Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

Aase Mellingen Langan

MRI-Based Radiomics Analysis for Predicting Treatment Outcome in Rectal Cancer

Master's thesis in MTNANO
Supervisor: Kathrine Røe Redalen
June 2020

NTNU
Norwegian University of
Science and Technology

Aase Mellingen Langan

MRI-Based Radiomics Analysis for Predicting Treatment Outcome in Rectal Cancer

Master's thesis in MTNANO
Supervisor: Kathrine Røe Redalen
June 2020

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

NTNU
Norwegian University of
Science and Technology

ABSTRACT

In an attempt to contribute to the investigation of radiomic features from magnetic resonance images (MRI) as potential biomarkers within the field of rectal cancer, this thesis had tree main objectives. First, to establish and optimize binary prediction models with progression free survival (PFS), tumor regression grade (TRG) and posttherapy pathological T-staging (ypT) as endpoints. The latter two are indicative of response to neoadjuvant chemoradiotherapy (nCRT). Second, to investigate the predictive and prognostic value of texture features in particular, and finally, determine the reproducibility of obtained results with respect to voxel dimension, tumor delineation and intensity discretization.

Combinations of four feature selector algorithms and six classifiers were evaluated. Shape, first-order statistical and texture features were derived from T2- and diffusion-weighted MRIs. Radiomic data from 81 individuals with confirmed rectal cancer, of which 35 received nCRT, was analysed.

The combination of Fisher score selector and Decision tree classifier achieved test scores measured in area under the receiver operator curve (AUC) of $62.2 \pm 5.9\%$ and $73.0 \pm 10.8\%$ when predicting PFS for all patients and the nCRT cohort, respectively. Both models selected first-order and texture features only. Across models, the small area high gray level emphasis texture feature appeared to be of relevance in predicting PFS. Prediction of TRG and ypT was achieved with test scores of about 80% and 90% AUC, respectively. Overall, texture and first-order features were well represented among those selected.

Values for test standard deviation were above 10% for a majority of models, and above 20% for some models considering the nCRT cohort. The small size and high dimensionality of this cohort may cause issues like over-fitting and poor ability to generalize.

As a preliminary investigation, reproducibility of obtained results was low. This may be influenced by the presence of correlated features. Evaluating correlation and removing redundant features accordingly are likely to render results more reliable and allow for features with predictive and prognostic value to more readily be identified.

SAMMENDRAG

Med et ønske om å bidra til forskningen om hvorvidt radiomicsparametre fra MR-bilder kan være biomarkører for endetarmskreft, hadde denne oppgaven tre hovedmål. For det første, å etablere og optimalisere binære prediksjonsmodeller med progresjonsfri overlevelse (PFS), tumorskrumpning (TRG) og T-stadieinndeling etter kjemoradioterapi (ypT) som endepunkter. De to sistnevnte indikerer effekten av preoperativ kjemoradioterapi (nCRT). For det andre, å undersøke i hvilken grad parametre som beskriver bildetekstur har prognostisk og prediktiv verdi. For det tredje, å evaluere utvalgte resultaters reproduserbarhet i forhold til voxeldimensjon, tumorsegmentering og diskretisering av bildenes intensitetsnivåer.

Kombinasjoner av fire seleksjonsalgoritmer og seks klassifikasjonsalgoritmer ble vurdert. Parametre som beskriver tumorens form og tekstur, i tillegg til første ordens statistiske parametre, ble beregnet fra T2- og diffusjonsvektede MR bilder. Radiomicdata fra 81 pasienter med bekreftet endetarmskreft ble analysert. Av disse fikk 35 pasienter nCRT.

Seleksjonsalgoritmen Fisher score i kombinasjon med klassifikasjonsalgoritmen Decision tree oppnådde følgende testresultat, målt som area under the receiver operator curve (AUC): 62.2 ± 5.9% og 73.0 ± 10.8% for prediksjon av PFS for henholdsvis alle pasienter, og pasienter som fikk nCRT. Begge modellene selekterte kun teksturparametre og første ordens statistiske paramerte. Basert på resultater fra flere modeller var en parameter som vektlegger små, høyintensive bilderegioner tilsynelatende av relevans i prediksjon av PFS. Prediksjon av TRG og ypT resulterte i AUC-verdier omkring henholdsvis 80% og 90%. Samlet sett var tekstur og første ordens statistiske parametre godt representert blant de selekterte.

Standardavvikene for test-verdier var omkring 10% for majoriteten, og over 20% for noen, for modellene som predikerte kun på bakgrunn av data fra pasienter med nCRT. Dette datasettet var lite og hadde samtidig høy dimensjonalitet på grunn av antallet parametre, noe som kan resultere i en overtilpasning av modellen, og en redusert evne til å generalisere.

Basert på denne undersøkelsen viste resultatene lav reproduserbarhet. Korrelasjon mellom parametre kan være en sentral årsak. En evaluering av i hvilken grad slik korrelasjon forekommer, og ekskludering av overflødige parametre, kan potensielt bidra til mer troverdige resultater. Dette vil også kunne gjøre det enklere å identifisere pålitelige parametre med prognostisk og prediktiv verdi.

ACKNOWLEDGMENTS

Above all, I want to thank my supervisor, Kathrine Røe Redalen, for supporting, helping and inspiring me throughout my work with this thesis. She managed to maintain good communication and collaboration even when we had to work from home. Thank you for always taking the time to explain and discuss matters I found difficult.

The same applies to Franziska Knuth, who generously shared her expertise in the fields of MRI, image analysis and Python-based programming. I am grateful for all our discussions.

Furthermore, I would like to express my great appreciation to Ahmed Albuni for support on using his program and for many interesting discussions.

Thank you also to remaining members of the Medical Radiation Physics group at NTNU, in particular Eline Furu Skjelbred and René Winter, for great discussions, input and feedback.

Finally, I want to thank my friends at MTNANO for five extraordinary years, and my family for their ability to make me excited about life.

CONTENTS

IV APPENDIX

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

MRI	Magnetic resonance imaging/image
ROI	Region of interest
VOI	Volume of interest
GLCM	Grey level co-occurrence matrix
GLRLM	Grey level run length matrix
GLSZM	Grey level size zone matrix
NGTDM	Neighbouring gray tone difference matrix
GLDM	Grey level dependence matrix
T2WI	T2-weighted imaging/image
DWI	Diffusion-weighted imaging/image
PFS	Progression free survival
MR	Magnetic resonance
RF	Radio frequency
TE	Time to echo
TR	Time to repetition
ADC	Apparent diffusion coefficient
MRF	Mesorectal fascia

nCRT	Neoadjuvant chemoradiotherapy
ypT	Posttherapy pathological T-staging
TRG	Tumor regression grade
CRT	Chemoradiotherapy
AJCC	American Joint Committee on Cancer
LARC	Locally advanced rectal cancer
MAD	Mean absolute deviation
IMC	Informational measure of correlation
IDM	Inverse difference moment
ID	Inverse difference
MCC	Maximal correlation coefficient
IDMN	Inverse difference moment normalized
MI	Mutual information
ReF	ReliefF
VT	Variance threshold
FS	Fisher score
RR	Ridge regression
OLS	Ordinary least squares
LR	Logistic regression
SVC	Support vector classifier
DT	Decision tree
ET	Extremely randomized trees
GBDT	Gradient boosting decision tree
LGBM	Light gradient boosting machine
CV	Cross validation
ROC	Receiver operating characteristic curve
AUC	Area under the receiver operator curve
PR	Precision-recall
FOV	Field of view
DICOM	Digital Imaging and Communications in Medicine
NifTI	Neuroimaging Informatics Technology Initiative
RV	Response variable
pCR	Pathological complete response
PCA	Principal component analysis
T1WI	T1-weighted imaging/image

Part I

BACKGROUND

1 INTRODUCTION

In 2018, colorectal cancer had the second highest incidence rate in Norway as well as in Europe, and third highest worldwide [1]. In present-day (2019) rectal cancer staging, magnetic resonance imaging (MRI) is the recommended image modality [2].

In the field of oncology, there is growing evidence of and attention towards tumor heterogeneity [3]. Observed differences in how patients respond to treatment motivates development of more personalized treatment options [3]. Increased knowledge on and quantification of tumor heterogeneity is considered to be crucial in this regard [4]. Being able to predict each tumor's response to a particular therapy would allow for a more exact tailoring of dose, duration and overall intensity of the cancer treatment [3].

Both inter- and intratumor heterogeneity exist. This may be due to differences in tumor origin (e.g. cell types involved, location and environment), as well as in the pattern of mutations occurring during tumor progression and in response to treatment [4].

Images acquired using MRI or other modalities are obtained for nearly all cancer patients [5], often several times during the course of diagnosis, cancer staging and treatment [3]. There exists a conception of such images containing unexploited, potentially highly useful information [3]. Image texture, defined in Chapter 4, is being explored as a source to such information and may be a way of evaluating tumor heterogeneity [3]. Parameters relevant for describing image texture, as well as other tumor features, can be extracted from MRIs in a process referred to as radiomics [3].

RADIOMICS Radiomics is defined as the process in which quantitative and high-dimensional features are extracted from medical images and then used to recognize patterns in the data [5]. Together with qualitative features as derived from medical images by experts, the quantitative, mathematically based features may be referred to as imaging biomarkers [6]. As such, radiomic features might provide information on normal as well as pathogenic biological processes [6].

Information provided by the radiomics analysis may be combined with other available data, for example genomic and clinical, to uncover patterns and typically assist in some sort of decision support, like predicting response to treatment [5]. As such, the field of radiomics contributes to personalized medicine. In the following, some important aspects of a typical radiomic analysis are introduced.

Prior to feature extraction, a region or volume of interest (denoted ROI and VOI, respectively) must be defined and subsequently segmented [5]. The VOI is typically whole tumor(s) or specific parts of the tumor with different physiology, so-called habitats [5]. Segmentation is done either manually by radiologists, by the use of automatic segmentation software, or by using a combination of the two [5].

In this thesis, the following three classes of radiomic features will be considered: first-order statistical features, shape-based features, and texture features [7] [8]. The latter includes features derived from the grey level cooccurence matrix (GLCM), grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM), neighbouring grey tone difference matrix (NGTDM), as well as the grey level dependence matrix (GLDM). Chapter 4 includes the necessary background on this topic.

The number of features extracted in a radiomics analysis may be large, typically in the order of hundreds [5] [9] or even thousands [10] [11]. The latter is often true especially when features are derived from MRIs of various scan types [11].

This increasing amount of available features may lead to over-fitting, especially when the number of patients is small [12]. As such, there is need for a selection process from which only the most relevant features remain [5]. The topic of feature selection is addressed in Chapter 5.

Finally, the dataset represented by selected features combined with chosen endpoint(s) may be used to train machine learning-based predictive and prognostic models [10]. Data mining refers to the process of thoroughly exploring large datasets with the goal of discovering intricate relationships [13], and is commonly performed within the field of radiomics [5]. Machine learning is the topic of Chapter 5.

In order for prediction models built from radiomics studies to be of use in evaluation of new input data, reproducibility must be assessed [14]. Various factors might add variability and thereby affect the reproducibility or robustness of radiomic biomarkers [14]. In particular, such factors include image acquisition approach, voxel dimensions, VOI segmentation, and intensity discretization [14] [12].

PROJECT AIM The overall aim of this thesis was to explore how a radiomics analysis of T2- and diffusion-weighted MRIs (denoted T2WIs and DWIs, respectively) may provide predictive and prognostic value within the field of rectal cancer, with response to preoperative treatment and progression free survival (PFS) as endpoints. The approach was threefold. First, binary prediction models based on radiomic data were established and attempted optimized by evaluating various combinations of selectors and classifiers. Second, the degree to which texture features in particular showed predictive or prognostic value was investigated. And, finally, a preliminary evaluation of reproducibility was performed.

The patient cohort considered in this thesis consisted of 81 individuals, all diagnosed with rectal cancer.

NOTE: If a reference is given at the end of a paragraph, it is implied that the reference was used throughout that particular paragraph.

2 MAGNETIC RESONANCE IMAGING

In this Chapter, the fundamental principles of MRI are first outlined in brief, based mainly on information found in [15] and [16]. Then, in Section 2.3, DWI is explained.

2.1 FUNDAMENTAL PRINCIPLES

MRI takes advantage of the fact that the hydrogen proton has spin. By virtue of being a charge in motion the hydrogen proton will posses a magnetic dipole moment. When exposed to a strong, static magnetic field as provided by the magnetic resonance (MR) machine, the dipole moments of the hydrogen protons attempt to align with this external field. Due to phenomena within quantum mechanics they are not quite able to align. This gives rise to a precession movement of the hydrogen protons around the direction of the external field. The angular frequency is commonly referred to as the Larmor frequency, expressed as the precessing hydrogen proton's gyromagnetic ratio multiplied by the magnetic field strength [15]:

$$f = 42.58 \cdot \text{magnetic field strength} \tag{2.1}$$

in units of MHz. Angular Larmor frequency is obtained by multiplication of 2π [15].

In addition to inducing precession, the external magnetic field will cause the hydrogen protons to be in one of two possible states, *almost* parallel or *almost* anti-parallel relative to the external field. Depending on field strength and temperature, there will be a slight majority of protons in one state. In MRI, this is usually the former state. The hydrogen protons all precess at the same frequency but are on average completely out of phase. Accordingly, with majority of protons in the parallel state, there will be a net magnetization precisely in the direction of the external field [16].

An illustration of both the spin and precession movement of protons can be seen from Figure 2.1.

In the following, a brief description of the main steps involved in acquiring an MRI is included. Subsections 2.1.1 and 2.1.2 are based on information from [15].

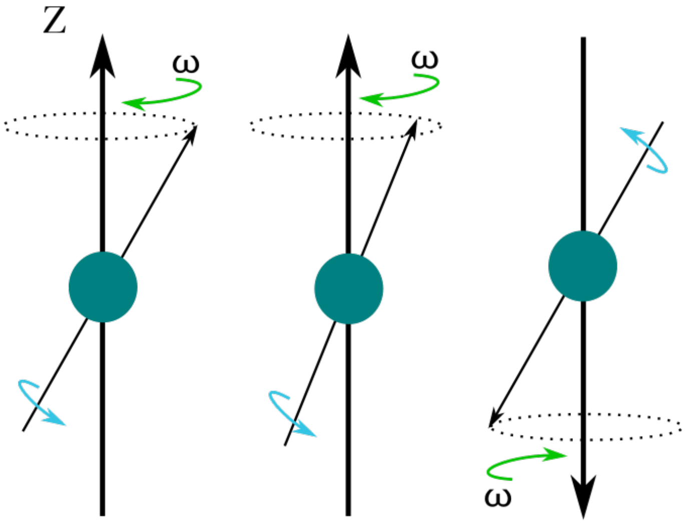


Figure 2.1: Protons precessing at the Larmor frequency ω, as evident by the green arrow, around the Z-direction, i.e. the direction of the external magnetic field $\mathbf{B_0}$. The blue arrow indicates spin. A majority of protons in the almost parallel state. Inspiration for this figure was found in [15], page 30.

2.1.1 *MRI sequences*

MR sequences are specifications of the steps involved in acquiring MRIs. Various approaches can be made based on the type of information the resulting MRI is expected to provide. Yet, four basic components are part of any standard MR sequence.

First, the previously mentioned static, external magnetic field, typically denoted $\mathbf{B_0}$, parallel to the z-axis. A common field strength is 1.5 Tesla.

Furthermore, a second, much weaker magnetic field is employed, namely the gradient field, giving rise to slight variations in the total magnetic field strength along the chosen gradient directions. This enables association of an individually detected signal with a specific position in the tissue being imaged.

Third, a radio frequency (RF) pulse is utilized. Prior to the pulse being emitted, the net magnetization of the hydrogen protons dipole moment was aligned parallel to $\mathbf{B_0}$ in the longitudinal direction, denoted $\mathbf{M_z}$. The component in the transverse direction, $\mathbf{M_{xy}}$, is zero. It is difficult to detect $\mathbf{M_z}$ particularly because it is not in motion and will therefore not induce current, and due to its magnitude being much smaller than that of $\mathbf{B_0}$.

Hence the RF, or excitation, pulse. The frequency of the pulse matches the precession frequency of the hydrogen protons, i.e. the Larmor frequency. Consequently, resonance will occur and thereby give rise to a short-lived transverse component. The RF pulse and subsequent resonance provide energy, causing the net magnetization to be "flipped" into the transverse plane. This means that the direction of the hydrogen protons spin axis or dipole moment are now positioned at a greater angle to the z-axis. In addition, they now precess with quite similar phases, resulting in a net magnetization in the transverse direction, $\mathbf{M_{xy}}$, according to vector addition. $\mathbf{M_{xy}}$ will rotate at the Larmor frequency about the direction of $\mathbf{B_0}$.

It is this transverse component of the net magnetization that is altered according to the gradient field: the applied gradient generates slight differences in the protons precession frequency, and subsequent dephasing and decrease in $\mathbf{M_{xy}}$. The extent to which $\mathbf{M_{xy}}$ is reduced can be linked to the position exposed to a gradient of matching magnitude.

The final step in the MRI sequence is signal detection. Current is induced in receiver coils due to the rotating $\mathbf{M_{xy}}$. Areas with different anatomy or function correspond to distinct signal intensities. Two main aspects in this regard are proton density and relaxation.

The former reflects the density of hydrogen protons able to interact as explained and give off signal. This will mainly be the case for hydrogen protons in water and fat. Relaxation is the net magnetization's return to its original state prior to being effected by the RF pulse. This process will begin right after emission of the RF pulse. It involves two concurrent phenomena: an increase in the longitudinal component and a decrease in the transverse component, referred to as T1 and T2 relaxation, respectively. T1 and T2 are both time constants for its associated relaxation process.

2.1.2 *T1 and T2*

Weakening of the transverse net magnetization component and increase of the longitudinal component will occur at a rate dependent on the hydrogen proton's environment. In T1 relaxation, the determining factor is whether or not the local magnetic field fluctuates at a frequency close to the Larmor frequency. In tissues were this is the case, the T1 relaxation will be rapid, corresponding to high-intensity signal.

T2 relaxation is based on dephasing of $\mathbf{M_{xy}}$. As mentioned, for there to be a net magnetization in the transverse plane, the spin axis must be in phase for a majority of the hydrogen protons. Accordingly, a dephasing is a weakening of the transverse component. Areas with high-frequency, low-amplitude magnetic field variations give rise to slow dephasing and corresponding T2 relaxation. This is the case for aqueous tissue and corresponds to high-intensity areas in the MR image. Conversely, rapid T2 relaxation is associated with low-frequency, high-amplitude magnetic field variations. In T2 relaxation, dephasing occurs solely due to impact on the hydrogen protons from nearby atoms and molecules. T2* relaxation refers to the situation in which other "external" effects like susceptibility of the different tissues as well as heterogeneity in the applied magnetic field are present.

An MRI can be T1-weighted or T2-weighted, implying that the sequence is constructed such that it is the T1 and T2 properties of the different tissues that are forming the basis of the image, respectively. T2WIs, as exemplified in Figure 2.2a, are considered in this thesis.

2.1.3 *Magnetic field gradients*

This Subsection is based on information provided by [16], mainly Chapter 4 and 7.

As mentioned, magnetic gradient fields are used in addition to the main magnetic field, $\mathbf{B_0}$. While $\mathbf{B_0}$ is of strength typically in the range of 1.5 T to 3 T, the gradient fields have units of mT. They are referred to as gradient fields because they give rise to linear variations in the total static, magnetic field strength. Gradients can be applied in either direction; x, y or in the direction of $\mathbf{B_0}$, z. This typically involves a subtraction or an addition to $\mathbf{B_0}$ based on the position along the corresponding axis. For instance, a gradient field in the y-direction, denoted $\mathbf{G_y}$, will add to or subtract from $\mathbf{B_0}$ according to distance along the y axis from the position at which $\mathbf{G_y} = 0$.

