, Chi Square feature selection and Fisher Score feature selection for the training set of data set 1. For model selection on the training set of data set 2, no models stood out from the other based on mean validation AUC. Several models gave mean AUCs close to the highest mean validation AUC, of which two of these models included the classifier LightGBM. The differences between the models that gave the highest AUCs, were generally high, meaning that the models were quite overfitted, with train scores close to 1. This was especially for the training set of data set 1.

5.3.4.1. Feature selection

One of the main goals with radiomics in oncology is to find biomarkers that are useful for predicting the state and treatment outcome of the patient. What features that were chosen the most during model selection with different model combinations of classifier and feature selectors were, not surprisingly, more dependent of feature selector than classifier. Wilcoxon was the feature selection method that chose less features several times compared to the other feature selectors.

For data set 1, two features were chosen often independently of model. These were the PET and texture feature Small Dependence Low Gray Level Emphasis (SDLGLE) and the shape feature Sphericity. SDLGLE belongs to the GLDM class and describes dependency between voxels of low intensity values, [29] while Sphericity describes the roundness of the tumour. These were chosen approximately half of the experiments.

For data set 2, three features were chosen more than half of the times. These were all first order PET-features, Skewness, Kurtosis and Uniformity. Both Skewness and Kurtosis describe the intensity value distribution of the sequences. Kurtosis describes how narrow the peak of the distribution is, while Skewness describes asymmetrical the distribution is. The peaks of most intensity distributions of the PET-sequences were located towards the lower SUV-values, as the tumour delineation in most PET-sequences cover more than the areas with higher intensity values. This is seen in Figure 5.2. These features might also be affected by the maximum SUV-values, how much of the surrounding tissue that is included in the tumour delineation and how well the tumour delineation was performed. The Uniformity PET-feature is the sum of squares of the SUV-values corrected for the number of voxels within the tumour mask [29]. A PET-sequence with larger areas of high SUV-values will have higher value of Uniformity. This gives that the features that were chosen most for data set 2 might have been chosen because they described the amount of high and low intensity values within the tumour.

For data set 2, more features were chosen a higher percentage of the time than for data set 1. This might be explained by that data set 2 contained fewer features than data set 1 but might also indicate that the features in this data were better predictors for the outcome. For both data sets, the majority of the 50 most chosen features were PET-features, this was especially the case for data set 2. A simple search for features that had been selected most from both model selections found eight features, either PET- or shape features. These included the shape features Sphericity, Elongation, Surface Volume Ratio, the NGTDM features Coarseness and Contrast and the GLRLM features Short Run Low Gray Level Emphasis and Gray Level Non Uniformity Normalized and the GLDM feature Large Dependence High Gray Level Emphasis.

This is interesting as PET-features seemed to be chosen more often than the features extracted from the other modalities in both data sets, even though none of these were among the ones that were chosen more than half of the time for each data set. Also, shape and NGTDM belong to classes that make up only a small fraction of the feature spaces for both data sets and would reduce the feature sets considerably if they showed to be important. That these eight features were among the most chosen in both data sets, indicate that they might explain PFS for anal cancer patients better than other features. That no CT-features were chosen once might indicate that the modified tumour mask might have removed too much information or that CT-features are not as important for predicting treatment outcome as PET- and shape features. Data set 2 did not contain ADC-features, so it is difficult to conclude whether ADC-features could prove to be important predictors, even though there were as many ADC- and CT-features among the most chosen features from data set 1.

One of the initial goals were to examine whether MR-features were good predictors for treatment outcome of anal cancer. Only ADC-features were examined, as the T2-weighted MR-sequences were not processed in a satisfactory way giving that they were given with different ranges of intensity values. The ADC-features were neither examined separately from features extracted from other image modalities, meaning that the results in this project did not really reflect the importance of ADC-features. The ADC-features did not particularly stand out from the other features during model selection, as PET-features were chosen more often. However, the highest validation AUCs from this model selection were not as low as expected and ADC-features were represented among the most chosen features, meaning that they were evaluated to stand out from the other features of provide information pertaining to the treatment outcome by the feature selectors. The inclusion of T2-weighted sequences of diffusion weighted MR-sequences and further exploration of these and ADC-features could better determine the importance of MR-features.

5.3.4.2. Intended validation
It was intended to extract the most chosen parameters and features to train models on the training sets and apply on the test sets. The test sets had not been used for model selection and would thus give an indication of the performance of the chosen models on unseen data. However, as the training sets were even smaller, containing 9 and 21 observations, in data set 1 and data set 2, it is not necessarily expected that these experiments would give satisfactory results.