$\mathbf{G_x}$, $\mathbf{G_y}$ and $\mathbf{G_z}$ are induced by the use of gradient coils. These can be employed in combination, thereby enabling linear field variations in any direction and orientation.

The precession frequency of the hydrogen protons is directly proportional to the magnetic field strength according to Equation 2.1. Thus, at a position exposed to a positive gradient the protons will be precessing faster than at the position of zero gradient, and vice versa. The overall result is a dephasing.

Magnetic field gradients are utilized in slice selection, spatial encoding and acquiring gradient echos.

SLICE SELECTION Choosing the area to excite and form an MR image from is a crucial step in MRI. In addition to allowing for selection of direction and orientation, the gradient coils also enable determination of slice thickness.

Slice selection is achieved by applying the RF excitation pulse in combination with a gradient field. The RF pulse will have a narrow bandwidth in the sense that it only contains frequencies slightly higher and slightly lower than the frequency of the precessing protons in the absence of any gradient. A smaller bandwidth give rise to a a narrower area of excited hydrogen protons. The gradient field will alter the protons precession frequency according to position along the chosen direction(s). Resonance between the RF pulse and precession movement only occurs if the frequencies of the two matches. In other words, at positions where the gradient field has decreased or increased the precession frequency to such an extent that this is no longer the case, hydrogen protons will not be excited.

SPATIAL ENCODING Spatial encoding involves relating detected signal to the position from where it originates from. In order to associate each pixel or intensity value in the 2D matrix with a position, spatial encoding along two directions is required. So-called phase encoding could be carried out in both directions, however utilizing frequency encoding in one of the two speeds up the image acquisition process.

Phase encoding is achieved by the use of a gradient along one of the two directions to be spatially encoded, say, the y-direction. Initially, following the RF pulse, the excited protons will have quite similar phases. As explained, the gradient will cause a dephasing. The extent to which the protons dephase depends on the distance along the y-axis from where the gradient is zero.

A second, frequency-encode gradient is applied along the other direction, the x-axis. This is done at a later time than the phase-encode gradient, simultaneously as detecting the signal. As a result, the obtained signal will be spatially encoded in two directions.

2.1.4 *Gradient echo and spin echo sequences*

MR signals are commonly referred to as echos. The time interval between emission of the RF pulse and detection of the echo is defined as time to echo (TE). In order to acquire enough signal to form an image, the steps involved in the MR sequence are typically repeated several times. The time between start points of two succeeding repetitions is defined as time to repetition (TR) [15].

The two main types of MRI sequences are the gradient echo and spin echo sequence [15].

GRADIENT ECHO In a gradient echo sequence, the area being imaged is exposed to a negative gradient field shortly after emission of the excitation RF pulse. Then the gradient is reversed, resulting in a rephasing taking place. An echo is detected from this rephasing process as well as the ensuing "natural" dephasing [15].

SPIN ECHO Spin echo involves the use of a second RF pulse, referred to as the refocusing pulse. Optimally, this results in an elimination of the static, external effects, which is favorable both due to avoidance of artifacts potentially created by these effects, as well as enhanced signal intensity [15].

2.2 IMAGE FORMATION

This Section is based on information provided by [16], mainly Chapter 4 and 7. One exception exists, at which the reference is given.

Simply put, the analogue signal as detected by the receiver coils is sampled to fit a matrix of the same dimension as the final image, defined as the raw data matrix. The signal is now digitized. However, this raw data matrix consists of values in k-space. Accordingly, reconstruction of the raw data into the final image involves Fourier transformation.

MR images are made up of pixels. The image size is determined by the pixel matrix. Some commonly used dimensions are 256 x 128 and 512 x 256. Each position in the matrix corresponds to a pixel represented by a number referring to the intensity of the signal at this position.

Each pixel corresponds to the amount of signal originating from a specific element of the area being imaged, a so-called voxel. The delineated area that is imaged is in the form of a 3D slice. The thickness of each voxel corresponds to the thickness of the slice. The surface area of each voxel is given by the field of view (often square) divided by the number of pixels/voxels in each direction as defined by the matrix dimension.

The value of each pixel lies within a range determined by the bits (denoted b in the relation below) of the image. The different values, reflecting the intensity at each position, are referred to as grey-values. Normally, black corresponds to 0 and the maximum grey-level value to white. The number of different grey-levels are calculated as follows: $2^b - 1$. MRIs are often of 12 bits [17].

2.3 DIFFUSION WEIGHTED MRI

In DWI, the mobility of hydrogen protons can be observed [18]. This is useful in particular because many pathological events like inflammation, or the presence of a tumor [19], involve changes that in turn may alter the degree to which water protons are free to diffuse [18]. For comparison, diffusive motion is completely unrestricted in pure water [20].

Magnetic gradient fields are employed to achieve image contrast that is diffusion-weighted [16]. The b-value parameter, depending on magnitude and duration of the gradient field, reflects the degree to which obtained contrast is determined by diffusion [16]. A DWI can be seen from Figure 2.2b.

DWI is achieved by using a minimum of two different b-values, allowing for calculation of the apparent diffusion coefficient (ADC) as belonging to each tissue or area in the image [19].

For a given b-value, signal strength from a tissue with diffusion coefficient D is given by [16]

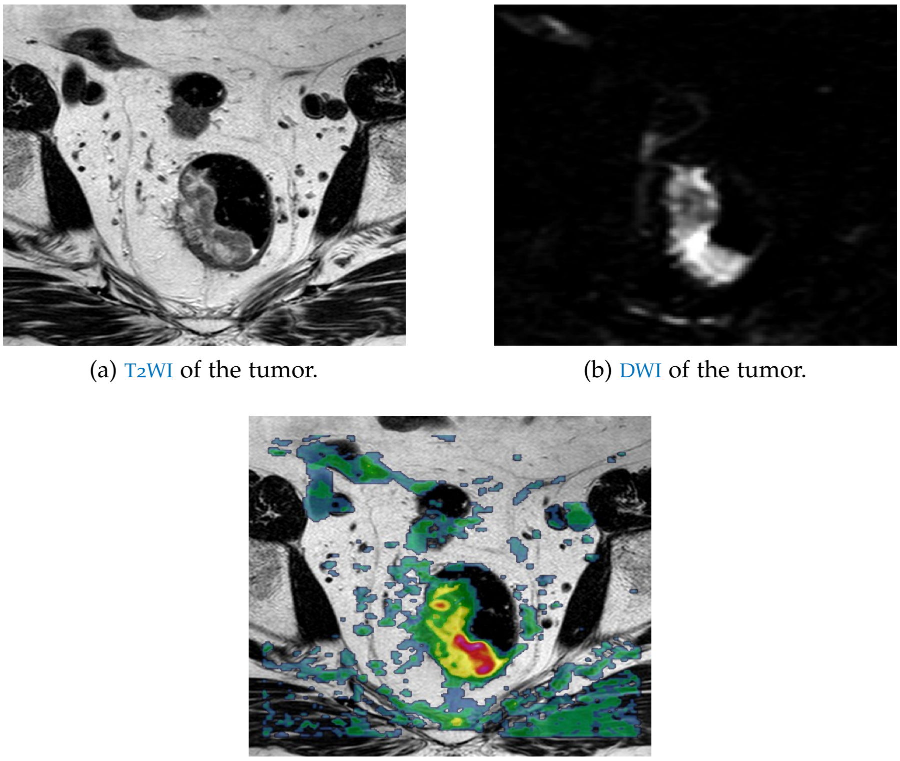
$$S(b) = S(0)e^{-bD}, \tag{2.2}$$

where $S(0)$ is the intensity with no diffusion-weighting [19]. Then, ADC associated with each voxel may be calculated as [19]

$$\text{ADC} = \frac{1}{b} \cdot ln \frac{S(0)}{S(b)} \tag{2.3}$$

Values for all voxels are typically presented in an ADC-map [19]. An ADC-map on top of a T2WI can be seen from Figure 2.2c.

DWIs and ADC-maps are used in correlation to get optimal information, for example in characterization of a lesion as either benign or malignant: the former appear dark (low intensity) in DWIs with high b-value and bright (high intensity) in the corresponding ADC-map, while the situation is reversed for malign lesions [19].



(a) T2WI of the tumor.

(b) DWI of the tumor.

(c) ADC-map on top of the T2WI of the tumor.

Figure 2.2: T2WI (a), DWI (b) and ADC-map on top of the T2WI (c), all showing the tumor of a patient included in the Hypoxia-mediated Rectal Cancer Aggressiveness (OxyTarget) study [21].

3 RECTAL CANCER AND MRI

Colorectal cancer include both colon and rectal cancer. Of all incidences of colorectal cancer, about $30 - 35\%$ is rectal cancer (2019) [10].

The rectum is the 12 cm to 18 cm long final part of the large intestine. In contrast to the colon, the rectum is surrounded by a layer of muscle, embedded in its walls. Some of the most important layers enclosing the rectal tube, from inner to outer, are the mucosa, the submucosa, the muscularis propria, and finally the mesorectum, bordered by the mesorectal fascia (MRF) [2]. The mesorectum may be of relevance in the development and spreading of rectal cancer due to the number of lymph nodes found here, typically between 5 and 20. Such lymph nodes may be targets in radiation therapy. A tumor is defined as locally advanced if it protrudes the MRF [22].

Cancer staging is the process of determining the degree to which the primary tumor has developed and spread [23]. The TNM standard [24], describing the tumor, regional lymph nodes, and degree of metastasis, respectively, is frequently used. The former is concerned with how far into the rectal wall the tumor extends. T1 and T2 refer to tumors having spread to the submucosa and muscularis propria, respectively. A T3 tumor extends into the mesorectum, while a T4 tumor penetrates the rectums entire cross-section and is thereby defined as locally advanced. T4b is descriptive of the tumor being in contact with nearby organs. Furthermore, N0 is indicative of no detected metastasis in nearby, or regional, lymph nodes. N1, N2 and the subdivisions within these stages correspond to increasing numbers of lymph nodes showing presence of metastasis. M1a and M1b denotes the case of metastasis to one or more distant organ(s) or position(s), respectively [2].

Based on stage, treatment of rectal cancer is normally attempted through surgery. So-called total mesorectal excision (resection of all layers enclosing the rectal canal until and including the mesorectum) is often performed [2]. Most times, surgery is done in combination with some form of preoperative therapy. This is referred to as neoadjuvant therapy and may involve radiation, chemotherapy, or a combination of both. Such preoperative measures are taken with the goal of reducing tumor size as well as enhancing possibility of a successful tumor removal during surgery [23].

In Europe, the following selected definitions and treatment standards prevails, according to [2]. In addition to a few other factors, a tumor staged as T1-T3, N0 and M0 is defined as low risk and therapy will consist of surgery only. High risk disease include T4-tumors and tumors with spreading to lymph nodes, i.e. locally advanced tumors.

In such cases, neoadjuvant chemoradiotherapy (nCRT) would typically be performed prior to surgery.

3.1 MRI IN RECTAL CANCER STAGING, TREATMENT PLANNING AND SURVEILLANCE

This section is based on information found in [2] unless otherwise stated.

MRI is the recommended image modality for rectal cancer staging. T2WIs are frequently used, with a narrow field of view and slice thickness below 3 mm. MRI is typically used in both primary staging and restaging, as well as in evaluation of local recurrence.

3.1.1 *Primary staging*

Primary staging is done prior to surgery and involves obtaining a detailed understanding of the particular case so that further actions can be carried out accordingly. Features related to high degree of tumor development, like invasion of the mesorectal vasculature and protrusion of the MRF, may be identified. Both location and morphology of the tumor can be described in great detail. An important descriptor is distance of the tumor to the anal verge. Tumors at a distance greater than 15 cm are defined as colon cancer. Furthermore, the extent to which the tumor protrudes the various tissues of the rectal wall is determined. This is crucial in T-staging.

The degree to which lymph nodes are involved, and thereby the N-stage, can be determined. MRI provides higher accuracy in T-staging than in N-staging. Nevertheless, useful insight regarding regional lymph nodes can be obtained from MRIs. In the case of tumors defined as T4, the organs and structures possibly invaded may be determined.

Based on features like the above mentioned, primary staging enables determination of whether or not neoadjuvant therapy is necessary, and is often crucial in planning of surgery.

3.1.2 *Restaging and evaluation of local recurrence*

Primary staging is also referred to as clinical staging, denoted cTNM. In pathological staging, denoted pTNM, information from analysing the resected tumor is obtained. Finally, posttherapy staging, denoted yTNM, consider the effect of preoperative treatment, and is also used when radiation- and/or chemotherapy is the only treatment [24].

In this thesis, both posttherapy pathological staging, denoted ypT, and Tumor Regression Grade (TRG), as explained next, are used as measures of the degree to which preoperative CRT was successful.

The TRG system exist in various forms, some ranging from 1 to 5 [2], others from 0 to 3 [25], where the lower number is indicative of complete removal and the higher of no or poor response. The latter system is called the 2010 American Joint Committee on Cancer (AJCC) TRG system [25], and will be used to assess response to nCRT in this thesis.

3.2 PREDICTING RESPONSE TO PREOPERATIVE CRT

In the field of rectal cancer, several studies on predicting response to nCRT in particular have been conducted. As mentioned, patients with locally advanced rectal cancer, LARC for short, is normally treated with nCRT prior to surgery. How successful this preoperative treatment is for LARC patients varies greatly, however [9].

To highlight the extremas: some experience complete tumor removal (about 15-27 % of patients receiving nCRT [26], [27]), while others are seemingly unaffected by the CRT [9].

From this it is clear that predicting response to nCRT could ideally limit the use of surgery and nCRT, respectively, to the cases in which it is actually of use. In particular, being able to select the patients experiencing complete tumor removal would be beneficial, as these patients might not need surgery at all [27].

4 IMAGE TEXTURE AND RADIOMIC FEATURES

Various definitions of image texture exists, each having a slightly different focus or generality. Some make higher demands as to what can be described as image texture than others.

A general definition is texture simply being spatial heterogeneity in pixel intensity [3]. While some include in their definition of texture features like mean, median and max/min values of intensity from a region of voxels [3] [17], others are more consistent with the idea that texture describes relationships between two or more voxels [5]. These two groups of features are referred to as first- and second-order statistical features, respectively [5]. Higher-order features are typically obtained by the means of filters or transforms and describe patterns in the VOI [5].

More specifically, the author of [28] defines texture in medical images as intrinsic heterogeneity or diversity at smaller scales than that of the object or structure being imaged. Such variations must be intrinsic in the sense that they exist due to actual heterogeneity or variation in the imaged structures, not merely caused by the instrument acquiring the image, such as noise.

The term texture will in this thesis apply to both second- and higher-order statistical features.

INTENSITY BINNING Discretisation or binning of the image intensities is commonly done prior to calculation of first-order and texture features, achieved according to either a defined bin width or, alternatively, a defined number of bins [6]. The latter will alter the contrast between two images, while a fixed bin width influence the image coarseness [6].

In this thesis, discretisation is done by defining a bin width, denoted w_b. Accordingly, the intensity of each voxel, X, is calculated as [6]

$$X_d = \frac{X - X_{min}}{w_b} + 1 \tag{4.1}$$

RADIOMIC FEATURES The remains of this chapter include theory on relevant radiomic features from each of the three classes; shape-based features, first-order statistical features and various texture features.

NOTE: All definitions presented in this chapter are based on the PyRadiomics package documentation [7] [8]. The same is true for associated descriptions and elaborations, unless otherwise stated.

All first-order statistical and texture features will be calculated from grey level values of voxels residing within the region of interest (ROI).

4.1 SHAPE FEATURES

Shape features describe the size and shape of the ROI. They are based on the binary mask used for segmentation of the ROI and thus independent of the intensity distribution within this region. Shape features may be three-dimensional or two-dimensional.

A network or mesh is created from the ROI to allow for calculation of shape features. Simply put, this mesh is made up by triangles where each triangle is defined by three adjacent points. The edge of the mesh consists of triangles defined by points positioned at the midpoint of the border between a voxel inside the ROI and a voxel excluded from the ROI. Note that the principal components defined below are obtained from the tumor directly, not from the mesh.

The shape features considered in this thesis are presented in Table 4.1, with corresponding descriptions for some features below.

Define:

- N_v = number of voxels in the ROI.
- N_f = number of triangles in the mesh.
- a_i, b_i, c_i = points constituting a triangle.
- O = point defined as image origin.
- λ_{major}, λ_{minor}, λ_{least} = greatest, next-greatest, and smallest principal component, respectively. Obtained by principal component analysis.

Sphericity quantifies the degree to which the tumor is shaped as a sphere. It ranges from 0 to 1, the latter suggesting a fully spherical tumor.

Obtained from a principal component, the **major**, **minor** and **least** axis length measure the greatest, next-greatest, and smallest axis length, respectively. The **elongation** feature ranges between 1, as indicative of a spherical, i.e. non-elongated tumor shape, and 0, as indicative of a 2D line, i.e. fully elongated. Similarly, **flatness** equals 1 when the tumor is spherical and non-flat, and 1 when the opposite is true.

FEATURES	DEFINITION
Mesh Volume, V	$\sum_{i=1}^{N_f} \frac{Oa_i \cdot (Ob_i \times Oc_i)}{6}$
Voxel Volume, V_{voxel}	$\sum_{k=1}^{N_v} V_k$
Surface Area, A	$\frac{1}{2} \sum_{i=1}^{N_f} \lvert a_i b_i \times a_i c_i \rvert$
Surface to Volume ratio	$\frac{A}{V}$
Sphericity	$\frac{\sqrt[3]{36\pi V^2}}{A}$
Major Axis Length	$4\sqrt{\lambda_{major}}$
Minor Axis Length	$4\sqrt{\lambda_{minor}}$
Least Axis Length	$4\sqrt{\lambda_{least}}$
Elongation	$\sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$
Flatness	$\sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$
Maximum 3D Diameter	
Maximum 2D Diameter, Slice	
Maximum 2D Diameter, Column	
Maximum 2D Diameter, Row	

Table 4.1: Features describing shape. Definitions are according to the PyRadiomics package documentation [7] [8].

4.2 FIRST-ORDER STATISTICAL FEATURES

The first-order statistical features can be obtained in a histogram analysis, in which the histogram displays the number of pixels in the whole image or ROI at each intensity level [4]. The features describe such intensity distributions without taking spatial relationships between voxels into account, thereby being first-order [29].

The first-order statistical features considered in this thesis are presented in Table 4.2, with corresponding descriptions below.

Define:

- $\mathbf{X}$ is a set of N_p voxels, all positioned within the ROI.
- $\mathbf{P}$(i) is a first-order histogram containing N_g intensity bins, and normalized as p(i) = $\frac{\mathbf{P}(i)}{N_p}$.
- c = constant defined if $\mathbf{X}$ contains negative values, resulting in positive intensities only.
- ϵ = some small, positive constant.

Percentiles describe the intensity value that a given percentage of all pixels are below, or less than [4]. The percentiles considered in this thesis are 10 % and 90 %, referred to as P_{10} and P_{90}, respectively [6].

FEATURES	DEFINITION
Maximum intensity	$max(\mathbf{X})$
Minimum intensity	$min(\mathbf{X})$
Range	$max(\mathbf{X}) - min(\mathbf{X})$
Median intensity	
Mean intensity	$\frac{1}{N_p}\sum_{i=1}^{N_p}\mathbf{X}(i)$
10th percentile	$\mathbf{P}_{10}$
90th percentile	$\mathbf{P}_{90}$
Interquartile range	$\mathbf{P}_{75} - \mathbf{P}_{25}$
Energy	$\sum_{i=1}^{N_p}(\mathbf{X}(i) + c)^2$
Total energy	$V_{voxel}\sum_{i=1}^{N_p}(\mathbf{X}(i) + c)^2$
Entropy	$-\sum_{i=1}^{N_g} p(i) log_2(p(i) + \epsilon)$
Variance	$\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^2$
Mean absolute deviation (MAD)	$\frac{1}{N_p}\sum_{i=1}^{N_p}\lvert\mathbf{X}(i) - \bar{X}\rvert$
Robust MAD	$\frac{1}{N_{10-90}}\sum_{i=1}^{N_{10-90}}\lvert\mathbf{X}_{10-90}(i) - \bar{X}_{10-90}\rvert$
Root mean square	$\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) + c)^2}$
Uniformity	$\sum_{i=1}^{N_g} p(i)^2$
Skewness	$\frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^3}{(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2})^3}$
Kurtosis	$\frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^4}{(\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i)-\bar{X})^2)^2}$

Table 4.2: First-order statistical features. Definitions are according to the PyRadiomics package documentation [7] [8].

High **energy** corresponds to a large number of high-intensity voxels. The **total energy** feature is obtained as the product of energy and voxel volume.

The **entropy** feature represents the degree to which an intensity histogram is irregular [4], or random [8].

The **variance** of **X** is the degree to which intensities are spread out from the mean intensity. Similar to variance, the **mean absolute deviation (MAD)** and the **robust MAD** are both measures of the degree to which intensities are spread out about the mean. In the calculation of robust MAD, only pixels within the range defined by $\mathbf{P}_{10}$ as lower limit and $\mathbf{P}_{90}$ as upper limit are considered.

Similar to energy, the **root mean square** reflects the magnitude of the total intensity.

Uniformity quantifies the extent to which the image consists of voxels with similar intensity values, in other words the homogeneity of the histogram.

Skewness describes the degree to which the intensity values are distributed in an asymmetric fashion about the mean intensity. A positively skewed distribution is characterized as being flat and spread-out for values to the right of the mean, while the left side often contains a broader peak [4], and vice versa.

Kurtosis reflects the shape of the intensity histogram [4]. Simply put, it describes whether the distribution is peaked around the mean intensity value (low kurtosis) or less peaked and more broad (high kurtosis).

4.3 Texture Features

Evaluating image texture may be viewed as a second order histogram analysis [4]. As mentioned, texture features take into account the spatial relationship between voxels [29]. From each of the matrices described below, various features can be derived. All features are calculated from voxels inside the ROI.

Define, for each ROI:

- N_g = number of intensity bins.
- $p(i, j)$ = the normalized matrix.
- N_p = number of voxels.
- ϵ = random, positive number of magnitude $\sim 10^{-16}$

Additional parameters are defined for each particular matrix.

4.3.1 *GLCM Features*

The grey level co-occurrence matrix (GLCM), denoted $\mathbf{P}(i, j)$, shows, for a given spacing between the two pixels as well as direction in the image, the number of pixel pairs found for each possible combination of intensity values [17].

The number of possible pixel pair combinations and thus GLCM size, is N_g x N_g. For a given spacing, δ, and angle, θ, each matrix element (i,j) is a value equal to the frequency at which this particular intensity combination i and j appears in the image.

The GLCM features considered in this thesis are presented in Table 4.3, with corresponding descriptions below.

Define, for each ROI:

- $p(i, j) = \frac{\mathbf{P}(i,j)}{\sum \mathbf{P}(i,j)}$
- $p_x(i) = \sum_{j=1}^{N_g} P(i,j)$ and $p_y(j) = \sum_{i=1}^{N_g} P(i,j)$ are marginal row and column probabilities, respectively.
- μ_x and σ_x = mean intensity and standard deviation of p_x, respectively.
- μ_y and σ_y = mean intensity and standard deviation of p_y, respectively.
- $HXY =$ joint entropy

The **autocorrelation** feature quantifies image coarseness. High **joint energy** indicates the presence of more homogeneous patterns, while high **joint entropy** is indicative of high variability among the pixel pairs calculated from the image.

Cluster prominence also quantifies variability, with a lower value indicating that values in the GLCM are closer to the mean, i.e. little variation. The same is true for the **cluster shade** feature. **Cluster tendency** quantifies the degree to which groups of voxels with similar intensities appear.

High **contrast** indicates larger differences in gray-levels for voxels in close vicinity of each other. The **correlation** feature quantifies correlation between the intensity value and the voxel.

Difference average evaluates the relation between pixel pairs of equal and dissimilar gray-level value, respectively. **Difference entropy** quantifies the degree to which the differences in intensities that appear close, vary. Finally, **difference variance** weights pixel pairs in which the intensity between the two voxels deviate from the mean difference, higher.

Informational measure of correlation (IMC) 1 and **2** both quantify the degree to which probability distributions for i and j are found to correlate. The **inverse difference moment (IDM)** feature quantifies homogeneity among nearby pixel pairs. The same is true for **inverse difference (ID)**. Similarly to IMC, the **maximal correlation coefficient (MCC)** describes texture complexity.

FEATURES	DEFINITION
Autocorrelation	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)ij$
Joint average	$\mu_x = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)i$
Joint energy	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p(i,j))^2$
Joint entropy	$-\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) log_2(p(i,j)+\varepsilon)$
Cluster prominence	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i+j-\mu_x-\mu_y)^4 p(i,j)$
Cluster shade	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i+j-\mu_x-\mu_y)^3 p(i,j)$
Cluster tendency	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i+j-\mu_x-\mu_y)^2 p(i,j)$
Contrast	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-j)^2 p(i,j)$
Correlation	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)ij-\mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}$
Difference average	$\sum_{k=0}^{N_g-1} kp_{x-y}(k)$
Difference entropy	$\sum_{k=0}^{N_g-1} p_{x-y}(k) log_2(p_{x-y}(k)+\varepsilon)$
Difference variance	$\sum_{k=0}^{N_g-1} (k-DA)^2 p_{x-y}(k)$
IMC 1	$\frac{HXY-HXY1}{max\{HX,HY\}}$
IMC 2	$\sqrt{1-exp(-2(HXY2-HXY)}$
IDM	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k^2}$
IDM normalized	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\frac{k^2}{N_g^2}}$
ID	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k}$
ID normalized	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\frac{k}{N_g}}$
MCC	$\sum_{k=0}^{N_g} \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$
Inverse variance	$\sum_{k=1}^{N_g-1} \frac{p_{x-y}(k)}{k^2}$
Maximum probability	$max(p(i,j))$
Sum average	$\sum_{k=2}^{2N_g} p_{x+y}(k)k$
Sum entropy	$\sum_{k=2}^{2N_g} p_{x+y}(k) log_2(p_{x+y}(k)+\varepsilon)$
Sum of squares	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu_x)^2 p(i,j)$

Table 4.3: Features derived from the GLCM. Definitions according to the PyRadiomics package documentation [7] [8].

Maximum probability considers the intensity value pair that appears most often in the GLCM, and equals the frequency of this occurrence.

Similar to the difference variance feature, **sum average** quantifies the relation between pixel pairs of lower and higher gray-levels, respectively. The **sum entropy** feature summarizes the differences in gray-level values between pairs in a neighbourhood.

Finally, the **sum of squares** feature quantify the deviation of gray-level value of neighbouring pixel pairs from the mean.

4.3.2 *GLSZM Features*

The grey level size zone matrix (GLSZM), denoted $\mathbf{P}(i,j)$, shows, for each intensity value, the number of adjacent or so-called connected voxels having the same intensity, thereby constituting a zone. The vertical and horizontal matrix dimension represents the grey level values and number of voxels in each zone, respectively. As such, each matrix element (i,j) corresponds to the number of zones with grey level i and j number of voxels. In a 2D image, each voxel is connected to 8 other voxels.

The GLSZM features considered in this thesis are presented in Table 4.4, with corresponding descriptions below.

Define, for each ROI:

- N_p = number of voxels.
- N_z = number of zones.
- N_s = number of different zone sizes.
- $p(i,j) = \frac{\mathbf{P}(i,j)}{N_z}$

The **small** and **large area emphasis** features quantify, respectively, the degree to which zones in the image are smaller, corresponding to finer texture, or larger, corresponding to courser texture.

High **gray level non-uniformity** indicates larger variance or heterogeneity in the different intensity values of the image. Similarly, a high **size-zone non-uniformity** indicates larger variance in the volumes of the different size zones of the image.

The **zone percentage** feature, defined as the number of zones relative to the number of voxels, is also a measure of texture coarseness.

Similarly to gray level and size-zone non-uniformity, the **gray level** and **zone variance** features quantify the variability in intensity values and volumes of the zones, respecitvely. In the corresponding equations, μ is defined as mean zone intensity and volume, respectively.

Zone entropy quantifies the degree to which the volumes and intensity values of the zones appear randomly, or more structured in a pattern.

FEATURES	DEFINITION
Small area emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)}{j^2}}{N_z}$
Large area emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\mathbf{P}(i,j)j^2}{N_z}$
Gray level non-uniformity	$\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_s}\mathbf{P}(i,j))^2}{N_z}$
Gray level non-uniformity normalized	$\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_s}\mathbf{P}(i,j))^2}{N_z^2}$
Size-zone non-uniformity	$\frac{\sum_{j=1}^{N_g}(\sum_{i=1}^{N_s}\mathbf{P}(i,j))^2}{N_z}$
Size-Zone non-uniformity Normalized	$\frac{\sum_{j=1}^{N_g}(\sum_{i=1}^{N_s}\mathbf{P}(i,j))^2}{N_z^2}$
Zone Percentage	$\frac{N_z}{N_p}$
Gray level variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}p(i,j)(i-\mu)^2$
Zone variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}p(i,j)(i-\mu)^2$
Zone entropy	$-\sum_{i=1}^{n_g}\sum_{j=1}^{N_s}p(i,j)log_2(p(i,j)+\epsilon)$
Low gray level zone emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$
High gray level zone emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\mathbf{P}(i,j)i^2}{N_z}$
Small area low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)}{i^2j^2}}{N_z}$
Small area high gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)i^2}{j^2}}{N_z}$
Large area low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)j^2}{i^2}}{N_z}$
Large area high gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\mathbf{P}(i,j)i^2j^2}{N_z}$

Table 4.4: Features derived from the GLSZM. Definitions according to the PyRadiomics package documentation [7] [8].

The information quantified by remaining features is given from the name. Each feature equals the distribution of all zones with attributes according to its name.

4.3.3 *GLRLM Features*

The grey level run length matrix (GLRLM), denoted $\mathbf{P}(i, j|\theta)$, contains in a similar manner as the GLSZM, for each intensity value, the length defined in number of pixels along the direction of θ having the same intensity. The vertical and horizontal matrix dimension represents the grey level values and the length of each run, respectively.

The GLRLM features considered in this thesis are presented in Table 4.1, with corresponding descriptions below.

Define, for each ROI:

- N_p = number of voxels.
- N_r = number of run lengths.
- $N_r(\theta)$ = number of runs along the direction of θ
- $p(i, j|\theta) = \frac{\mathbf{P}(i,j|\theta)}{N_r(\theta)}$

The analogy to features derived from the GLSZM is apparent, now with the focus being run length instead of zone size.

The **short** and **long run emphasis** features quantify the degree to which runs are shorter or longer, as corresponding to finer or coarser texture, respectively.

Gray level non-uniformity is defined similarly to when this feature is extracted from the GLSZM as outlined above, now calculated based on the intensity levels of the runs. Likewise, **run length non-uniformity** quantifies the degree to which lengths of the runs calculated from the VOI vary.

Again, similarly to the zone percentage feature from the GLSZM, **run percentage** is another measure for texture coarseness. Furthermore, **gray level** and **run variance** describes the variability in intensity and length of the runs, respectively, similarly to the two non-uniformity features. In the corresponding equations, μ is defined as mean run intensity and length, respectively.

Run entropy quantifies randomness both with respect to length and intensity of runs.

The information quantified by remaining features is given from the name. Each feature equals the distribution of all runs with attributes according to its name.

FEATURES	DEFINITION
Short run emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j\vert\theta)}{j^2}}{N_r(\theta)}$
Long run emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j\vert\theta)j^2}{N_r(\theta)}$
Gray level non-uniformity	$\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_r}\mathbf{P}(i,j\vert\theta))^2}{N_r(\theta)}$
Gray level non-uniformity normalized	$\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_r}\mathbf{P}(i,j\vert\theta))^2}{N_r(\theta)^2}$
Run length non-uniformity	$\frac{\sum_{j=1}^{N_r}(\sum_{i=1}^{N_g}\mathbf{P}(i,j\vert\theta))^2}{N_r(\theta)}$
Run length non-uniformity normalized	$\frac{\sum_{j=1}^{N_r}(\sum_{i=1}^{N_g}\mathbf{P}(i,j\vert\theta))^2}{N_r(\theta)^2}$
Run percentage	$\frac{N_r(\theta)}{N_p}$
Gray level variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,\ j\vert\theta)(j-\mu)^2$
Run variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,\ j\vert\theta)(j-\mu)^2$
Run entropy	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,\ j\vert\theta)log_2(p(i,\ j\vert\theta)+\epsilon)$
Low gray level run emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j)}{i^2}}{N_r(\theta)}$
High gray level run emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j)i^2}{N_r(\theta)}$
Short run low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j)}{i^2j^2}}{N_r(\theta)}$
Short run high gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j)i^2}{j^2}}{N_r(\theta)}$
Long run low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j)j^2}{i^2}}{N_r(\theta)}$
Long run high gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j)i^2j^2}{N_r(\theta)}$

Table 4.5: Features derived from the GLRLM. Definitions according to the PyRadiomics package documentation [7] [8].

4.3.4 *NGTDM Features*

The neighbouring gray tone difference matrix (NGTDM) contains, for each intensity value i in the image, the following parameters: first, the number of voxels in the ROI having intensity i, denoted n_i. Second, the probability of intensity i, expressed as $p_i = \frac{n_i}{N_v}$, N_v being the total number of voxels in the ROI. Finally, the absolute difference between i and the average intensity of this particular voxels surroundings or neighbourhood, defined by a distance δ, summed up for each value of i. This sum is denoted s_i and defined as $\sum^{n_i} |i - \bar{A}_i|$.

The NGTDM features considered in this thesis are presented in Table 4.6, with corresponding descriptions below.

In addition to parameters defined above, let

- $N_{g,p}$ = the number of gray levels present in the ROI, i.e. having non-zero p_i
- $N_{v,p}$ = the number of voxels in the ROI with a minimum of one neighbor

FEATURES	DEFINITION
Coarseness	$\frac{1}{\sum_{i=1}^{N_g} p_i s_i}$
Contrast	$(\frac{1}{N_{g,p}(N_{g,p}-1)} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_i, p_j (i-j)^2)(\frac{1}{N_{v,p}} \sum_{i=1}^{N_g} s_i)$
Busyness	$\frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \lvert i p_i - j p_j \rvert}$
Complexity	$\frac{1}{N_{v,p}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \lvert i-j \rvert \frac{p_i s_i + p_j s_j}{p_i + p_j}$
Strength	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p_i + p_j)(i-j)^2}{\sum_{i=1}^{N_g} s_i}$

Table 4.6: Features derived from the NGTDM. Definitions according to the PyRadiomics package documentation [7] [8].

Coarseness quantifies how quickly gray levels are found to change, when moving from the voxel in question and out into its neighborhood. Low coarseness suggests rapid change and corresponding fine, heterogeneous texture.

High **contrast** is associated with high $N_{g,p}$ and low coarseness. Similarly to coarseness, **busyness** quantifies the difference between the gray-level value of each pixel and the surrounding neighborhood. Likewise, **complexity** measures the degree to which the ROI is heterogeneous with respect to intensity.

Finally, **strength** represents the degree to which larger characteristics in the ROI appear, as evident by less abrupt local gray-level changes, and higher coarseness.

4.3.5 *GLDM Features*

Two voxels i and j separated by a distance δ are defined as dependent if $|i - j| \leq \alpha$. The grey level dependence matrix (GLDM), denoted $\mathbf{P}(i, j)$, shows the frequency of each combination of intensity i and number of dependent voxels j occurring in the image. Each dependency can be viewed as a dependency zone.

The GLDM features considered in this thesis are presented in Table 4.7, with corresponding descriptions below.

Define, for each ROI:

- $N_d =$ the number of different dependencies with respect to size
- $N_z = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \mathbf{P}(i, j)$
- $\mu_i = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} ip(i, j)$
- $\mu_j = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} ip(i, j)$

The analogy to features derived from the GLSZM and GLRLM is apparent, now with the focus being voxel dependency instead of zone size and run length, respectively.

The **small** and **large dependence emphasis** quantify the degree to which smaller or larger dependency zones appear in the ROI. As for the GLSZM and GLRLM, these features are representative of texture coarseness. Larger dependencies suggest coarse texture.

Gray level non-uniformity is defined as for the GLSZM and GLRLM however now calculated from dependency intensities. High **dependence non-uniformity** is indicative of greater variance in the sizes of dependency zones. Similarly, **gray level** and **dependence variance** also quantify variability in intensity and dependency zone sizes, respectively.

Dependency entropy quantifies randomness with respect to both gray level value and dependency zone size.

The information quantified by remaining features is given from the name. Each feature equals the distribution of all dependencies with attributes according to its name.

FEATURES	DEFINITION
Small dependence emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)/i^2}{N_z}$
Large dependence emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)j^2}{N_z}$
Gray level non-uniformity	$\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_d}\mathbf{P}(i,j))^2}{N_z}$
Dependence non-uniformity	$\frac{\sum_{j=1}^{N_d}(\sum_{i=1}^{N_g}\mathbf{P}(i,j))^2}{N_z}$
Dependence non-uniformity normalized	$\frac{\sum_{j=1}^{N_d}(\sum_{i=1}^{N_g}\mathbf{P}(i,j))^2}{N_z^2}$
Gray level variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}p(i,j)(i-\mu_i)^2$
Dependence variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}p(i,j)(j-\mu_j)^2$
Dependence entropy	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}p(i,j)log_2(p(i,j)+\epsilon)$
Low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)/i^2}{N_z}$
High gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)i^2}{N_z}$
Small dependence low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2j^2}}{N_z}$
Small dependence high gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)i^2}{j^2}}{N_z}$
Large dependence low gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)j^2}{i^2}}{N_z}$
Large dependence high gray level emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)j^2i^2}{N_z}$

Table 4.7: Features derived from the GLDM. Definitions according to the PyRadiomics package documentation [7] [8].

5 MACHINE LEARNING

Two of the primary approaches to machine learning are supervised learning and unsupervised learning. They differ in the sense that the former involves providing the algorithm with both the dataset and the corresponding output, i.e. the dataset is labeled. In this context, the goal of the learning process is for an algorithm to predict the output when given just data and no right answer. Supervised learning is conducted mainly as either classification (organization of data into two or more distinct classes) or regression (model establishment founded on relationships in the data and subsequent prediction of an (often continuous) variable [30]). On the contrary, unsupervised learning involves algorithms recognizing tendencies and trends in unlabeled datasets, and on that basis organize the data into clusters [31].

Most algorithms used for predictive purposes have parameters that must be pre-defined [32], so-called hyper-parameters. Such parameters may impact the bias-variance tradeoff, i.e. whether the model is over-fit (high variance, low bias) or under-fit (low variance, high bias) [33]. Bias is here defined as the deviation of estimated performance from the true value [32]. The more a model is trained to fit the data, the more complex it will be, and correspondingly the higher the risk of over-fitting [34].

In this chapter, methods for feature selection and classification are first introduced. Then, in Section 5.3, approaches for evaluation of model performance are outlined.

5.1 FEATURE SELECTION

Feature selection can be viewed as a process in which a subset of relevant features is defined [35]. Here, relevant implies being useful in categorizing the data relative to some endpoint or class [36]. Feature selection is particularly important when dealing with large datasats, in which typically only a fraction of features are useful, the rest just adding to the complexity of the problem [36].