5.3.4.3. Software and reproducibility of project

Both *PyRadiomics* and *biorad* were downloaded from source in GitHub, giving that it might not be possible to download the specific versions used in this project for reproduction if this should be desirable. *biorad* was also under more intensive development than *PyRadiomics* during this project such that it became necessary that the author of the software provided extensions for model selection when model selection was to be performed on the data sets.

biorad was also developed for feature extraction and model selection of images belonging to patients with head and neck cancer. The parameters were only changed to suit the size of the data sets for patients with anal cancer. Thus, there is potential for exploring and evaluation other choices of included images, image processing, feature extraction parameters, model selection parameters and choice of metric to potentially improve the performance of classification of treatment outcome for patients with anal cancer.

6. Conclusion

Two data sets were created with shape, PET-, CT- and ADC-features extracted from image sequences of patients with anal cancer. Data set 1 contained features extracted from PET-sequences, CT-sequences and ADC-maps from 36 patients and data set 2 contained features extracted from PET- and CT-sequences from 81 patients. The predictive performance of the extracted features was examined by model selection with *biorad*, using 10 classifiers and five feature selectors. The effect of different tumour masks, different discretization levels with different bin widths and different extraction dimensions were examined by computing correlations between individual features and patient features extracted with different settings.

The features extracted from the same image but with different discretization levels were mostly highly correlated, indicating that it could be enough to extract features from images of one discretization level for some features. This was especially the case for extracted CT-features and features belonging to the first order feature class and the texture NGTDM feature class.

The tumour masks were created by excluding areas with HU-values outside the range of [-150, 200] that had been dilated in 2D. The modified tumour mask did not seem to adapt well to resampled image sequences. It was also a concern that the chosen lower threshold of unwanted HU-values (-150) was too high, possibly removing information unnecessarily.

Correlation values between features extracted in different dimensions using the original tumour masks indicated that extraction in 2D was an acceptable choice, as they were mostly highly correlated. This was especially the case for shape features and features extracted from CT and ADC-sequences using the original tumour mask.

It should be noted that the features that were examined belonged to 36 patients and might not be representative for features extracted from image sequences of other patients with anal cancer.

Model selection did not give higher validation AUCs than the no information rate. The no information rates for the training sets of data set 1 and data set 2, were 0.85 and 0.83, respectively. The highest mean validation AUC that was obtained for the training set of data set 1 was 0.71. This was found by using the classifier LDA and Fisher Score feature selection. The highest mean validation AUC that was obtained for the training set of data set 2 was 0.77. This was found by using three selector-classifier combinations; Decision Tree with Chi Square feature selection, LightGBM with Mutual Information feature selection and LightGBM with Wilcoxon feature selection.

Eight PET- and shape features were among the most selected features from each model selection. Most of the features that were among the most selected features from both model selections were PET-features. That T2-weighted MR-sequences were excluded from the data set, and that the ADC-features were only present in one of the data sets and was not examined separately makes it difficult to reach conclusions about their predictive performance.

These results are preliminary and indicative due to the low number of patients. They show potential as the validation scores were not as far apart from the no information rate as initially feared. More patients, different image processing steps, different extraction choices like filters, discretization levels, different tumour masks, further reduction of features and the inclusion of T2-weighted images could improve the predictive performance of PFS.

7. References

[1] A. Midtfjord, "Prediction of treatment outcome of head and throat cancer using radiomics of PET/CT images," NMBU, Ås, 2018.

[2] G. S. R. E. Langberg, «biorad,» Software, downloaded from source 28.03.2019, https://github.com/GSEL9/biorad.

[3] M. Cabrol, "Unpublished: From images to a data set - making a data set from CT-, PET- and MR-images," Ås, 2018.

[4] I. Gardin, V. Grégoire, G. D., H. Kirisli, D. Pasquier, J. Thariat and P. Vera, "Radiomics: Principles and radiotherapy applications," 29 March 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1040842818302166. [Accessed 14 May 2019].

[5] B. Roald, T. Sauer and O. Klepp, "kreft. In: Store medisinske leksion," 29 October 2018. [Online]. Available: https://sml.snl.no/kreft. [Accessed 14 May 2019].