Various feature selection methods exist. Choosing the optimal method for a particular problem is crucial, notably to avoid discarding features that provide important information. Methods may be categorized based on the criteria defined as allowing for inclusion in the selected feature subset, or, alternatively, based on how the selection process relates to model construction. The latter distinguishes between filter, wrapper and embedded methods. Relief based selection methods, the topic of Subsection 5.1.4, are classified as filter methods [36].

Furthermore, selection methods may be categorized as either univariate or multivariate. In the former, each feature is evaluated separately, without taking the possibility of it depending on other features into account [35]. Mutual information (MI) is an example of a univariate method [37]. On the other hand, the Relief-based selection method ReliefF (ReF) considers feature dependency, thereby being a multivariate selection method [35].

5.1.1 *Variance threshold*

This subsection is based on information provided as documentation of the Scikit-Learn package (v. 0.22.2) [37] [38].

Selecting radiomic features based on the degree to which each feature is found to vary across samples is a straightforward approach, realized by defining a variance threshold (VT). As a minimum, this method excludes features found to be equal in all samples, in other words having $Var[X] = 0$.

5.1.2 *Mutual Information*

MI is an approach used for evaluating dependency between random variables [39]. Feature selection may be performed based on MI by evaluating the degree to which a feature is relevant in determining target class [40]. In other words, selecting features that best describe or provide information on the target class [40]. An MI score equal to zero indicates that the two variables are independent [39], and the feature is defined as not relevant in predicting class.

5.1.3 *Fisher Score*

A feature may be assigned a Fisher score (FS) based on an evaluation of the extent to which the value of this feature varies between classes. Higher-scoring features appear close in value for instances belonging to the same class, with a large difference for instances in different classes [41].

Each feature can be assigned a score according to the following expression [41]:

$$F = \frac{\sum_{i=1}^{c} n_i \left(\mu_i - \mu\right)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \tag{5.1}$$

Here, n_i is the number of instances belonging to each class, with c being the number of classes. Furthermore, μ_i and σ_i^2 are the mean and variance, respectively, of the particular feature in class i. Finally, μ is equal to the mean feature value across all classes.

5.1.4 *Relief and ReliefF*

Relief is a feature weighting method in which weights are assigned based on the features ability to distinguish between similar samples or instances [42]. To test this ability, the feature value corresponding to a given instance, I, is compared with two of its nearest neighbours, namely a near-hit (same binary outcome as I) and a near-miss (opposite binary outcome as I) [42]. A feature with values of I and near-hit being similar, while I and near-miss are different, is considered relevant [42].

The remains of this subsection is based on [36].

Features are ranked or weighted from -1 to $+1$, $+1$ being the best. For each randomly chosen instance, referred to as the target R_i, a comparison is made between R_i and pairs of neighbouring hit, H, and miss, M, instances, respectively. In each such comparison, all features are considered. The weighting of a feature A, initially being equal to zero, is updated as follows:

$$W[A] = W[A] - \frac{diff(A, R_i, H)}{m} + \frac{diff(A, R_i, M)}{m} \tag{5.2}$$

where m is the number of times a new R_i is chosen from the dataset, and the difference function, with I_2 equal to either H or M, defined as

$$diff(A, R_i, I_2) = \frac{|value(A, R_i) - value(A, I_2)|}{max(A) - min(A)} \tag{5.3}$$

The normalization is done in order to get weights ranging from 0 to 1, and then from -1 to 1 when dividing by m in Equation 5.2.

5.1.4.1 *ReliefF*

Several Relief-based selection methods or algorithms exists. ReF is the most commonly used, and has to a large extent replaced Relief. While this initial algorithm only can be used on binary classification problems, ReF is applicable for datasets with endpoints consisting of more than two classes.

Furthermore, instead of simply evaluating R_i on the basis of one hit and one miss, ReF enables evaluation according to a number of k nearest neighbors, in other words k hits and k misses, where k is specified by the user. A third main difference between Relief and ReF is that m is equal to the total number of instances in the training dataset.

5.2 CLASSIFICATION

5.2.1 *Ridge regression*

Ridge regression (RR) is a method for regularization of linear models, developed with the goal of addressing shortcomings associated with ordinary least squares (OLS) regression [43]. For a number N of training instances (x_i, y_i), OLS or linear regression involves estimating parameters $\theta = (\theta_0, \theta_1, ..., \theta_p)^T$ that minimize the following cost function [44]:

$$C_{linear} = \sum_{i=1}^{N}(y_i - f(x_i))^2, \tag{5.4}$$

with $f(x_i) = \theta_0 + \sum_{j=1}^{p} x_i\theta_j$ being the linear model. This approach is susceptible to collinearity, which occurs when several variables represent highly similar information, and resulting in high variance when introduced to new data [45]. I. e., the model is over-fit.

To reduce variability while still keeping bias to a minimum, RR involves a shrinkage of the coefficients [43]. This is achieved by adding a regularization term [46] to the cost function [44]:

$$C_{ridge} = C_{linear} + \alpha \sum_{j=1}^{p} \theta_j^2 \tag{5.5}$$

where the regularization parameter α can be optimized to best fit the problem at hand.

Ridge regression may be used for both binary and multiclass classification purposes, the former involving defining classes as $\{-1, 1\}$ and predict class according to the sign of the regression result [47].

5.2.2 *Logistic regression*

Logistic regression (LR) is frequently used for classification purposes, and, like RR, typically involve regularization [46].

Considering a binary classification problem with $y_i = \{0, 1\}$, the probabilities of y_i given some x_i is expressed as [46] [48]:

$$Pr(y_i = 1|x_i) = f(x_i, \theta) = \frac{1}{1 + exp(\theta_0 + \theta^T x_i)} \tag{5.6}$$

$$Pr(y_i = 0|x_i) = 1 - f(x_i, \theta) \tag{5.7}$$

with θ defined as in Subsection 5.2.1. The cost function, i.e. the negative log-likelihood [49], for this classification problem with N training instances may be expressed as [46] [48]:

$$C_{logistic} = f(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(f(x_i, \theta)) + (1 + y_i) log(1 - f(x_i, \theta)) \tag{5.8}$$

A regularization term is often added to Equation 5.8 typically in one of the following ways [46]:

$$C = f(\theta) + \beta \sum_{i=1}^{N} |\theta_i| \tag{5.9}$$

$$C = f(\theta) + \alpha \sum_{i=1}^{N} \theta_i^2 \tag{5.10}$$

referred to as ℓ_1- and ℓ_2-regularization, respectively. Note, from Eq. 5.5, that RR is linear regression with ℓ_2-regularization by definition.

5.2.3 *Support vector machine*

Support vector machine based methods involve mapping of the data set in such a way that classification according to two or more defined classes, typically separated by a hyperplane, is possible [50].

Lets first consider a binary classification problem with $y_i \in -1, 1$ and data that is linearly separable, implying that the hyperplane H is given as [51]:

$$H : w \cdot x - b = 0 \tag{5.11}$$

where x is the input dataset, w is a weight vector and b the bias. Furthermore, supporting hyperplanes, one on either side of H, may be defined on the basis of the following expressions for each instance x_i [51]:

$$\begin{cases} w \cdot x_i - b \geq 1, \ y_i = 1 \\ w \cdot x_i - b \leq -1, \ y_i = -1 \end{cases} \tag{5.12}$$

Instances appearing on ($w \cdot x_i - b = 1$ or $w \cdot x_i - b = -1$) or close to the supporting hyperplanes are referred to as support vectors. The separation, or margin, between the two classes corresponds to the distance between the two hyperplanes, expressed as $\frac{2}{||w||}$ [51].

Optimal values for w and b must satisfy Equation 5.12, rephrased as $y_i(w \cdot x_i - b) - 1 \geq 0$, for all x_i. Furthermore, the margin should be maximized. As such, optimal w and b satisfy both of the following [51]:

$$\begin{cases} \text{minimize } \frac{||w||^2}{2} \\ y_i(w \cdot x_i - b) \geq 1, \text{ for all } i = 1, 2, ..., m \end{cases} \tag{5.13}$$

m being the number of instances in the training set.

However, most datasets cannot be perfectly linearly separated. A possible approach then could be to allow for some minimal degree of misclassification, realised by including a slack variable $\xi_i \geq 0$ in Eq. 5.13, now expressed as [51]:

$$\begin{cases} \text{minimize } \frac{||w||^2}{2} + C\sum_{i=1}^{m} \xi_i, \; C > 0 \\ y_i(w \cdot x_i - b) \geq 1 - \xi_i, \text{ for all } i = 1, 2, ..., m \end{cases} \tag{5.14}$$

where C is the regularization parameter. Note that the higher the value of C, the less regularization is posed on the problem [52].

For some classification problems this is still not sufficient, however. The approach may then be to map the dataset into a higher-dimensional space by means of a nonlinear transformation. If the dataset has acquired sufficiently high dimensionality, it may now be linearly separable. The mapping is achieved by the use of a kernel, $K(x, y)$ [51].

The support vector classifier (SVC) algorithm is used in this thesis.

5.2.4 *Tree-Based Methods*

5.2.4.1 *Decision tree*

A Decision tree (DT) model is built on the basis of a hierarchy of descriptors with the goal of ending up at the most suitable "answer", or class, for each sample in the dataset [53].

Simply put, a common approach is as follows: 1) define the root node (i.e. the point at which the entire dataset is initially split into subsets [54]), 2) define a decision so that the split provides maximal information, and then repeat these steps for each of the subsequent decision nodes (i.e. sub-nodes being split into additional sub-nodes [54]) [53]. Leaf nodes terminate the hierarchy, thereby providing the class to which the particular instance belong [54].

Consider a binary classification problem with $y_i \in \{0, 1\}$. For each node m corresponding to a subset R_m containing N_m instances, I, the proportion of instances belonging to each class may be expressed as [55]:

$$p_{m,y_i} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i) \tag{5.15}$$

The N_m instances in node m are classified according to the value of y_i for which Equation 5.15 is maximized [55].

In order to define optimal variables and decisions at each split, node impurity is evaluated [56]. Node impurity quantifies the homogeneity or bias of the node, the optimal choice in variable and split corresponds to minimal impurity [56]. Two frequently used measures of impurity are gini index and cross-entropy [55]:

$$\text{Gini Index} = 2p(1-p) \tag{5.16}$$

$$\text{Cross-entropy} = -p\log p - (1-p)\log(1-p) \tag{5.17}$$

where p is Equation 5.15 corresponding to class $y_i = 1$.

5.2.4.2 *Random forest*

A random forest is established based on DTs and two concepts referred to as bootstrapping and bagging [57]. The former involves random selection with replacement of instances from the training dataset to yield a bootstrap dataset of the same size as the training set [34]. Several bootstrap datasets are made, and for each one, a tree is built to fit the data [57]. The tree nodes are defined based on the variable and split found to be optimal from a randomly selected subset of the available variables [57].

Bagging, also called bootstrap aggregation, involves averaging predictions made by each model, in this case trees, built from each of the bootstrap datasets [58] [57]. As such, the variance of the final prediction is reduced [58].

Bagging is an ensemble method, characterized by the use of several random and thus at best unbiased models to yield a final prediction [57] [59].

5.2.4.3 *Extremely Randomized Trees*

In an attempt to improve the accuracy of the random forest further, the Extremely randomized trees (ET) method with increased randomization levels, was proposed [59].

Contrary to the random forest approach where bootstrap datasets are created, the whole training set is used to built each tree, or stump [59].

As for random forests, tree nodes are defined based on a randomly selected subset of variables. However, the split or more precisely the cut-off value in the case of numerical variables, belonging to each variable is not optimized but rather chosen at random. Then, defined by their random cut-off value, the optimal variable is selected for the node [59].

5.2.5 *Boosting*

Boosting is a committee-based machine learning approach in which several sub-optimal models are employed to generate a final, powerful "committee" [60]. As such, it resembles bagging [57]. The main difference is that in boosting, the weak models evolve over time [57].

Boosting is utilized for both classification and regression purposes, and are typically based on DTs although other learning methods can also be suitable [60]. AdaBoost and gradient boosting are two commonly used boosting methods [60]. The latter is described next.

5.2.5.1 *Gradient Boosting*

Gradient boosting involves a model predicting the error of the previous model, which in turn predicts the error of its former model, and so on. This results in a consistent improvement (boosting) of the error (gradient), i.e. the prediction error is gradually reduced [61].

Considering $m = 1, 2, ..., M$ weak or interim classifiers. The gradient boosting involves, for each current tree denoted $f_{m-1}(x)$, determination of the optimal parameters defining the next tree, $f_m(x)$ [60].

5.2.5.2 *Light gradient boosting machine*

The computational complexity of the gradient boosting decision tree (GBDT) method is proportional to the number of instances in the dataset, as well as the number of features. Accordingly, when dealing with large datasets, this method is time-intensive. The approach to overcome this issue was to introduce to GBDT ways in which data could be sampled, thereby reducing the overall dataset size [62].

As such, the light gradient boosting machine (LGBM) was developed. In short, LGBM combines GBDT with gradient-based One-Side Sampling and Exclusive Feature Bundling. The former involves favoring instances with the largest gradients, i.e. the instances with a tendency to be misclassified, defined as "under-trained". The Exclusive Feature Bundling algorithm groups features that seldom have nonzero values at the same time, defined as exclusive features. The features now appear in bundles, thereby reducing overall complexity [62].

5.3 EVALUATING MODEL PERFORMANCE

A crucial step in model selection and evaluation is to define the hyperparameter configuration best fit to the problem at hand [63]. Examples are the k number of nearest neighbours that the ReF algorithm evaluates, and the regularization parameter λ in RR.

5.3.1 *Cross-validation*

Samples used to evaluate model performance should ideally not have been included in the data with which the model was trained. For smaller datasets however, it can be problematic to isolate the test data both because it may be too small to provide trustworthy evaluations, and because it renders the train dataset scarce. Resampling can offer a solution, of which bootstrapping is an example. Another common method is cross-validation (CV) [32].

K-fold CV involves dividing the training data into k number of groups or *folds*. Then, each of the k groups are in turn defined as the validation set while the others are used as training sets. So for each k, the other folds are trained to fit this particular fold [34].

Overall performance of the model is then defined as e.g. mean performance of all k validations [34]. Measures for evaluating model performance are outlined next.

5.3.2 *The Confusion Matrix*

Considering a binary classification problem, a confusion matrix shows, for a given classifier and test data, the number of true positives (TP) and negatives (TN), as well as false positives (FP) and negatives (FN) [64]. When the test data is imbalanced the minority and majority class are typically defined as the positive and negative class, respectively [65]. The confusion matrix forms the basis of several measures of classifier performance [64].

Accuracy or error rate (1 - accuracy) are commonly used, accuracy being defined as $\frac{TP+TN}{P+N}$, with P and N the number of positives and negatives in the test data, respecitvely [66] [64]. A shortcoming of these metrics is the falsely high performance that may be obtained for imbalanced datasets, e.g. models consistently predicting the majority class [66].

5.3.3 *The Receiver Operating Characteristic Curve*

The receiver operating characteristic curve (ROC) is defined by plotting true positive rate, TPR $= \frac{TP}{P}$, versus false positive rate, FPR $= \frac{FP}{N}$ TPR is equal to the sensitivity of the model, while FPR = 1 - specificity [64].

For scoring classifiers considering a binary classification problem, the ROC may be obtained by varying the score threshold that separates instances into the two respective classes. Accordingly, choosing a threshold involves defining the optimal trade-off between a high TPR while maintaining a low FPR [64].

The area under the ROC, referred to as AUC, is a measure used to compare the prediction ability of classifiers. A large area reflects high TPR and low FPR [64], or, equivalently, high sensitivity and specificity [55]. As such, the classifiers performance can be determined from a single number [64]. AUC values range between 0 and 1, and random guessing corresponds to an AUC of 0.5. [64].

For imbalanced data sets in which the number of instances in the minority class is low, the AUC can be unreliable as a measure of model performance[66]. Then, Precision-Recall (PR) have proved more suitable, especially due to its focus on the positive, i.e. minority, class [66].

5.3.4 *The Precision-Recall Curve*

The PR curve is defined by plotting precision, defined as $\frac{TP}{TP+FP}$, versus recall, which is equal to the TPR [64]. A well-performing classifier is associated with both high precision and recall, and the PR AUC may be used as a measure of performance [66]. Note that hereinafter, AUC will refer to the ROC AUC.

For a given scoring threshold, the F-score is a commonly used measure, defined as [66]

$$F_\beta = (1+\beta^2)\,\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{5.18}$$

where β is often set equal to 1, yielding the F_1 measure [66].

Part II

METHODS

Chapter 6 includes information on acquisition and processing of the MRIs, as well as clinical characteristics of the patient cohort.

In Chapter 7, eleven experiments designed in an attempt to achieve the objectives of this thesis are outlined.

6 DATA ACQUISITION AND PREPARATION

This thesis was part of the Functional MRI of Hypoxia-mediated Rectal Cancer Aggressiveness (OxyTarget) study [21], which aim was to identify image biomarkers related to metastasis-free survival and response to CRT in rectal cancer [67].

192 individuals were enrolled into the study between October 2013 and December 2017. A total of 111 individuals were excluded from further analysis in this thesis due to rectal cancer not being histologically confirmed, standards for image acquisition or quality not met, or other problems experienced during image acquisition and processing [68].

35 patients received preoperative treatment, of which 32 received CRT, while 3 patients received radiotherapy only. Considering that radiotherapy will have the greatest effect on tumor response, while chemotherapy prevents spreading of disease, all 35 patients will be included in what is referred to as the nCRT cohort in the remains of this thesis. Both chemotherapy and radiotherapy were performed as per clinical guidelines [69].

T2WIs and DWIs were obtained for all 81 patients prior to treatment. Acquisition protocols are included in Section 6.2.

6.1 CLINICAL FACTORS

Initial staging was done according to the TNM standard, edition 7 [24]. For the nCRT cohort, two measures were used to evaluate response to this nCRT as mentioned in Section 3.1.2; posttherapy pathological T-staging, denoted ypT according to [24] as above, and TRG [25]. Clinical characteristics of the patient cohort are summarized in Table 6.1.

PFS was defined as time from study enrollment to some event, either local recurrence, metastasis or death. When used as binary endpoint, i.e. event occurring or not occurring (corresponding to $y = 1$ and $y = 0$, respectively), no time frame was defined. This means that the endpoint was assigned based on whether or not an event had occurred during the time from inclusion to the study until January 20th 2020, when the study ended. Since patients were enrolled at different times, from October 2013 to December 2017, a more accurate measure would have been PFS three years after inclusion, for example. Events tend to occur early, however, so the difference between these two definitions of PFS is small; only two patients had first event occurring at a time longer than three years post inclusion.

Number of participants	81
Sex (women, men)	28 (35%), 53 (65%)
Age (mean)	64
Initial T-stage (AJCC)	
1	0
2	12
3	41
4	28
Treatment	
Surgery alone	37
Preoperative CRT	36*
Palliative care alone	7
Palliative care and surgery	1
Response to nCRT	
ypT $(0-1, 2-4)$	(8, 27)
TRG $(0-1, 2-3)$	(12, 23)
Survival	
PFS (event, no event)	(32, 49)

Table 6.1: Clinical characteristics of the study cohort. An event refer to local recurrence, metastasis or death. No event indicates PFS. * One patient did not undergo surgery, and is not included in the nCRT cohort.

6.2 MR IMAGES

6.2.1 *Acquisition protocol*

MRIs were acquired using a Philips Achieva 1.5T machine from Philips Healthcare, Best, The Netherlands [68].

For each patient, T2WI was performed using a $180 \times 180\,\text{mm}^2$ field of view (FOV) and a 512×512 matrix dimension. The thickness of each slice was 2.50 mm, yielding a voxel size of $0.35 \times 0.35 \times 2.50\,\text{mm}^3$. Each slice was positioned perpendicular to the tumor axis.

DWIs were acquired using seven different b-values; $b = 0, 25, 50, 100, 500, 1000, 1300\,\text{s/mm}^2$. FOV was $160 \times 160\,\text{mm}^2$, matrix dimension 128×128, and slice thickness 4.00 mm, yielding a voxel size of $1.25 \times 1.25 \times 4.00\,\text{mm}^3$. The spacing between slices was 4.30 mm.