[6] F. Valvo, E. Ciurlia, B. Avuzzi, R. Doci, M. Ducreux, F. Roelofsen, A. Roth, A. Trama and C. B. J. F. Wittekind, "Cancer of the anal region," March 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1040842817302408. [Accessed 14 May 2019].

[7] O. Klepp, "analcancer. In: Store medisinske leksikon," 13 February 2009. [Online]. Available: https://sml.snl.no/analcancer. [Accessed 14 May 2019].

[8] M. A. Haidekker, "Computed Tomography. In: Medical Imaging Technology," 18 April 2013. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4614-7073-1_3. [Accessed 30 March 2019].

[9] Oslo universitetssykehus, "CT på Radiumhospitalet," 11 September 2018. [Online]. Available: https://oslo-universitetssykehus.no/behandlinger/ct?sted=avdeling-for-radiologi-og-nukleermedisin-radiumhospitalet. [Accessed 3 April 2019].

[10] M. A. Haidekker, "X-Ray Projection Imaging In: Medical Imaging Technology," 18 April 2013. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4614-7073-1_2. [Accessed 31 April 2019].

[11] J. Lilley, Nuclear Physics Principles and Applications, Chichester: John Wiley & Sons, 2001.

[12] T. E. o. E. Britannica, "Pair production," 5 March 2008. [Online]. Available: https://www.britannica.com/science/pair-production. [Accessed 1 April 2019].

[13] Principles and practice of PET and PET/CT, 2 ed., Philadelphia: Lippinkott Williams & Wilkins, a Wolters Kluwer Business, 2009.

[14] I. Bickle and K. Greenway, "Hounsfield Unit," 2019. [Online]. Available: https://radiopaedia.org/articles/hounsfield-unit. [Accessed 3 April 2019].

[15] M. Khalil, "Positron Emission Tomography (PET):Basic Principles. In Khalil M. (eds) Basic Sciences of Nuclear Medicine," 2010. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-540-85962-8_11. [Accessed 2 April 2019].

[16] M. A. Haidekker, "Nuclear Imaging In: Medical Imaging Technology," 18 April 2013. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4614-7073-1_4. [Accessed 1 April 2019].

[17] Oslo universitetssykehus, "PET-undersøkelse," 7 December 2016. [Online]. Available: https://oslo-universitetssykehus.no/behandlinger/pet-undersokelse. [Accessed 3 April 2019].

[18] G. Saha, "Data Acquisition and Corrections. In: Basics of PET Imaging," 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-16423-6_3. [Accessed 1 April 2019].

[19] G. Saha, "Topics of interest. In_ Basics of PET Imaging," Springer, 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-16423-6_14. [Accessed 1 April 2019].

[20] M. Brekke and A. Borthne, "MR-undersøkelse In: Store medisinske leksikon," 22 Februar 2019. [Online]. Available: https://sml.snl.no/MR-unders%C3%B8kelse. [Accessed 4 April 2019].

[21] Oslo universitetssykehus, "MR-undersøkelse på Radiumhospitalet," 21 August 2018. [Online]. Available: https://oslo-universitetssykehus.no/behandlinger/mr-undersokelse?sted=radiumhospitalet. [Accessed 3 April 2019].

[22] D. McRobbie, MRI From picture to Proton, 2 ed., New York: Cambridge University Press, 2006.

[23] R. H. Hashemi and W. G. Bradley Jr., MRI the basics, Baltimore: Williams & Wilkins, 1997.

[24] A. Quayym, "Diffusion-weighted Imaging in the Abdomen and Pelvis: Concepts and Applications," 1 October 2009. [Online]. Available: https://pubs.rsna.org/doi/10.1148/rg.296095521. [Accessed 18 March 2019].

[25] D. Bell and A. Goel, "b values," 2018. [Online]. Available: https://radiopaedia.org/articles/b-values-1?lang=us. [Accessed 18 March 2019].

[26] J. Sánchez-González and J. Lafuente-Mártinez, "Diffusion-Weighted Imaging: Acquisition and Biophysical Basis. In: Diffusion MRI Outside the Brain," 2012. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-21052-5_1. [Accessed 18 March 2019].

[27] Stanford University, "Calculation of measured ADC maps," [Online]. Available: http://rsl.stanford.edu/moseley/tensorcalc/tensorcalc/Steps/ADCmaps/adc.html. [Accessed 18 March 2019].

[28] L. A. Sánchez-González J., "Quantification and Postprocessing of DWI. In: Diffusion MRI Outside the Brain," 11 08 2011. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-642-21052-5_3. [Accessed 18 March 2019].

[29] Pyradiomics Community, "PyRadiomics documentation," 2017. [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/. [Accessed 14 May 2019].

[30] G. L. Løvås, Statistikk for universitet og høgskoler, Oslo: Universitetsforlaget, 3.

[31] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Roin, S. Pieper and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107," 2017.

[32] OncoLink, "All about anal cancer," 9 January 2019. [Online]. Available: https://www.oncolink.org/cancers/gastrointestinal/anal-cancer/all-about-anal-cancer. [Accessed 22 February 2019].

[33] J. Ovrum, "Intensitetsmodulert stråleterapi (IMRT) av prostatakreft. En analyse basert på ekvivalent uniform dose (EUD)," UiO, 2016.

[34] E. Malinen, *HU-values without clinical information,* Personal communication, 2019.

[35] E. Rusten, *Removal of air from CT-sequences,* E-mail, 12.03.2019.

[36] R. T. H. Leijenaar, S. Carvalho, E. R. Velasquez, W. J. C. van Elmpt, P. C., O. S. Hoekstra, C. J. Hoekstra, R. Boellaard, A. L. A. J. Dekker, R. J. Gillies, H. J. W. L. Aerts and P. Lambin, "Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and interobserver variability," *Acta Oncologica, 52:7, 1391-1397,* 2013.

[37] A. Zwanenburg, S. Leger, M. Vallières and S. Löck, "Image biomarker standardisation initiative," 21 December 2016. [Online]. Available: https://arxiv.org/abs/1612.07003. [Accessed 25 February 2019].

[38] K. Lloyd, P. L. Rosin, D. Marchall and e. al, "Machine Visions and Applications," May 2017. [Online]. Available: https://link.springer.com/article/10.1007%2Fs00138-017-0830-x. [Accessed 10 May 2019].

[39] National Cancer Institute, "Complete response," [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/complete-response. [Accessed 12 April 2019].

[40] E. Malinen, *Progression free survival and image processing,* Phone conversation, 12.04.2019.

[41] M. Shafiq-ul-Hassan, K. Latifi, G. Zhang, G. Ullah, R. Gillies and E. Moros, "Voxel size and gray level normalization of CT radiomic features in lung cancer," 12 July 2018. [Online]. Available: https://www.nature.com/articles/s41598-018-28895-9. [Accessed 15 May 2019].

Appendices

A.1. Installed Python packages

alabaster	0.7.12	pypi_0	pypi
babel	2.6.0	pypi_0	pypi
blas	1.0	mkl	anaconda
bottleneck	1.2.1	py36h452e1ab_1001	conda-forge
ca-certificates	2018.11.29	ha4d7672_0	conda-forge
81elief81	2018.11.29	py36_1000	conda-forge
chardet	3.0.4	pypi_0	pypi
colorama	0.4.1	pypi_0	pypi
configspace	0.4.8	pypi_0	pypi
cycler	0.10.0	py_1	conda-forge
cython	0.29.5	pypi_0	pypi
decorator	4.3.2	py_0	conda-forge
docopt	0.6.2	pypi_0	pypi
docutils	0.14	pypi_0	pypi
emcee	2.2.1	pypi_0	pypi
freetype	2.9.1	h5db478b_1005	conda-forge
future	0.17.1	py36_1000	conda-forge
81elief	0.3.1	pypi_0	pypi
hyperopt	0.1.1	py_0	conda-forge
icc_rt	2019.0.0	h0cc432a_1	anaconda
icu	58.2	ha66f8fd_1	anaconda
idna	2.8	pypi_0	pypi
imagesize	1.1.0	pypi_0	pypi
imbalanced-learn	0.4.3	py_0	conda-forge
intel-openmp	2019.1	144	anaconda
jinja2	2.10	pypi_0	pypi
joblib	0.13.1	py36_0	anaconda
jpeg	9c	hfa6e2cd_1001	conda-forge
kiwisolver	1.0.1	py36he980bc4_1002	conda-forge
libiconv	1.15	hfa6e2cd_1004	conda-forge
libpng	1.6.36	h7602738_1000	conda-forge
libtiff	4.0.10	h36446d0_1001	conda-forge
libxml2	2.9.3	vc14_9	conda-forge
lightgbm	2.2.3	pypi_0	pypi
llvmlite	0.27.0	py36h6538335_0	numba
markupsafe	1.1.0	pypi_0	pypi
matplotlib	3.0.2	py36_1002	conda-forge
matplotlib-base	3.0.2	py36h3e3dc42_1002	conda-forge
mifs	0.0.1.dev0	py36_0	pchrapka
mkl	2019.1	144	anaconda
mkl_fft	1.0.10	py36h14836fe_0	anaconda
mkl_random	1.0.2	py36h343c172_0	anaconda
mlxtend	0.13.0	py_1	conda-forge
natsort	5.5.0	py36_0	anaconda
81elief81	2.2	py_1	conda-forge
nose	1.3.7	pypi_0	pypi
numba	0.42.0	np115py36h6538335_0	numba
numpy	1.15.4	py36h19fb1c0_0	anaconda
numpy-base	1.15.4	py36hc3f5095_0	anaconda
openssl	1.1.1a	hfa6e2cd_1000	conda-forge
os	0.1.3	0	jmcmurray