6.2.2 *Delineation*

Manual delineation of whole tumors was performed on the T2WIs by two radiologists with 14 and 7 years of experience, respectively [68]. These delineations were fit to the DWIs as explained in Subsection 6.2.3.

The delineations performed by the two radiologists resulted in two binary masks, referred to as mask1 and mask2, in which voxels outside the ROI were indicated by zero's.

6.2.3 *Preprocessing*

REGISTRATION When imaging a tumor using different acquisition schemes, registration (also referred to as coregistration) is necessary in order to combine the information provided by the different images. Registration involves a mapping of the image being transformed, called the moving image, so that it fits the image chosen as a reference. As such, differences in how the patient is positioned for the two respective acquisition schemes are taken into account [70].

For this purpose, the SimpleElastix package Release 01, a part of SimpleITK, was used [71]. Registration was performed by Franziska Knuth at The Norwegian University of Science and Technology (NTNU). A brief summary of the process is presented in the next few paragraphs.

Raw images were of the Digital Imaging and Communications in Medicine (DICOM) format, each capturing a 2D slice. DICOM images are sorted into stacks and converted to the 3D Neuroimaging Informatics Technology Initiative (NifTI) format. This was performed for the T2WIs, and for DWIs corresponding to each of the seven b-values, yielding eight NifTI images per patient.

Registration was performed in several rounds. First, to account for movement of the patient in the course of acquiring all seven b-values, registration was conducted considering the DWIs only.

Next, the DWI with $b = 0$ was defined as moving image and registered to fit the T2WI. No mask was used. The transformation, consisting of a deformation vector field, defined from this registration process was then used directly to register the remaining DWIs.

Finally, delineation masks were obtained for the DWIs by again using the transformation, this time in reverse. As such, the mask defined on the T2WI was registered to fit the DWIs, yielding, for each of the two delineations, a single mask used for DWIs of all b-values.

RESAMPLING Two sets of images were obtained, hereinafter referred to as Dataset1 and Dataset2. For the former, no adjustments of the voxel dimension were performed, and resolution is as described in Subsection 6.2.1. The voxels of the images in Dataset2 were resampled to yield the isotropic size of $1 \times 1 \times 1\,\text{mm}^3$.

CROPPING All images in both datasets were cropped so that a $10 \times 10\,\mathrm{mm}^2$ margin enclosed the union of the two masks. The scripts for both resampling and cropping were written by Franziska Knuth.

6.2.4 *Intensity Discretization*

Prior to feature extraction as will be described shortly, in Subsection 7.1.1, discretization or binning of image intensities were performed according to a defined bin width. Voxel intensities are calculated according to Equation 4.1 Chapter 4. The intensity binning tool is an integrated part of the PyRadiomics package [7] [8], further described in the next chapter.

The authors of [72] argue that the total number of intensity bins should be in the range of $8 - 128$ when extracting texture features. Their argument is based on [73] in which ^{18}F-FDG PET scans are considered.

This turned out to be difficult to ensure, however. Images considered in this thesis were not normalized. T2WIs consisted of intensities in the range of about $20 - 2000$. For the DWIs, the variation was quite high, both between images with the same b-value, and (as expected) across b-values. Images with $b = 0\,\mathrm{s/mm^2}$ consisted of larger grey-value ranges, some up to 4500, while greater b-values yielded lower numbers, also as expected.

In an attempt to take this into account, two different bin widths were tested: 25 and 35.

7 PREDICTION MODEL OPTIMIZATION

Experiments were designed in an attempt to achieve the following three aims: first of all, establish high-performing binary prediction models based on radiomic data while taking issues regarding overfitting, poor regularization and imbalanced datasets into account. Second, investigate whether texture features in particular were selected as relevant. And, finally, evaluate reproducibility with respect to voxel size, intensity discretization, and VOI segmentation.

Accordingly, the eleven experiments presented in Section 7.2 were performed. The Python-based Biorad program, Section 7.1, was used for this purpose.

Four combinations of patient cohort and response variable (RV) were considered, as presented in Table 7.1. These were: 1) the all patients cohort (n = 81) with RV = PFS, 2) the nCRT cohort (n = 35) with RV = PFS, 3) the nCRT cohort with RV = TRG, and 4) the nCRT cohort with RV = ypT. Note that PFS is defined in Section 6.1, while TRG and ypT both are defined in Section 3.1.

Samples	**RV**	y = 1	y = 0
All patients	PFS	32	49
nCRT cohort	PFS	14	21
nCRT cohort	TRG	12	23
nCRT cohort	ypT	8	27

Table 7.1: The different combinations of samples and binary RV used to train prediction models. The columns denoted $y = 1$ and $y = 0$ refer to the number of samples belonging to each class. For PFS, 1 and 0 denotes event and no event, respectively. For TRG and ypT, 1 and 0 denotes good and bad response, respectively.

7.1 BIORAD

Both extraction of radiomic features as well as building and testing of models were achieved using Biorad. The first version of this program was developed by Geir Severin Rakh Elvatun Langberg [74], and the second version by Ahmed Albuni, both at the Norwegian University of Life Sciences: NMBU. The latter version was used in this thesis and is available from GitHub as https://github.com/ahmedalbuni/biorad.

Biorad is based on the Python programming language, version 3.6.2, and consists of two main parts: a tool for radiomics feature extraction, and a tool for model building and comparison. These are the topic of Subsections 7.1.1 and 7.1.2, respectively.

7.1.1 *Radiomic feature extraction*

The feature extraction tool is based on the PyRadiomics package [7] [8]. All features are calculated accordingly, as described in Sections 4.1 - 4.3. Segmentation of the VOI using a binary mask, as well as binning of intensities, are both integrated in the PyRadiomics software.

Shape, first-order and texture features were extracted from both T2WIs and DWIs . Note that shape features are calculated based on the binary mask and are thus independent of the images.

107 features were extracted from the T2WIs: 14 shape features, 18 first-order features, and the remaining described texture; 16 from the GLSZM, 16 from the GLRLM, 5 from the NGTDM, 14 from the GLDM and 24 from the GLCM.

For all DWIs, the same mask is used, thereby resulting in the same exact shape features for all seven images. As such, when features from the DWIs were combined, the 107 features including those describing shape were only extracted from one of the images. From the remaining DWIs, 93 features (first-order and texture) were derived.

Considering all eight images (one T2WI and seven DWIs) belonging to each patient, a total number of 772 features were derived for each patient.

NOTATION The output comma-separated values (csv) file containing all radiomic features (n = 772) corresponding to each patient is denoted $\mathbf{X}_0$. After inclusion of endpoint (PFS, TRG or ypT) and deletion of the column containing patient ID numbers, the feature file is referred to as $\mathbf{X}$.

Table 7.2 shows an overview of the various feature files that will be analysed in the experiments presented in Section 7.2. Note that for $\mathbf{X}_{Tb5}$, the second highest b-value (denoted b5) is chosen in order to derive features from highly diffusion weighted images while avoiding too much noise. The latter may appear in DWIs with $b = 1300\,\mathrm{s/mm^2}$.

Note that no investigation of feature correlation was performed prior to feature selection and classification.

7.1.2 *Features selection and classification*

Prediction models were established based on feature selectors and classifiers as described in Sections 5.1 - 5.2. In addition, the case of no feature selection was evaluated.

ALGORITHMS AND HYPERPARAMETERS All algorithms are based solely on Scikit-Learn Software, except LGBM, FS and ReF. The LGBM

Notation	Features	n
$\mathbf{X}$	T2WIs & all DWIs	772
$\mathbf{X}_T$	T2WIs	107
$\mathbf{X}_{Tb5}$	T2WIs & DWIs (b5)	214
$\mathbf{X}_s$	Shape features	28
$\mathbf{X}^t_{Tb5}$	Texture features from $\mathbf{X}_{Tb5}$	150

Table 7.2: Description of the features included in each input file used in Section 7.2. n = number of features per patient.

algorithm is part of a Microsoft Open Source Project [75], while the FS algorithm is from the Skfeature package. The ReF algorithm is part of the Scikit-Rebate project, accessible from GitHub as `https://github.com/EpistasisLab/scikit-rebate`.

All selectors and classifiers considered required optimization of at least one hyper-parameter. In Table 7.3, the parameters for which Biorad required a range to be specified by the user are presented. Note that a range was required, i.e. the program did not allow for fixed numbers.

Here, k is the number of top-scoring features to select, n the number of nearest neighbors, α_{RR}, C_{LR} and C_{SVC} regularization parameters, MTD = maximum tree depth, MNI = minimum number of instances at a leaf node, and MNL = maximum number of leaf nodes per tree. Note that $C_{LR} = \frac{1}{2\alpha_{RR}}$ and $C_{SVC} = \frac{1}{\alpha_{RR}}$ [47] [52].

Both gini index and entropy was evaluated as impurity measure for the DT and ET classifier.

Remaining parameters required by each algorithm were left to default value of which information can be found by accessing the reference specified in Table 7.3. Note that the abbreviations MTD, MNI and MNL are used here only.

Together, the parameters in Table 7.3 constituted the hyper-parameter space from which configurations were established and evaluated. These values were chosen based on available documentation on the various algorithms as well as results from test runs. Performance, overfitting and ability to generalize were considered. Model complexity and the number of features to select were attempted tuned thereafter.

As evident from Table 7.2 however, the dimensionality of inputs varied greatly, from 28 to 772 number of features per patient. Accordingly, defining parameter ranges that performed well with all inputs were difficult. Where minor adjustments were made, it will be specified.

MEASURE OF PERFORMANCE The measure used to evaluate and compare prediction models was AUC as described in Subsection 5.3.3. In addition, selected experiments performed on the nCRT cohort were

Algorithm	Ref.	Parameter	$n = 81$	$n = 35$
VT	[37]	Threshold	0.1 - 0.9	0.1 - 0.9
MI	[37]	k	1 - 15	1 - 9
MI	[37]	k	1 - 15	1 - 9
ReF		k	1 - 15	1 - 9
ReF		n	1 - 3	1 - 3
RR	[47]	α_{RR}	1 - 5	1 - 5
LR	[47]	C_{LR}	1 - 4	1 - 4
SVC	[52]	C_{SVC}	1 - 4	1 - 4
DT	[76]	MTD	10, 20, 30	10, 20, 30
DT		MNI	1 - 15	1 - 15
ET	[77]	MNI	5 - 15	5 - 15
LGBM	[75]	MTD	5 - 30	5 - 30
LGBM		MNI	1 - 15	1 - 15
LGBM		MNL	3 - 20	3 - 20

Table 7.3: Values defining the hyper-parameter space. All notations for parameters are defined in the text above. $n = 81$ and $n = 35$ refer to the all patient and nCRT cohort, respecitvely.

repeated with the F1 score (Subsection 5.3.4) in an attempt to account for these datasets being imbalanced.

Model performance was evaluated as achieving a high score with low associated standard deviation (std) and ability to generalize well. The latter corresponds to small differences between train and test scores.

MODEL ESTABLISHMENT AND COMPARISON The approach for evaluating hyper-parameter configurations and prediction models are summarized in the following.

Each model, performing both feature selection (F) and classification (CLF), is expressed as [74]

$$\lambda(\phi, \cdot) = \lambda_{CLF}(\phi_{CLF}, \lambda_F(\phi_F, \cdot)), \tag{7.1}$$

where ϕ is a particular configuration of hyper-parameters considering both the selector and classifier, in other words $\phi = \phi_{CLF} \cup \phi_F$.

CV is used to evaluate the model with different hyper-parameter configurations. In all experiments conducted, the number of configurations to be tested was equal to 80.

The parameter space as shown in Table 7.3 is sampled by a random search function, rather than a more comprehensive grid search [78]. More specifically, the RandomizedSearchCV function from Scikit-

Learn is used with five stratified CV folds [78] [79]. All the available data is included, i.e. no validation set is isolated.

Mean test and train scores, as well as the optimal ϕ, are obtained from the CV schema.

By iterating through the set of prediction models (combinations of selector and classifier), optimal hyper-parameter configuration and performance (both training and testing) was specified for each model. As such, the various combinations of selector and classifier can be compared.

Results according to test performance are presented using heatmaps. Furthermore, for each model, results include values corresponding to the optimal ϕ and the list of selected features. For MI and ReF selectors, the ranking corresponding to each feature is provided.

7.2 EXPERIMENTS

In addition to the random search for the optimal hyper-parameter combinations, model evaluation and comparison was performed in view of several aspects, as evident by Subsections 7.2.1 - 7.2.2.

Features derived from images of different scan types were used, corresponding to $\mathbf{X}$, $\mathbf{X}_T$ and $\mathbf{X}_{Tb5}$ in Table 7.2. Features of different type were considered separately, corresponding to $\mathbf{X}_s$ and $\mathbf{X}^t_{Tb5}$ in Table 7.2. Finally, reproducibility of the result with respect to voxel resampling, bin width and VOI delineation were evaluated.

As described in Section 6.2.3, two sets of images were prepared. Dataset1 contained images of original voxel size, while in Dataset2, the voxels were resampled to yield the isotropic size of $1 \times 1 \times 1\,\text{mm}^3$. Mask1 and mask2 refer to the two VOI delineations.

Experiments 1 and 3 were performed on all four combinations of patient cohort and RV, as presented in Table 7.1. Remaining experiments were only performed with the all patient cohort (n = 81). Experiment 2 was not performed with the nCRT cohort (n = 35) due to the great difference between number of features and number of samples. Furthermore, $\mathbf{X}$ was not considered in experiments $6 - 11$ due to the potentially large number of redundant features.

Results from experiments $1 - 5$ are presented in Chapters 8 and 9. Results from experiments $6 - 11$ are presented in Chapter 10.

In Table 7.4, an overview of the experiments that were performed are presented. Details are given in Subsections 7.2.1 - 7.2.2.

NOTE: "FIXED" PARAMETERS In Table 7.4, note the following for experiments in which the parameters were "Fixed": as mentioned, Biorad required a range designated each parameter, so keeping parameters fixed was not possible. Still, attempts were made to achieve this in order to allow for comparison of experiments related to reproducibility. This involved setting ranges corresponding to the parameters in question equal to only two values, or excluding parameters if default value was the wanted one.

Exp.	Input	Dataset	BW	Mask	Param.
1	$\mathbf{X}_T$	1	25	1	7.3
2	$\mathbf{X}$	1	25	1	7.3
3	$\mathbf{X}_{Tb5}$	1	25	1	7.3
4	$\mathbf{X}_s$	1	25	1	7.3
5	$\mathbf{X}^t_{Tb5}$	1	25	1	7.3
6	$\mathbf{X}_T$, $\mathbf{X}_{Tb5}$	2	25	1	7.3
7	$\mathbf{X}_{Tb5}$	2	25	1	"Fixed"
8	$\mathbf{X}_T$, $\mathbf{X}_{Tb5}$	1	35	1	7.3
9	$\mathbf{X}_{Tb5}$	1	35	1	"Fixed"
10	$\mathbf{X}_T$, $\mathbf{X}_{Tb5}$	1	25	2	7.3
11	$\mathbf{X}_{Tb5}$	1	25	2	"Fixed"

Table 7.4: Summary of experiments. "Fixed" parameters (param.) refer to attempts of achieving this despite Biorad requiring a range. In these experiments, the aim was to evaluate the effect of resampling (Exp. 7), bin width (Exp. 9) and delineation mask (Exp. 11) relative to the situation in experiment 3. Experiments below the line seek to investigate reproducibility. BW = Bin width.

7.2.1 *Scan Type and Number of Input Features*

In experiment $1-3$, the effect of including features from DWIs was investigated. All were performed on Dataset1. Mask1 and a bin width of 25 was used.

EXPERIMENT 1 Features from T2WIs only, i.e. input was $\mathbf{X}_T$.

EXPERIMENT 2 Features from T2WIs and all seven DWIs, i.e. input was $\mathbf{X}$.

EXPERIMENT 3 Features from T2WIs as well as DWIs with b = $1000\,\text{s}/\text{mm}^2$, i.e. input was $\mathbf{X}_{Tb5}$.

7.2.2 *Feature type*

Experiment 4 and 5 were conducted to investigate the effect of considering only shape features, $\mathbf{X}_s$ (n = 28), and only texture features, $\mathbf{X}^t_{Tb5}$ (n = 150), respectively. Both were performed on Dataset1. Mask1 and a bin width of 25 was used in all runs.

EXPERIMENT 4 $\mathbf{X}_s$ was input.

EXPERIMENT 5 $\mathbf{X}^t_{Tb5}$ was input.

7.2.3 *Voxel resampling*

To investigate the effect of non-isotropic vs. fully isotropic voxels on reproducibility of the results, experiment 6 was performed. As before, mask1 and a bin width of 25 was used.

EXPERIMENT 6 Experiment 1 and 3 were repeated, this time on features derived from images in Dataset2.

EXPERIMENT 7 Experiment 3 was repeated, this time on $\mathbf{X}_{Tb5}$ obtained from Dataset2. The exact parameters of a specific, selected model from experiment 3 was now used with Dataset2. This allowed for comparison of this particular model across the two datasets, but none of the other selector and classifier combinations. See the paragraph above on "fixed" parameters. Details regarding the chosen model and corresponding parameters are given in Chapter 10.

7.2.4 *Bin width*

So far, a bin width of 25 has been used. Experiment 7 and 8 were performed to investigate the effect of bin width on reproducibility of the results. Both experiments were conducted with bin width equal to 35. Performed on Dataset1, and mask1 was used in all runs.

EXPERIMENT 8 Experiment 1 and 3 were repeated, this time on features derived from images binned to a width of 35.

EXPERIMENT 9 Similarly to experiment 7; experiment 3 was repeated, this time on $\mathbf{X}_{Tb5}$ obtained from images (Dataset1) binned to a width of 35. The exact parameters of a specific, selected model from experiment 3 was used. This allowed for comparison of this particular model across the two datasets, but none of the other selector and classifier combinations. See the paragraph above on "fixed" parameters. Details regarding the chosen model and corresponding parameters are given in Chapter 10.

7.2.5 *Segmentation mask*

So far mask1 has been used. In experiment 9 and 10, the effect of segmentation mask was investigated. The same approach as in Subsections 7.2.3 and 7.2.4 was chosen. Performed on Dataset1, using a bin width of 25 in all runs.

EXPERIMENT 10 Experiment 1 and 3 were repeated, this time on features derived from images in which the ROI was defined by mask2.

EXPERIMENT 11 Similarly to experiments 7 and 9, experiment 3 was repeated, this time on features derived from images with the ROI defined by mask2. The exact parameters of a specific, selected model from experiment 3 was used. This allowed for comparison of this particular model across the two datasets, but none of the other selector and classifier combinations. See the paragraph above on "fixed" parameters. Details regarding the chosen model and corresponding parameters are given in Chapter 10.

Part III

RESULTS AND DISCUSSION

In Chapters 8 and 9, results from experiments $1-5$ are presented. The former includes results for models predicting PFS, while the topic of the latter chapter is predicting response to nCRT.

In Chapter 10, results from experiments $6-11$ are presented, with the goal of investigating reproducibility.

Note that across experiments, test standard deviation was quite high. Large differences between test and train scores as evident of poor generalization ability were observed. These trends were present in particular from analysis with the nCRT cohort containing only 35 patients.

A discussion on methods and obtained results are included in Chapter 11.

8 PREDICTING PROGRESSION FREE SURVIVAL

In this chapter, results from experiments $1-5$ for models predicting PFS are presented. It consists of two main sections. In the former, results for models trained on radiomic data from all 81 patients are outlined. In the latter, Section 8.2, results for models considering the nCRT cohort only (n = 35) are presented. Before proceeding, a few remarks are made. These apply to Chapters 8 - 10.