packaging	19.0	pypi_0	pypi
pandas	0.24.1	py36ha925a31_0	anaconda
pip	19.0.2	py36_0	conda-forge
psutil	5.5.1	pypi_0	pypi
pybind11	2.2.4	py36he980bc4_1000	conda-forge
pydoe	0.3.8	pypi_0	pypi
pykwalify	1.7.0	pypi_0	pypi
pymongo	3.7.1	py36h6538335_1000	conda-forge
pynisher	0.5.0	pypi_0	pypi
pynrrd	0.3.6	pypi_0	pypi
pyparsing	2.3.1	py_0	conda-forge
pyqt	5.9.2	py36ha878b3d_0	anaconda
pyradiomics	0+unknown	pypi_0	pypi
pyrfr	0.8.0	pypi_0	pypi
python	3.6.6	he025d50_0	conda-forge
python-dateutil	2.8.0	py_0	conda-forge
python-utils	2.3.0	py_1	conda-forge
pytz	2018.9	py36_0	anaconda
pywavelets	1.0.0	pypi_0	pypi
pyyaml	3.13	pypi_0	pypi
qt	5.9.7	vc14h73c81de_0	
qutil	3.2.1	6	jmcmurray
82elief	0.1.2	pypi_0	pypi
requests	2.21.0	pypi_0	pypi
scikit-feature	1.0.0	py36_0	pchrapka
scikit-learn	0.20.2	py36h343c172_0	
scipy	1.2.0	py36h29ff71c_0	
setuptools	40.8.0	py36_0	conda-forge
simpleitk	1.2.0	pypi_0	pypi
sip	4.19.13	py36ha925a31_0	anaconda
six	1.12.0	py36_1000	conda-forge
smac	0.10.0	pypi_0	pypi
snowballstemmer	1.2.1	pypi_0	pypi
sobol-seq	0.1.2	pypi_0	pypi
sphinx	1.8.4	pypi_0	pypi
sphinx-rtd-theme	0.4.3	pypi_0	pypi
sphinxcontrib-websupport	1.1.0	pypi_0	pypi
sqlite	3.26.0	he774522_0	anaconda
statsmodels	0.9.0	pypi_0	pypi
swig	3.0.12	2	conda-forge
tornado	5.1.1	py36hfa6e2cd_1000	conda-forge
typing	3.6.6	pypi_0	pypi
urllib3	1.24.1	pypi_0	pypi
vc	14.1	h21ff451_3	anaconda
vs2015_runtime	15.5.2	3	anaconda
wheel	0.32.3	py36_0	conda-forge
wincertstore	0.2	py36_1002	conda-forge
xgboost	0.82	pypi_0	pypi
xlrd	1.2.0	py36_0	anaconda
zlib	1.2.11	h2fa13f4_1004	conda-forge