NOTE 1: Models that employ VT as feature selector are not given much consideration in the ensuing chapters due to the large number of features typically being selected. Accordingly, such models would be prone to over-fit when trained with datasets of relatively small sizes, as is the case in this thesis. The same is true for models not performing any feature selection. Overall high train scores as well as test std were associated with these models. However, note that in the calculation of selection rates, the VT algorithm is included.

NOTE 2: In some experiments, test AUC values below 50% were obtained. Overall high test std was typically associated with these experiments. This topic will not be given much attention in Chapters 8 - 10, but will be discussed in Chapter 11.

8.1 PREDICTING PFS FOR THE ALL PATIENTS COHORT

8.1.1 *Evaluating performance*

EXPERIMENT 1 In experiment 1, models were trained to predict PFS based on 107 features per patient derived from T2WIs. A selection of models are presented in Table 8.1. Corresponding selected hyper-parameters are included in Table A.1.

In Figure 8.1a, a heatmap showing performance of the various combinations of feature selector and classifier is included. Note that LGBM in combination with FS had training score of 100.0% AUC, i.e. apparently not generalizing well and therefore not included in Table 8.1. The train score of SVC combined with MI was 89.5 ± 2.1%, so the same applies to this model.

No model had AUC score below 50%. Both RR and LR had train scores slightly ($\sim 4\%$) below test score when combined with FS as well as ReF selectors. This occurred in some other experiments as well. A discussion on this topic is included in Chapter 11.

Each selector algorithm is part of six models. Mean test std for the six models in which ReF was used as feature selector was 15.5%. Similarly, mean values for test std of models consisting of the MI and FS selectors, respectively, were 11.1% and 11.6%.

Model	AUC,	
	Test (%)	**Train** (%)
FS and SVC	63.0 ± 11.8	63.4 ± 5.6
FS and DT	62.2 ± 5.9	77.1 ± 4.8
FS and ET	62.1 ± 14.7	69.6 ± 3.6

Table 8.1: Model performance in prediction of PFS for all patients. Selected combinations of feature selector and classifier that performed relatively well in experiment 1. Hyper-parameters for all models are presented in Table A.1.

EXPERIMENT 2 In experiment 2, the number of features per patient was 772, derived from T2WIs and all DWIs. The heatmap showing performance of the various models is presented in Figure A.1. Models performing overall well are included in Table 8.2 with corresponding selected hyper-parameters given in Table A.2 and A.3. As in experiment 1, FS in combination with LGBM had training score of 100% AUC and are thus not included in the table below.

Two models had AUC value below 50%. Two models had train scores slightly below test scores.

Mean values for test std of models consisting of the ReF, MI and FS selectors, respectively, were 12.4%, 11.0% and 10.8%.

Model	AUC,	
	Test (%)	**Train** (%)
FS and DT	68.6 ± 8.9	86.4 ± 3.1
MI and LR	68.2 ± 13.5	60.6 ± 2.2
FS and ET	67.0 ± 10.1	72.7 ± 1.9
MI and ET	64.9 ± 12.9	70.0 ± 6.6

Table 8.2: Model performance in prediction of PFS for all patients. Selected combinations of feature selector and classifier that performed relatively well in experiment 2. Hyper-parameters for all models are presented in Tables A.2 and A.3.

EXPERIMENT 3 In experiment 3, features derived from T2WIs and the DWIs with second highest b-value, equal to 1000 s/mm², were analysed, the latter referred to as DWIs (b5). The heatmap with test scores from this experiment can be seen in Figure 8.1b. In Table 8.3 models performing overall well are presented. The LGBM classifier over-fit when combined with both MI and FS selectors, as evident by the 100% train score.

One combination of feature selector and classifier had AUC values below 50%, as evident from Figure 8.1b. Train scores were consistently higher than test scores across models.

Mean values for test std of models consisting of the ReF, MI and FS selectors, respectively, were 10.0%, 15.0% and 10.4%.

Model	**AUC**,	
	Test (%)	**Train** (%)
MI and ET	67.5 ± 15.0	76.0 ± 3.7
FS and ET	62.7 ± 13.3	91.7 ± 4.2
MI and SVC	62.4 ± 16.2	72.7 ± 3.9
MI and DT	61.0 ± 13.8	86.7 ± 3.0
FS and DT	60.6 ± 11.0	72.4 ± 2.9
FS and LR	59.1 ± 9.6	60.5 ± 1.4

Table 8.3: Model performance in prediction of PFS for all patients. Selected combinations of feature selector and classifier that performed relatively well in experiment 3. Hyper-parameters for all models are presented in Tables A.4 and A.5.

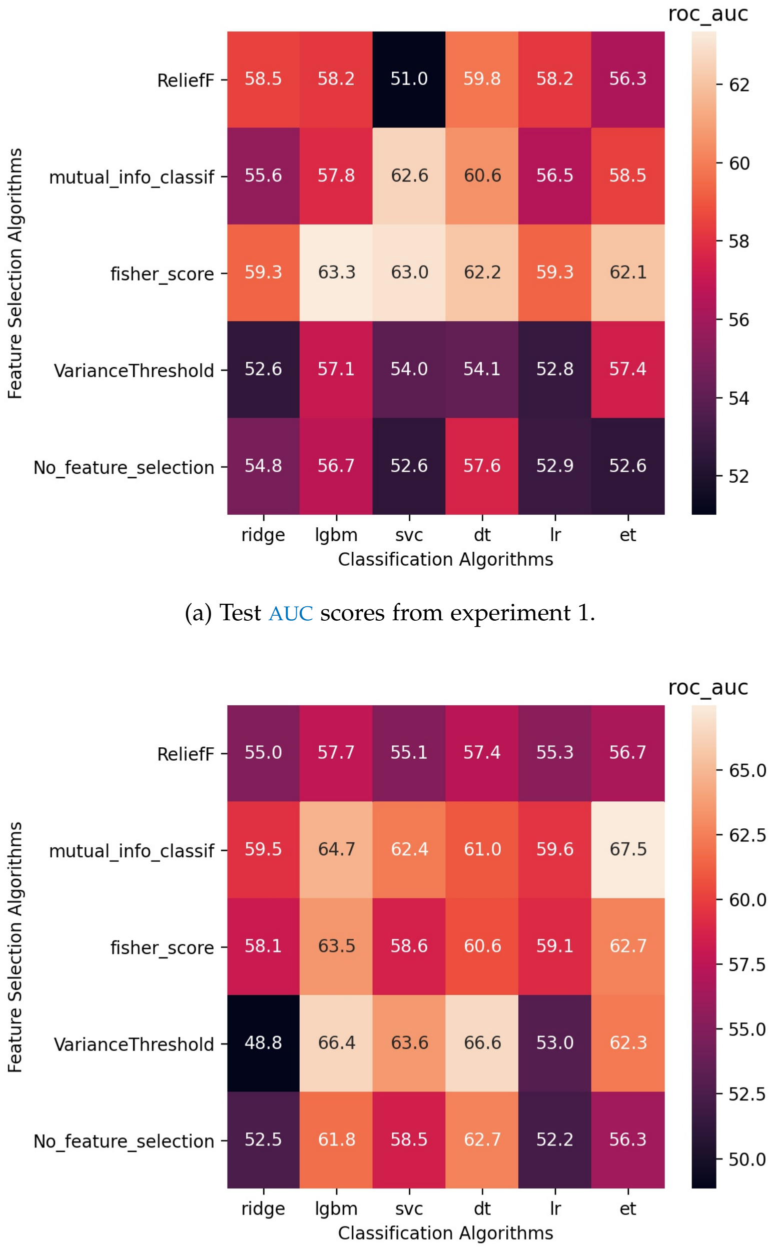
8.1.2 *The predictive value of texture features*

The selection frequency of a feature corresponds to the number of times it was selected relative to the number of times it could have been selected [74]. It is obtained by dividing the number of times a feature appears in an experiment with the number of selector and classifier combinations, i.e. 24.

Figure 8.2 shows selection frequency, or rate, for features selected more than 8 out of the 24 times, i.e. having selections rates > 0.33. This value was chosen in order to show rates for a decent number of features.

As evident from Figures 8.2b and 8.2c, no feature had selection rate above 0.50 in experiment 2 or 3. For the former, this may be expected considering that features from all seven DWIs were included without any prior investigation of feature correlation. Accordingly, several features probably represent overlapping information, rendering selection rates for each individual feature low.

Higher rates were expected in experiment 3 considering features from DWIs (b5) only were included in addition to the T2WI features. However, comparing Figures 8.2b and 8.2c, rates appearing in the former are slightly more constant than for the latter (experiment 3), with a larger difference between the two or three highest rates and the remaining ones. This may suggest a somewhat lower correlation



Feature Selection Algorithms	ridge	lgbm	svc	dt	lr	et
ReliefF	58.5	58.2	51.0	59.8	58.2	56.3
mutual_info_classif	55.6	57.8	62.6	60.6	56.5	58.5
fisher_score	59.3	63.3	63.0	62.2	59.3	62.1
VarianceThreshold	52.6	57.1	54.0	54.1	52.8	57.4
No_feature_selection	54.8	56.7	52.6	57.6	52.9	52.6

(a) Test AUC scores from experiment 1.

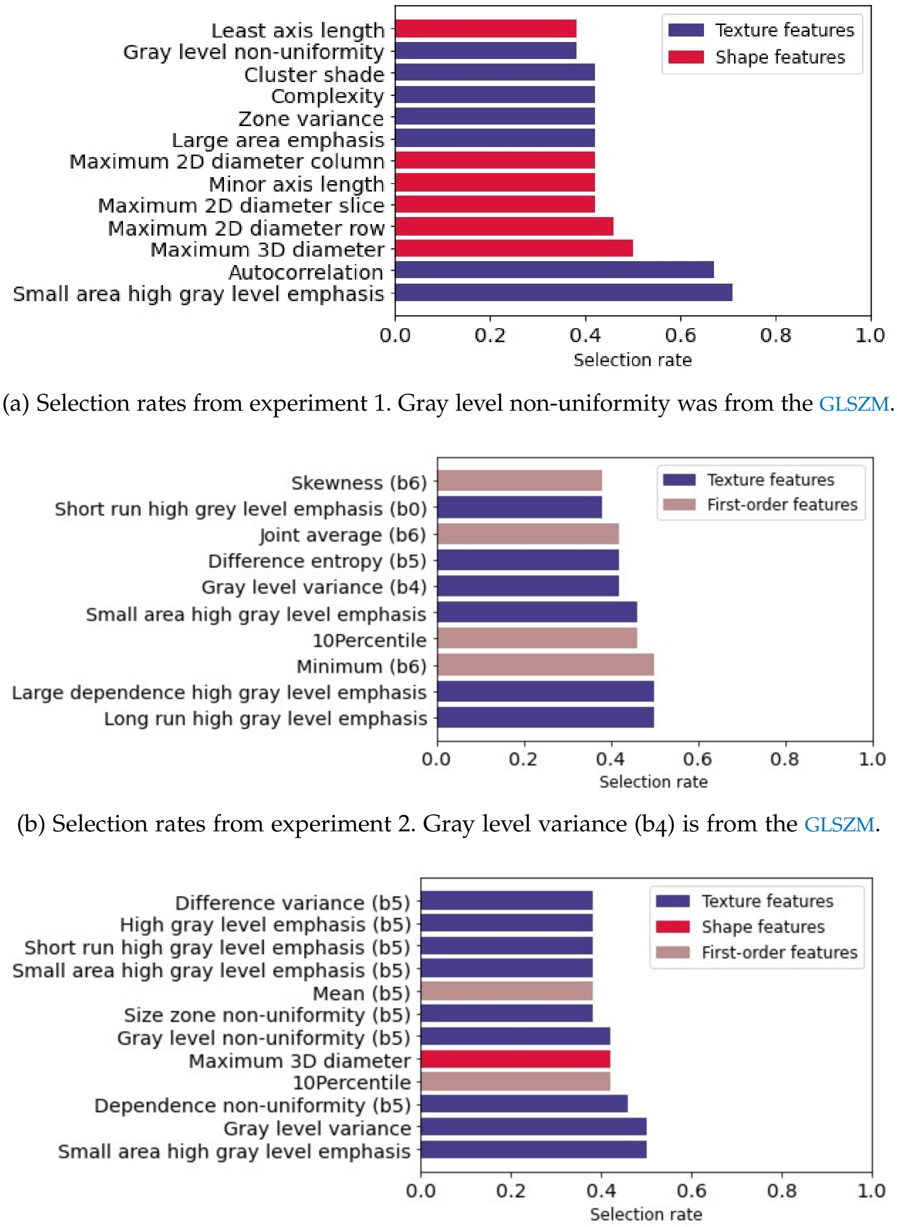
Feature Selection Algorithms	ridge	lgbm	svc	dt	lr	et
ReliefF	55.0	57.7	55.1	57.4	55.3	56.7
mutual_info_classif	59.5	64.7	62.4	61.0	59.6	67.5
fisher_score	58.1	63.5	58.6	60.6	59.1	62.7
VarianceThreshold	48.8	66.4	63.6	66.6	53.0	62.3
No_feature_selection	52.5	61.8	58.5	62.7	52.2	56.3

(b) Test AUC scores from experiment 3.

Figure 8.1: Performance, measured in AUC, for combinations of feature selector and classifiers from experiment 1 (a) and 3 (b).

between features in experiment 3. A discussion on the topic of feature correlation, its effect on selection rate, as well as features potentially representing overlapping information is included in Chapter 11.

From Figure 8.2 it can be seen that small area high grey level emphasis from the GLSZM was the most selected feature in both experiment 1 and 3, together with gray level variance from the GLRLM for the latter experiment. Small area high gray level emphasis was selected at a rate equal to 0.46 in experiment 2.



(a) Selection rates from experiment 1. Gray level non-uniformity was from the GLSZM.

(b) Selection rates from experiment 2. Gray level variance (b4) is from the GLSZM.

(c) Selection rates from experiment 3. Gray level variance was from the GLRLM, while gray level non-uniformity (b5) and high gray level emphasis (b5) were from the GLDM.

Figure 8.2: Feature selection rates from experiment 1 (a), 2 (b), and 3 (c). Remaining features had selection rate ≤ 0.33. Information on which matrix each texture feature was derived from can be found in Chapter 4.

feature scores Each feature selector algorithm except VT designated a score to each input feature. The Biorad program allowed for retrieval of the scores imposed by the MI and ReF selectors. The FS could unfortunately not be obtained.

Consider again Table 8.3 with selected models from experiment 3. The same two top-scoring features were selected by the MI selector in combination with both the ET classifier and SVC, respectively. These were small area high gray level emphasis from the GLSZM and dependence non-uniformity (b5) from the GLDM, scored by the MI algorithm as 0.17 and 0.16, respectively.

EXPERIMENT 4 AND 5 In these experiments, the predictive performance of models based on only shape, $\mathbf{X}_s$, or texture features, $\mathbf{X}^t_{Tb5}$, respectively, were evaluated. Heatmaps showing test performance of the different models are included in Figures A.2 and A.3.

Of the models predicting PFS solely from texture features (experiment 5), the ones based on LGBM or DT classifiers with high test scores generalized poorly, evident by large differences between test and train scores. Mean values for test std of models using the ReF, MI and FS selectors, respectively, were 11.0%, 13.2% and 12.8%. MI combined with the SVC performed relatively well, with test and train scores of 60.4 ± 10.4% and 66.8 ± 6.7%, respectively. A single feature was selected by this model, namely small area high gray level emphasis from the GLSZM. Across models, this feature was selected at a rate of 0.42. The same rate was observed for small area high gray level emphasis calculated from the DWI (b5).

Relatively high selection rates were observed in both experiments. Note that $\mathbf{X}_s$ only contains 28 features per patient, so higher selection rates are expected. $\mathbf{X}^t_{Tb5}$ contains 150 features per patient. In Figures 8.3 and 8.4, features selected at a rate greater than 0.46, i.e. at least 12 out of the 24 times, are included.

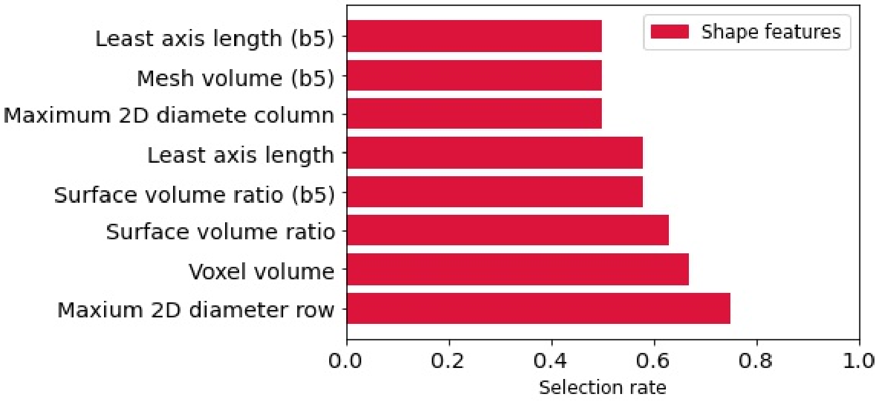


Figure 8.3: Selection rates from experiment 4. Remaining features had selection rate ≤ 0.46.

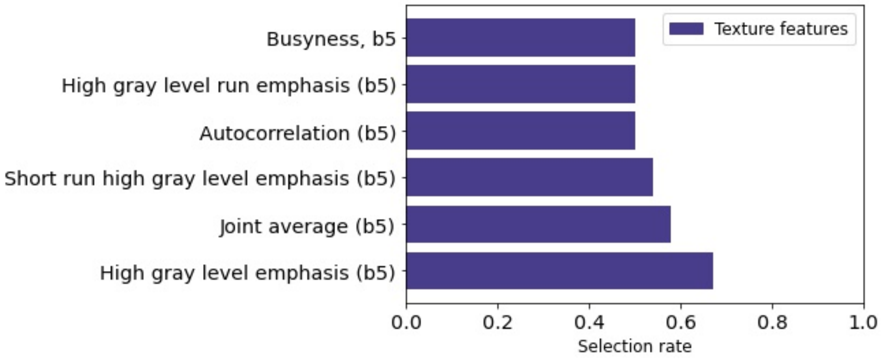


Figure 8.4: Selection rates from experiment 5. Remaining features had selection rate ≤ 0.46. Information on which matrix each feature was derived from can be found in Chapter 4.

8.2 PREDICTING PFS FOR THE NCRT COHORT

From experiment 1 and 3, test scores were overall higher than for experiments run with radiomic data from all 81 patients.

EXPERIMENT 1 Selected models performing overall well with features derived from T2WIs are included in Table 8.4. The heatmap corresponding to experiment 1 is shown in Figure 8.5a. The MI selector in combination with both LGBM and SVC had train scores of $98.6 \pm 3.2\%$ and $97.6 \pm 3.2\%$, respectively, i.e. clearly not generalizing well and thus not included in the table below.

Mean values for test std of the six models consisting of the ReF, MI and FS selectors, respectively, were 23.3%, 20.3% and 19.4%.

Model	**AUC**,	
	Test (%)	**Train** (%)
MI and ET	76.0 ± 15.9	81.6 ± 3.1
MI and DT	74.7 ± 17.3	73.2 ± 6.7
MI and RR	70.0 ± 26.1	75.7 ± 6.1

Table 8.4: Model performance in prediction of PFS for the nCRT cohort. Selected combinations of feature selector and classifier that performed relatively well in experiment 1. Hyper-parameters for both models are presented in Table A.6.

EXPERIMENT 3 Inclusion of features from DWIs (b5) overall increased test scores. However, the highest-scoring model, FS in combination with LGBM, was maximally over-fit, with train score of 100%. Similarly, MI and FS selectors with SVC had train scores of $99.1 \pm 1.2\%$ and $96.8 \pm 2.2\%$. Scores for selected models performing overall well are presented in Table 8.5, and the heatmap shown in Figure 8.5b.