B.1. Python scripts: Cropping functions

```python
# -*- coding: utf-8 -*-

"""Cropping functions"""
import nrrd
import numpy as np


__author__ = 'Maria Cabrol'
__email__ = 'mariacab@nmbu.no'


def slices_w_tumour(img_path):
    """
    Finds first and slice in image that contain tumour delineation

    :param   img_path: path of cropped image

    :return: int, start_tumour and end_tuour. Position of first and last slice in image (z-
    direction) that contain
             tumour delineation
    """
    img, _ = nrrd.read(img_path)
    img_end = np.shape(img)[2]-1

    if tumour_all_slices(img_path):
        return 0, img_end

    else:
        z = 0
        empty = True
        while empty:
            if len(np.unique(img[:, :, z])) > 1:
                empty = False
                z -= 1
            z += 1
        start_tumour = z

        z = img_end
        empty = True
        while empty:
            if len(np.unique(img[:, :, z])) > 1:
                empty = False
                z += 1
            z -= 1
        end_tumour = z

        return start_tumour, end_tumour

def tumour_crop(mask_path, img):
    """
    Crop image such that only slices with tumour delineation are contained in image

    :param mask_path: (str), path of mask
    :param img:       nrrd_type, image to crop

    :return: None
    """

    tumour_start, tumour_end = slices_w_tumour(mask_path)

    return img[:, :, tumour_start:tumour_end + 1]
```

```python
def slice_crop(img, start=0, end=0):
    """
    Crop images in z-direction, e.g. removes chosen slices.

    :param img:   (image read by nrrd), image to crop
    :param start: (int), number of slices from beginning to crop
    :param end:   (int), number of slices from end to crop

    :return: cropped image
    """

    if start < 0 or end < 0:
        raise ValueError('Both start and end must have value >= 0')

    if end == 0:
        cropped_img = img[:, :, start:]
    else:
        cropped_img = img[:, :, start:-end]

    return cropped_img
```

B.2. Python scripts: CT modified tumour mask

```python
# -*- coding: utf-8 -*-

"""Create mask for CT-images that leave out air and abnormally high intensity values"""

import os
import re
import nrrd
import numpy as np


__author__ = 'Maria Cabrol'
__email__  = 'mariacab@nmbu.no'


def create_dir(dir_name):
    """
    dir_name: string, name of new directory.

    OBS: Creates new directory in current working directory.
    Code from Stack overflow
    """

    try:
        os.mkdir(dir_name)
    except OSError:
        print('Creation of directory %s failed' % dir_name)
    else:
        print('Successfully created the directory %s' % dir_name)
    return None


def dilate_mask_square_2d (mask, dist):
    """
    Dilates mask in x- and y-direction. square shape.
    Requires that the tumour is not near the edges in the slices

    :param mask:    mask to dilate
    :param dist:    number of pixels to dilate from mask contour
    :return:        dilated image
    """

    dilated_mask = mask.copy()      # New mask
    temp_mask = mask.copy()         # Temporary mask
    struct = struct_square_3d()[1] # Structuring element

    # Remove outer layer of pixels dist times
    for d in range(dist): # Find contour that surrounds the mask (outer layer of
                          mask pixel wise), and its coordinates
        shell = contour_3d(temp_mask)
        shell_coors = np.where(shell == 1)

        # Expand mask
        for coor in range(len(shell_coors[0])):

            i, j, k = shell_coors[0][coor], shell_coors[1][coor], shell_coors[2][coor]

            dilated_mask[i-1:i+2, j-1:j+2, k] = dilated_mask[i-1:i+2, j-1:j+2, k][struct == 1]
= 1

            # Update temporary mask
        temp_mask = dilated_mask.copy()

    return dilated_mask
```

85

```python
os.chdir('E://Master/Cropped')
# Create directory where all images are to be stored
dir_name = 'ROI_PETCT2'  # masks are dilated 2 times, -150 lowest value that is not set to
zero

create_dir(f'{dir_name}')

for mask_file in os.listdir('E://Master/Cropped/ROI_cropped'):
    # only patient ID should contain numbers in file_name
    tag = re.search(r'\d+', mask_file).group()

    ct_img, _ = nrrd.read(f'Base/PETCT_cropped/petct{tag}_cropped.nrrd')
    ct_mask, header = nrrd.read(f'ROI_cropped/{mask_file}')

    cropped = ct_img * ct_mask

    # Create mask for low and high intensity values within tumour area. The low values
represent air,
    # and the high values are not considered to give clinical information
    air_mask = np.zeros(np.shape(ct_img))
    air_mask[cropped < -150] = 1
    air_mask[cropped > 200] = 1

    air_mask_new = dilate_mask_square_2d(air_mask, 2)
    # air_mask_new = air_mask.copy()

    ct_mask[air_mask_new == 1] = 0
    ct_mask[cropped > 200] = 0

    nrrd.write(f'E://Master/Cropped/{dir_name}/ct_mask{tag}_cropped.nrrd', ct_mask,
header=header)

os.chdir('../')
```

B.3. Python scripts: Computation of fixed bin width

```python
# -*- coding: utf-8 -*-

import os
import re
import nrrd
import numpy as np
import pandas as pd

__author__ = 'Maria Cabrol'
__email__ = 'mariacab@nmbu.no'


def bin_widths_tumour(image_path, mask_path, n_bins, normalization=False,
remove_outliers=False):
    """
    From biorad2, author Severin Langberg
    Compute bin width for tumour histogram

    :param image_path:      path to image directory
    :param mask_path:       path to mask directory
    :param n_bins:          desired number of bins
    :param normalization:   boolean, default False, if entire image is to be normalized
    :param remove_outliers: boolean, default False, if outliers are to be removed from the
images after normalization

    :return: bin_width
    """

    if not normalization and remove_outliers:
        raise ValueError('Set normalization to true before removing outliers')

    # Lists to store min and max values
    img_min = np.zeros(len(os.listdir(image_path)))
    img_max = np.zeros(len(os.listdir(image_path)))

    for i, image_file in enumerate(os.listdir(image_path)):
        img, _ = nrrd.read(f'{image_path}/{image_file}')

        # Find patient ID for finding correct mask
        tag = re.search(r'\d+', image_file).group()

        for mask_file in os.listdir(mask_path):
            if tag in mask_file:
                mask, _ = nrrd.read(f'{mask_path}/{mask_file}')

        if normalization:
            # Normalize image
            norm_img = (img-np.mean(img))/np.std(img)

            if remove_outliers:
                # Remove outliers
                norm_img[norm_img > 3] = 3
                norm_img[norm_img < -3] = -3

            img_max[i] = np.max(norm_img[mask != 0])
            img_min[i] = np.min(norm_img[mask != 0])

        else:
            img_max[i] = np.max(img[mask != 0])
            img_min[i] = np.min(img[mask != 0])

    return (np.mean(np.ceil(img_max)) - np.mean(np.floor(img_min))) / n_bins
```

C.1. *PyRadiomics:* Example of parameter file

```
# -*- coding: utf-8 -*-
#
# PyRadiomics version: unknown, PyRadiomics was downloaded from source

setting:
  # Histogram bin size and intensity value quantification (PET with
  # discretization level 64).
  binWidth: 0.2886
  # All image intensity values > 0: no need to shift voxels.
  voxelArrayShift: 0
  # Distance to neighbour: Specifies the distances between the center voxel and
  # the neighbor, for which angles should be generated. Only affects the GLCM
  # and NGTDM feature classes. The GLSZM and GLRLM feature classes use a fixed
  # distance of 1 (infinity norm) to define neighbours.
  distances: [1]
  # Value of ROI in mask (is binary).
  label: 1
  # Disable output of additional information from the extraction.
  additionalInfo: False
  # Only extract features from the slices (the xy-plane).
  force2Ddimension: 0

imageType:
  # Original filter: No filter applied.
  Original: {}


featureClass:
  #shape:

  # Enable all possible features in firstorder class.
  firstorder:
  # Enable specific features in glcm class.
  glcm:
    - 'Autocorrelation'
    - 'JointAverage'
    - 'ClusterProminence'
    - 'ClusterShade'
    - 'ClusterTendency'
    - 'Contrast'
    - 'Correlation'
    - 'DifferenceAverage'
    - 'DifferenceEntropy'
    - 'DifferenceVariance'
    - 'JointEnergy'
    - 'JointEntropy'
    - 'Imc1'
    - 'Imc2'
    - 'Idm'
    - 'MCC'
    - 'Idmn'
    - 'Id'
    - 'Idn'
    - 'InverseVariance'
    - 'MaximumProbability'
    - 'SumEntropy'
    - 'SumSquares'

  # Enable all possible features in glrlm class.
  glrlm:
  # Enable all possible features in glszm class.
  glszm:
  # Enable all possible features in gldm class.
  gldm:
  # Enable all possible features in ngtdm class.
  ngtdm:
```

C.2. *PyRadiomics:* List of extracted feature

Shape:

```
'original_shape_Elongation',
'original_shape_Flatness',
'original_shape_LeastAxisLength',
'original_shape_MajorAxisLength',
'original_shape_Maximum2DDiameterColumn',
'original_shape_Maximum2DDiameterRow',
'original_shape_Maximum2DDiameterSlice',
'original_shape_Maximum3DDiameter',
'original_shape_MeshVolume',
'original_shape_MinorAxisLength',
'original_shape_Sphericity',
'original_shape_SurfaceArea',
'original_shape_SurfaceVolumeRatio',
'original_shape_VoxelVolume']
```