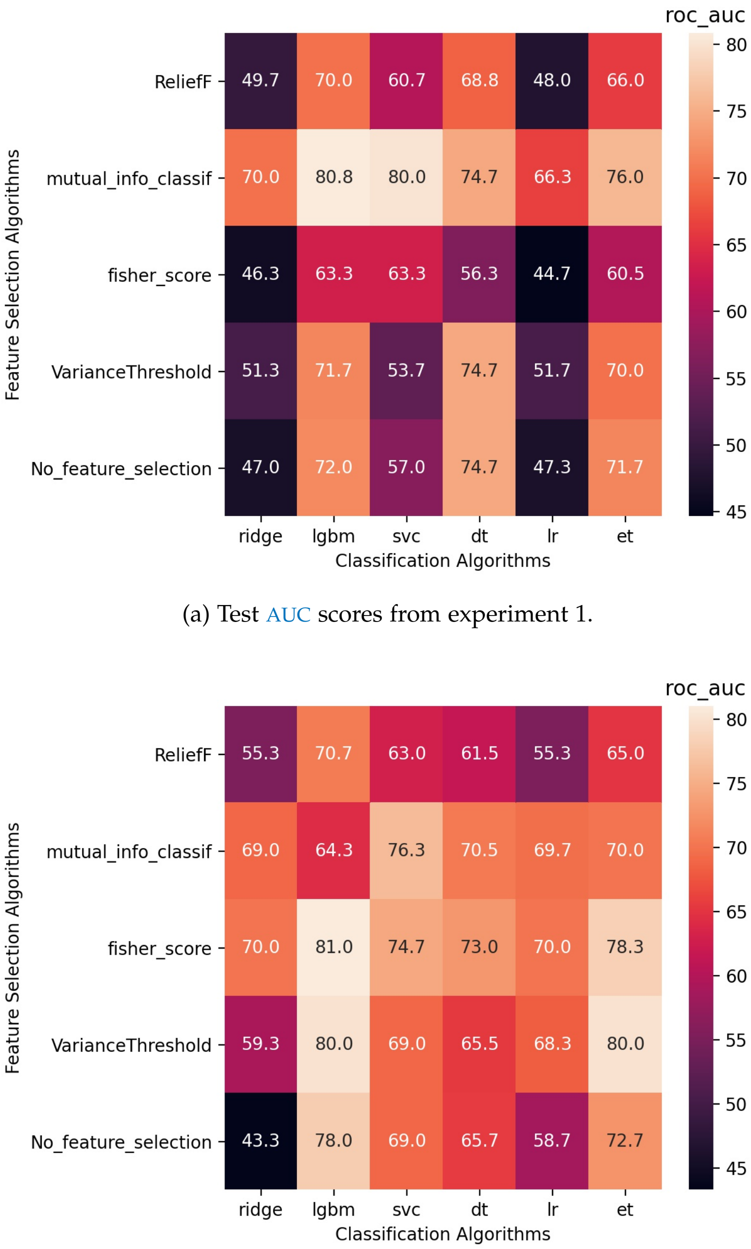
Mean values for test std of the six models consisting of the ReF, MI and FS selectors, respectively, were 19.0%, 19.7% and 17.0%. Again, test std is high. Accordingly, predictions made by the models considered in this section are unreliable.

Model	**AUC**,	
	Test (%)	**Train** (%)
FS and ET	78.3 ± 18.0	84.6 ± 6.1
FS and DT	73.0 ± 10.8	80.7 ± 3.0

Table 8.5: Model performance in prediction of PFS for the nCRT cohort. Selected combinations of feature selector and classifier that performed relatively well in experiment 3. Hyper-parameters for both models are presented in Table A.7.

8.2.1 *The predictive value of texture features*

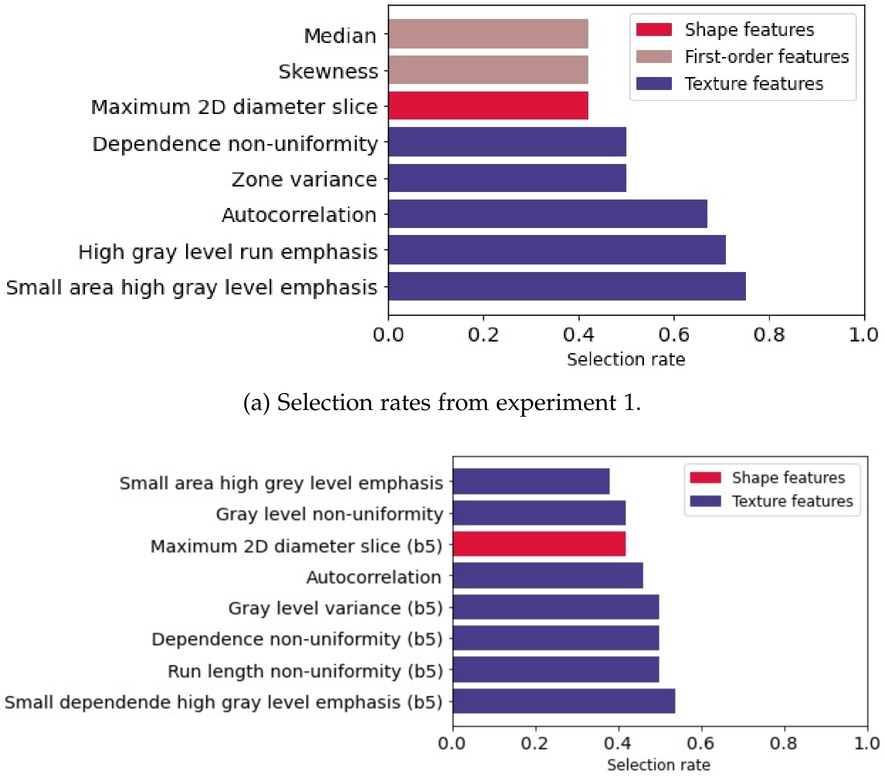
Figure 8.6 shows selection frequency for features with rates > 0.33. The same trend as for the all patients cohort is apparent, with overall lower and more similar rates across features for experiment 3. Furthermore, small area high gray level emphasis from the GLSZM is most frequently selected in experiment 1, now at a rate of 0.75. In experiment 3, this feature is selected at a rate equal to 0.38.



(a) Test AUC scores from experiment 1.

(b) Test AUC scores from experiment 3.

Figure 8.5: Performance measured in AUC, for combinations of feature selector and classifiers from experiment 1 (a) and 3 (b), considering the nCRT cohort.



(a) Selection rates from experiment 1.

(b) Selection rates from experiment 3. Both gray level variance (b5) and gray level non-uniformity (b5) are from the GLDM.

Figure 8.6: Predicting PFS in the nCRT cohort. Selection rates for experiment 1 (a) and 3 (b). Remaining features had selection rate ≤ 0.33. Information on which matrix each texture feature was derived from can be found in Chapter 4.

9 PREDICTING RESPONSE TO PREOPERATIVE CRT

In this chapter, results from experiments 1 and 3 for models predicting response to nCRT are presented. It consists of two main sections, corresponding to the two metrics used to evaluate response. In the former, results for models predicting TRG are presented. In section 9.2, ypT is used as endpoint.

9.1 PREDICTING TRG

9.1.1 *Evaluating performance*

EXPERIMENT 1 Models were trained on radiomic data derived from the T2WIs only. In Table 9.1, a selection of models that performed well are presented. The corresponding heatmap is shown in Figure 9.1a.

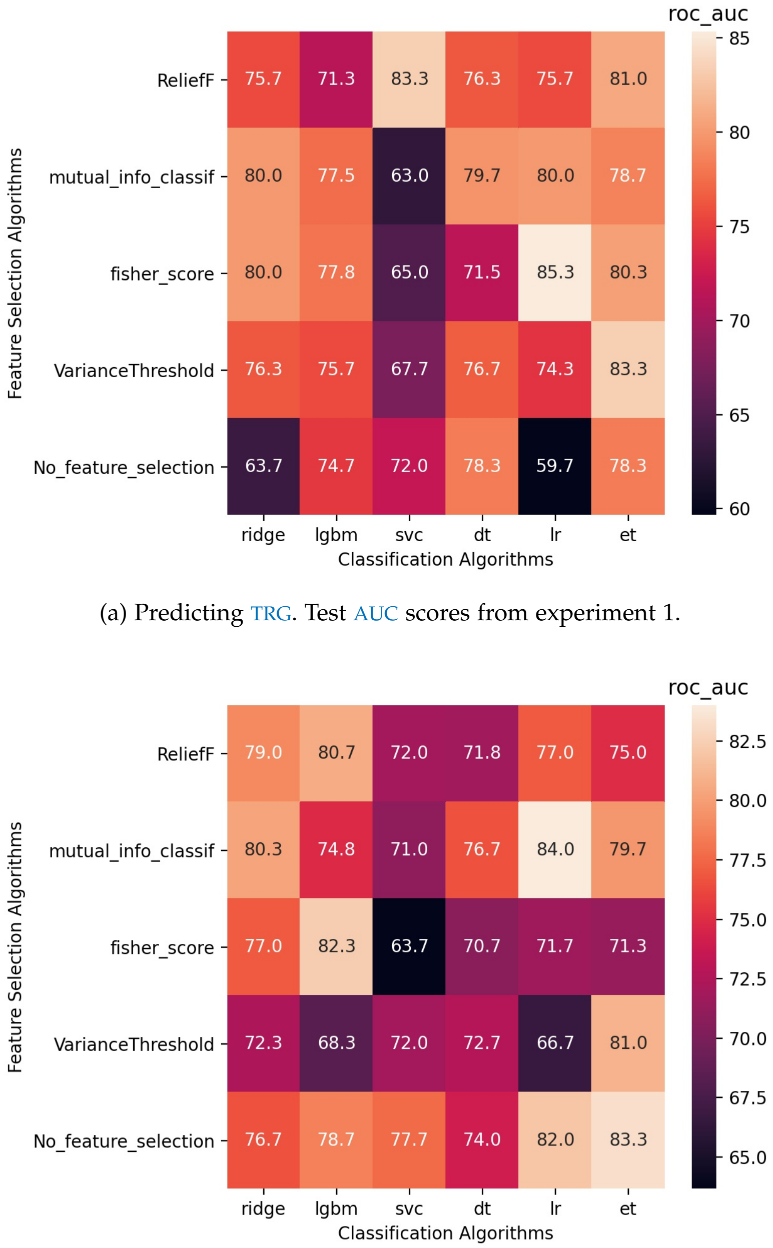
Mean values for test std of the six models using the ReF, MI and FS algorithm as feature selector, respectively, were 13.0%, 12.4% and 13.3%.

Model	**AUC**,	
	Test (%)	**Train** (%)
FS and LR	85.3 ± 11.1	75.2 ± 3.0
ReF and SVC	83.3 ± 7.3	95.3 ± 2.4
ReF and ET	81.0 ± 15.8	80.6 ± 4.6
FS and ET	80.3 ± 12.7	83.5 ± 1.7
MI and DT	79.7 ± 4.1	85.8 ± 2.9

Table 9.1: Model performance in prediction of TRG. Selected combinations of feature selector and classifier that performed relatively well in experiment 1. Hyper-parameters for all models are presented in Tables A.8 and A.9.

EXPERIMENT 3 Results from experiment 3 are presented in Table 9.2, with corresponding heatmap shown in Figure 9.1b. Both ReF and FS had train scores of ∼ 100% when combined with the LGBM classifier. Mean values for test std of the six models consisting of the ReF, MI and FS selectors, respectively, were 12.5%, 15.6% and 17.8%.

Across models in both experiment 1 and 3, values for test std were lower than for models predicting PFS in the nCRT cohort, as evident from Section 8.2.



(a) Predicting TRG. Test AUC scores from experiment 1.

(b) Predicting TRG. Test AUC scores from experiment 3.

Figure 9.1: Performance measured in AUC, for combinations of feature selector and classifiers predicting TRG, from experiment 1 (a) and 3 (b).

9.1.2 *The predictive value of texture features*

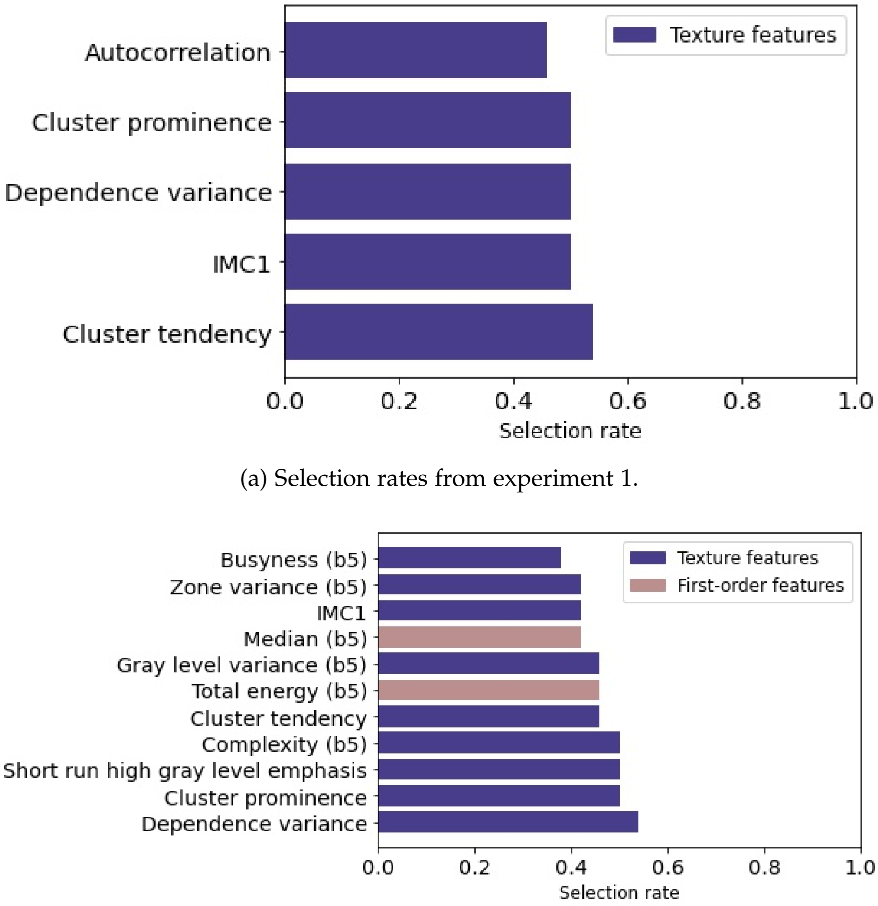
As evident from Figure 9.2a, a relatively low number of features were selected at a rate above 0.33 in experiment 1. The highest selection rate was as low as 0.54. The latter is also true for experiment 3 as shown in Figure 9.2b, however a greater number of features were selected at rates between 0.54 and 0.33. This high number of features

Model	**AUC**, **Test** (%)	**Train** (%)
MI and LR	84.0 ± 16.9	87.2 ± 5.5
ReF and LGBM	80.7 ± 3.4	98.4 ± 2.6
MI and RR	80.3 ± 21.1	85.5 ± 3.8
MI and ET	79.7 ± 12.9	86.4 ± 8.4
MI and DT	76.7 ± 4.6	79.0 ± 2.6

Table 9.2: Model performance in prediction of TRG. Selected combinations of feature selector and classifier that performed relatively well in experiment 3. Hyper-parameters for all models are presented in Table A.10 and A.11.

selected at similar rates is, as mentioned in Chapter 8, suggestive of features representing overlapping information and thereby leaving some redundant.

The two features cluster prominence from the GLCM and dependence variance from the GLDM were selected at relatively high rates in both experiments; 0.54 and 0.50 in experiment 1, 0.50 and 0.54 in experiment 2, respectively.



(a) Selection rates from experiment 1.

(b) Selection rates from experiment 3. Gray level variance (b5) was from the GLSZM.

Figure 9.2: Predicting TRG. Selection rates for experiment 1 (a) and 3 (b). Remaining features had selection rate ≤ 0.33. Information on which matrix each texture feature was derived from can be found in Chapter 4.

9.2 PREDICTING YPT

9.2.1 *Evaluating performance*

As evident from Figure 9.3, overall high AUC scores were associated with experiments considering the nCRT cohort with ypT as endpoint. Scores of some selected models from experiment 1 and 3 are presented in table 9.3.

From experiment 1, mean values for test std of the six models consisting of the ReF, MI and FS selectors, respectively, were 6.0%, 7.8% and 15.2%. From experiment 3, these respective values were 13.0%, 6.7% and 20.6%. Lower values as well as larger variations in test std were observed compared to models trained to predict PFS or TRG in the nCRT cohort.

An important factor in causing such inconsistent and presumably over-optimistic results is the imbalance of the dataset. As seen from

Exp.	Model	AUC, Test (%)	Train (%)
1	ReF and ET	98.0 ± 4.0	94.0 ± 2.1
1	ReF and SVC	98.0 ± 4.0	97.8 ± 1.2
1	MI and DT	94.7 ± 3.1	97.0 ± 1.8
1	ReF and LGBM	94.3 ± 7.9	99.2 ± 1.3
3	MI and DT	96.3 ± 3.2	98.5 ± 0.3
3	MI and ET	94.7 ± 6.9	96.2 ± 2.4
3	MI and RR	94.7 ± 6.9	88.0 ± 11.9
3	FS and DT	79.0 ± 19.1	88.6 ± 2.5

Table 9.3: Model performance in prediction of ypT. Selected combinations of feature selector and classifier that performed relatively well in experiments (Exp.) 1 and 3. Hyper-parameters for all models are presented in Tables A.13 - A.15.

Table 7.1, only 12 of the 35 patients responded well ($y = 1$) according to the TRG metric, while the remaining 23 responded poorly ($y = 0$). When ypT is used as endpoint, the imbalance is even greater, with 8 responding well and 27 responding poorly. Accordingly, it might be difficult to obtain trustworthy results using AUC as metric. A discussion on this topic is included in Chapter 11.

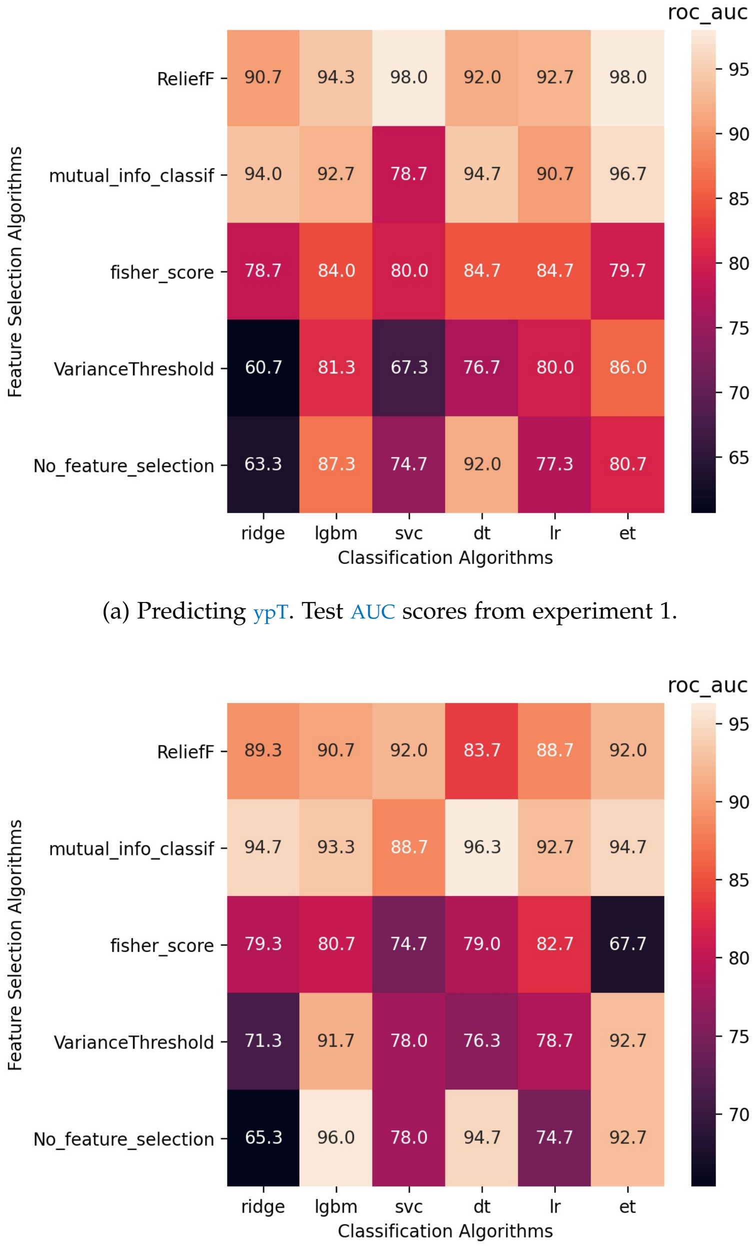
Experiment 1 and 3 were repeated using F1 as performance metric in an attempt to better account for few positive, i.e. $y = 1$, samples. Heatmaps are shown in Figure 9.4. The range corresponding to the minimum number of samples in each leaf node for the ET classifier was changed from $5 - 15$ to $0 - 15$ due to higher performance being observed when using this range during test runs.