Firstorder and texture:

```
'original_firstorder_10Percentile',
'original_firstorder_90Percentile',
'original_firstorder_Energy',
'original_firstorder_Entropy',
'original_firstorder_InterquartileRange',
'original_firstorder_Kurtosis',
'original_firstorder_Maximum',
'original_firstorder_MeanAbsoluteDeviation',
'original_firstorder_Mean',
'original_firstorder_Median',
'original_firstorder_Minimum',
'original_firstorder_Range',
'original_firstorder_RobustMeanAbsoluteDeviation',
'original_firstorder_RootMeanSquared',
'original_firstorder_Skewness',
'original_firstorder_TotalEnergy',
'original_firstorder_Uniformity',
'original_firstorder_Variance',
'original_glcm_Autocorrelation',
'original_glcm_JointAverage',
'original_glcm_ClusterProminence',
'original_glcm_ClusterShade',
'original_glcm_ClusterTendency',
'original_glcm_Contrast',
'original_glcm_Correlation',
'original_glcm_DifferenceAverage',
'original_glcm_DifferenceEntropy',
'original_glcm_DifferenceVariance',
'original_glcm_JointEnergy',
'original_glcm_JointEntropy',
'original_glcm_Imc1',
'original_glcm_Imc2',
'original_glcm_Idm',
'original_glcm_MCC',
'original_glcm_Idmn',
```

```
'original_glcm_Id',
'original_glcm_Idn',
'original_glcm_InverseVariance',
'original_glcm_MaximumProbability',
'original_glcm_SumEntropy',
'original_glcm_SumSquares',
'original_glrlm_GrayLevelNonUniformity',
'original_glrlm_GrayLevelNonUniformityNormalized',
'original_glrlm_GrayLevelVariance',
'original_glrlm_HighGrayLevelRunEmphasis',
'original_glrlm_LongRunEmphasis',
'original_glrlm_LongRunHighGrayLevelEmphasis',
'original_glrlm_LongRunLowGrayLevelEmphasis',
'original_glrlm_LowGrayLevelRunEmphasis',
'original_glrlm_RunEntropy',
'original_glrlm_RunLengthNonUniformity',
'original_glrlm_RunLengthNonUniformityNormalized',
'original_glrlm_RunPercentage',
'original_glrlm_RunVariance',
'original_glrlm_ShortRunEmphasis',
'original_glrlm_ShortRunHighGrayLevelEmphasis',
'original_glrlm_ShortRunLowGrayLevelEmphasis',
'original_glszm_GrayLevelNonUniformity',
'original_glszm_GrayLevelNonUniformityNormalized',
'original_glszm_GrayLevelVariance',
'original_glszm_HighGrayLevelZoneEmphasis',
'original_glszm_LargeAreaEmphasis',
'original_glszm_LargeAreaHighGrayLevelEmphasis',
'original_glszm_LargeAreaLowGrayLevelEmphasis',
'original_glszm_LowGrayLevelZoneEmphasis',
'original_glszm_SizeZoneNonUniformity',
'original_glszm_SizeZoneNonUniformityNormalized',
'original_glszm_SmallAreaEmphasis',
'original_glszm_SmallAreaHighGrayLevelEmphasis',
'original_glszm_SmallAreaLowGrayLevelEmphasis',
'original_glszm_ZoneEntropy',
'original_glszm_ZonePercentage',
'original_glszm_ZoneVariance',
'original_gldm_DependenceEntropy',
'original_gldm_DependenceNonUniformity',
'original_gldm_DependenceNonUniformityNormalized',
'original_gldm_DependenceVariance',
'original_gldm_GrayLevelNonUniformity',
'original_gldm_GrayLevelVariance',
'original_gldm_HighGrayLevelEmphasis',
'original_gldm_LargeDependenceEmphasis',
'original_gldm_LargeDependenceHighGrayLevelEmphasis',
'original_gldm_LargeDependenceLowGrayLevelEmphasis',
'original_gldm_LowGrayLevelEmphasis',
'original_gldm_SmallDependenceEmphasis',
'original_gldm_SmallDependenceHighGrayLevelEmphasis',
'original_gldm_SmallDependenceLowGrayLevelEmphasis',
'original_ngtdm_Busyness',
'original_ngtdm_Coarseness',
'original_ngtdm_Complexity,
'original_ngtdm_Contrast,
'original_ngtdm_Strength,
```