With the F1 score, mean values for test std increased: 25.5%, 24.1% and 26.5% for the six models using the ReF, MI and FS feature selector, respectively, from experiment 1. From experiment 3, the respective values were 20.9%, 15.4% and 34.1%.

9.2.2 *The predictive value of texture features*

Selection rates from experiment 1 and 3 when using AUC as scoring metric are presented in Figure 9.5. The same trends as for models predicting TRG were observed: a relatively low number of features selected in experiment 1, with top rates just slightly above 0.5. For experiment 3, the high number of features with rates between 0.33 and 0.50 were again observed.

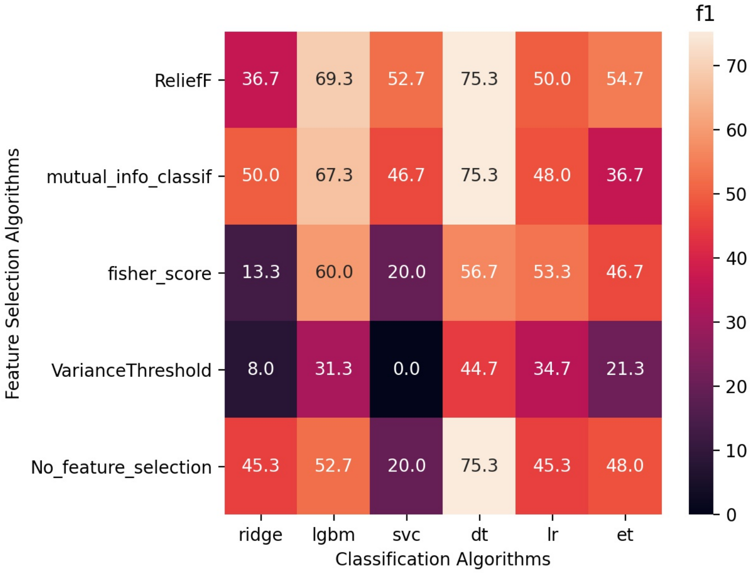
The only feature selected at rates > 0.33 in *both* experiments was small area high gray level emphasis from the GLSZM at rates of 0.46 and 0.38 for experiments 1 and 3, respectively.



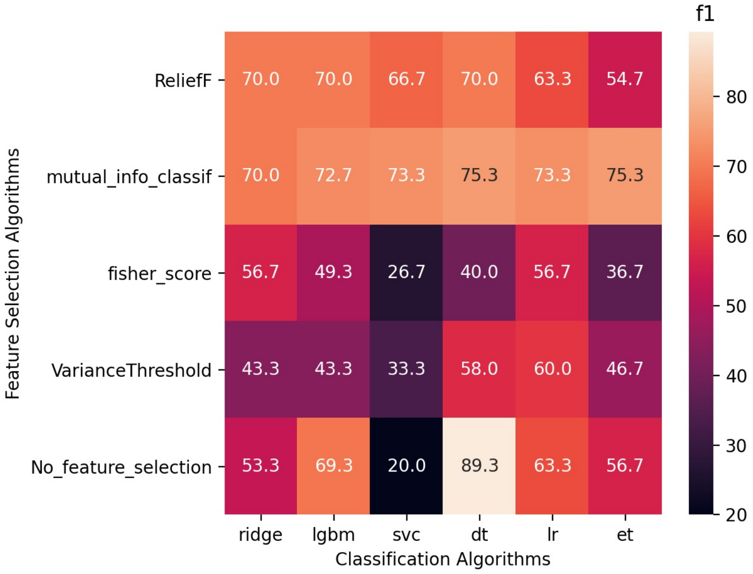
(a) Predicting ypT. Test AUC scores from experiment 1.

(b) Predicting ypT. Test AUC scores from experiment 1.

Figure 9.3: Performance, measured in AUC, for combinations of feature selector and classifiers predicting ypT, from experiment 1 (a) and 3 (b).

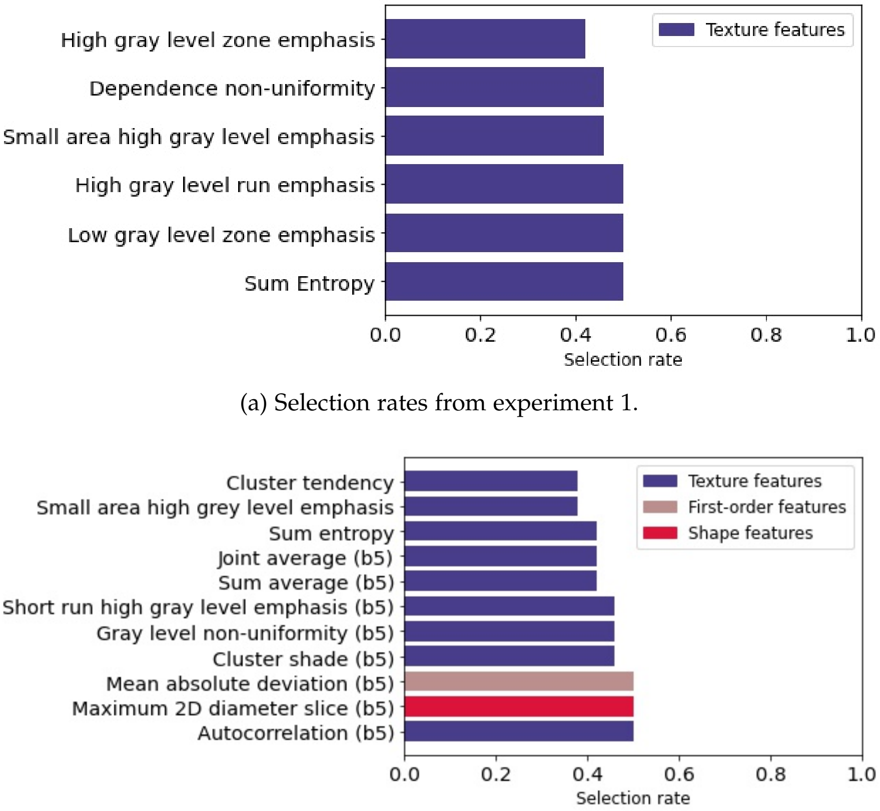


(a) Predicting ypT. Test F1 scores from experiment 1.



(b) Predicting ypT. Test F1 scores from experiment 3.

Figure 9.4: Performance measured as F1 score, for combinations of feature selector and classifiers predicting ypT, from experiment 1 (a) and 3 (b).



(a) Selection rates from experiment 1.

(b) Selection rates from experiment 3. Gray level non-uniformity (b5) was calculated from the GLRLM.

Figure 9.5: Predicting ypT, performance measured in AUC. Selection rates from experiment 1 (a) and 3 (b). Remaining features had selection rate ≤ 0.33. Information on which matrix each texture feature was derived from can be found in Chapter 4.

10 REPRODUCIBILITY OF RADIOMIC RESULTS

Results included in this chapter offer a first evaluation of reproducibility. It is divided into three main sections based on the different aspects in view of which reproducibility is evaluated. Namely, voxel resampling (10.1), intensity binning (10.2) and VOI delineation (10.3). In Section 10.4, the findings in terms of selected features are summarized in an attempt to assess feature robustness.

The same approach was chosen for sections 10.1 - 10.3. In experiments 6, 8 and 10, features derived from T2WIs alone were analysed, before doing a second run in which features from the DWIs (b5), were included. I.e., experiment 1 and 3 were repeated, respectively. In experiment 6, Dataset2 was analysed instead of Dataset1. In experiment 8, a bin width of 35 was used instead of 25, and in experiment 10, mask2 instead of mask1 defined the VOI. This allowed for a first evaluation of reproducibility.

In order to take this investigation of reproducibility one step further, experiment 7, 9 and 11 were performed. In these experiments, the model consisting of FS selector and LR classifier were considered. In experiment 3 as presented in Section 8.1, this model had number of selected features $k = 4$ and regularization parameter $C_{LR} = 3$. Selected features were gray level variance, run percentage (b5) and short run emphasis (b5) from the GLRLM, and dependence non-uniformity normalized (b5) from the GLDM, evident from Table A.5.

When achieving equal values for parameters k and C_{LR} for this model in experiment 7, 9 and 11, performance could be compared across experiments. This model was chosen due to its relatively low std and small difference between train and test score in experiment 3, as evident from Table 8.3.

All models considered in this chapter were trained to predict PFS using radiomic data from all 81 patients.

10.1 VOXEL RESAMPLING

In experiments 6 and 7, features derived from Dataset2 were analysed. Images in this dataset had isotropic voxels with size $1 \times 1 \times 1\,\text{mm}^3$.

10.1.1 *Experiment 6*

EVALUATING PERFORMANCE Heatmaps showing performance of models from experiment 6 are included in Figures A.4 and A.5. Comparing with those shown in Figure 8.1 for Dataset1, performance was all over quite similar between the two datasets. The highest scores when only considering T2WIs were about 63% AUC for both datasets, and increased by a few percentage when including features from DWIs (b5).

Mean values for test std of the six models consisting of the ReF, MI and FS selectors, respectively, were 14.5%, 13.9% and 14.1% in the experiment considering T2WI-derived features only. When $\mathbf{X}_{Tb5}$ was input, values for mean test std were 10.0%, 10.7% and 12.8%, respectively.

SELECTED FEATURES Differences were greater when considering feature selection rates. In particular, in the first run when only considering features from T2WIs, only two features had rates higher than 0.33, and these were as low as 0.50 and 0.46. The features in question were cluster tendency from the GLCM and complexity from the NGTDM, respectively.

The most frequently selected feature in experiment 1 for Dataset1 was small area high grey level emphasis from the GLSZM with selection rate 0.71, as evident from Figure 8.2a. With Dataset2, this feature was selected at the much lower rate of 0.25.

When including features from the DWI (b5) the situation was somewhat more similar to that for Dataset1, as can be seen from Figure 10.1. Five of the same features as in experiment 3 for Dataset1 were selected, as evident from Table 10.2 below. Small area high gray level emphasis was again selected at a rate of only 0.25.

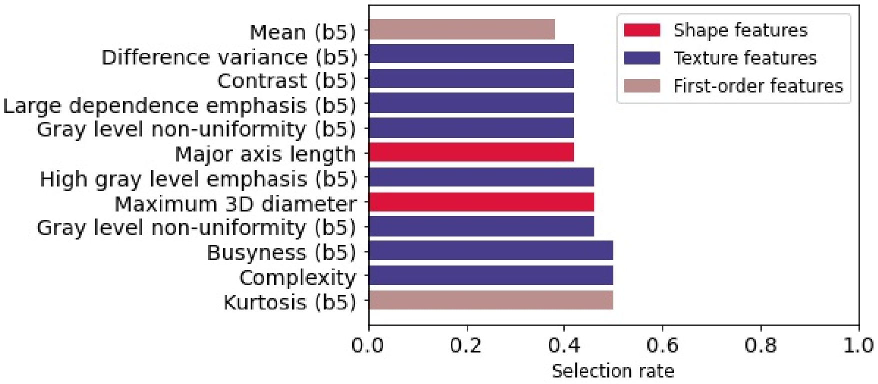


Figure 10.1: Predicting PFS, all patients. Selection rates obtained in experiment 6 from analysis performed on features from both T2WIs and DWIs (b5). Remaining features had selection rate ≤ 0.33. Gray level non-uniformity (b5) was from the GLSZM, while high gray level emphasis (b5) and gray level non-uniformity (b5) were from the GLDM. Information on which matrix each of the remaining texture feature was derived from can be found in Chapter 4.

10.1.2 *Experiment 7*

With $k = 4$ and $C_{LR} = 3$, the model consisting of FS selector and LR classifier performed poorly, with test and train scores of $51.1 \pm 13.7\%$ and $64.8 \pm 1.7\%$, respectively. Selected features were small area low gray level emphasis from the GLSZM, short run low gray level emphasis (b5) from the GLRLM, busyness (b5) from the NGTDM and gray level non-uniformity (b5) from the GLDM.

10.2 BIN WIDTH

In experiments 8 and 9, features were again derived from images in Dataset1, however this time from images with intensities binned to a width of 35 instead of 25

10.2.1 *Experiment 8*

The approach for this experiment was similar to that in Section 10.1, as explained in the introduction of this chapter.

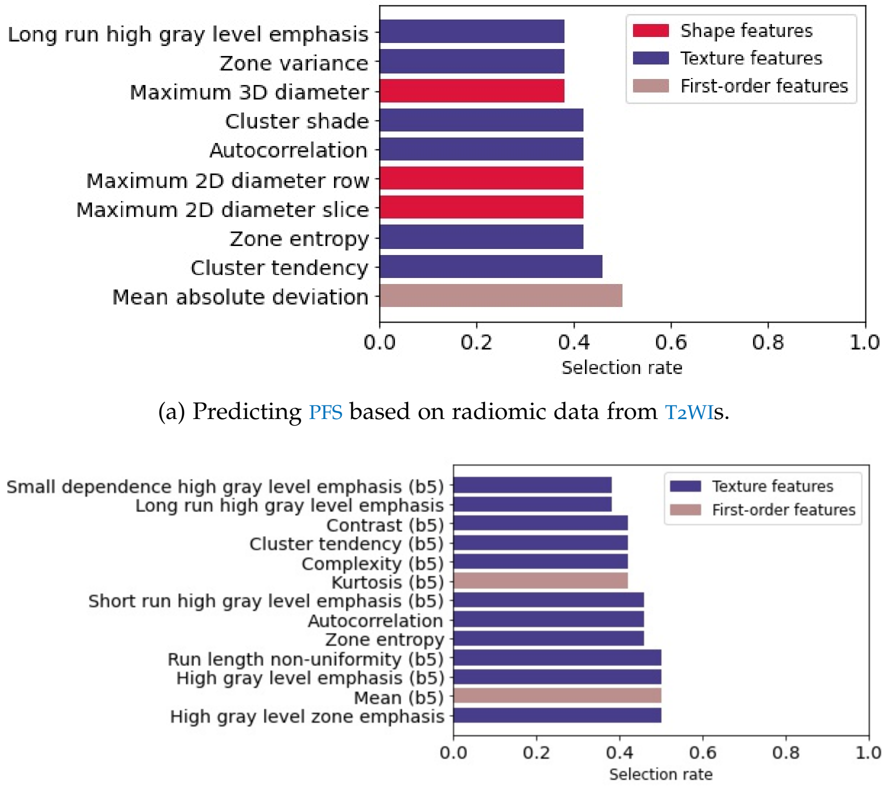
EVALUATING PERFORMANCE Test scores for the various models from experiment 8 can be seen from the heatmaps presented in Figures A.6 and A.7. When features from T2WIs only were input, only one model had test score above 60%. The lowest score was $43.8 \pm 8.3\%$ for the LR classifier with no feature selection. Inclusion of features from DWIs (b5) overall increased test scores, with only two scores below 50%.

Mean values for test std of the six models consisting of the RF, MI and FS selectors, respectively, were 10.9%, 11.9% and 13.3% in the experiment considering T2WI-derived features only. When $\mathbf{X}_{Tb5}$ was input, the respectively values for mean test std were 9.5%, 9.0% and 11.2%.

SELECTED FEATURES Selection rates for experiment 8 are presented in Figure 10.2. As evident from Table 10.1 below, six feature were selected at a rate greater than 0.33 in both experiment 1 and 8, when only analysing features from T2WI.

10.2.2 *Experiment 9*

With $k = 4$ and $C_{LR} = 3$, the model consisting of FS selector and LR classifier again performed quite poorly, with test and train scores of $55.6 \pm 9.9\%$ and $62.6 \pm 6.1\%$, respectively. Selected features were now zone entropy from the GLSZM, low gray level run emphasis (b5) and run length non-uniformity (b5) from the GLRLM, as well as complexity (b5) from the NGTDM.



(a) Predicting PFS based on radiomic data from T2WIs.

(b) Predicting PFS based on radiomic data from T2WIs and DWIs (b5). High gray level emphasis (b5) was from the GLDM.

Figure 10.2: Predicting PFS, all patients. Selection rates from experiment 8, when analysing features from T2WIs only (a), and when including features derived from DWIs (b5) (b). Remaining features had selection rate ≤ 0.33. Information on which matrix each texture feature was derived from can be found in Chapter 4.

10.3 SEGMENTATION MASK

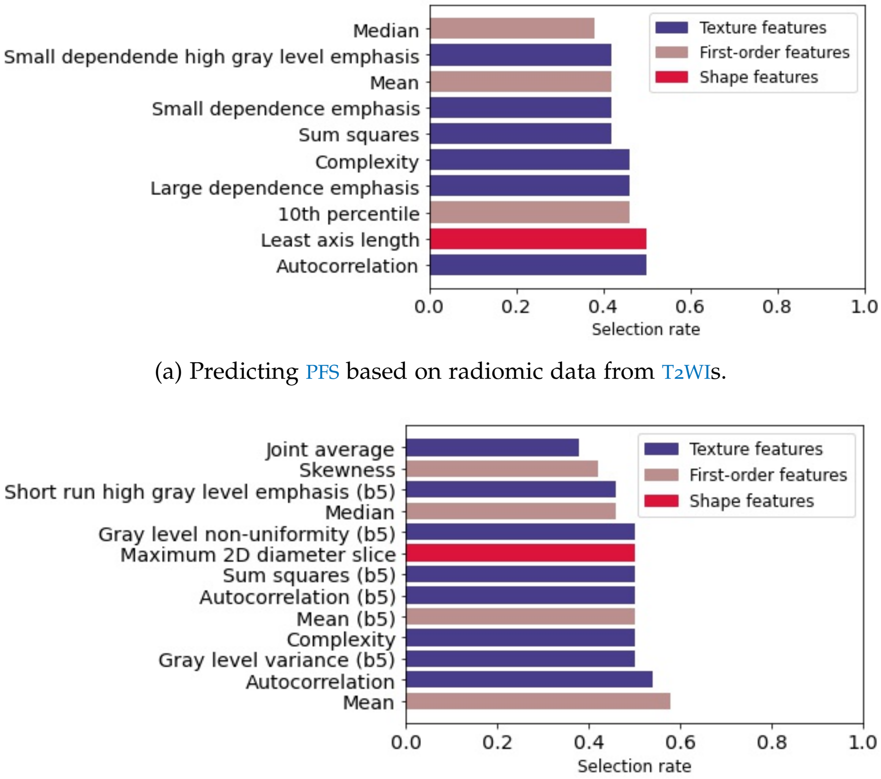
The two final experiments were performed with the aim of evaluating the effect of segmentation mask defining the VOI on the radiomic results. Mask1 was used in experiments $1-9$. In experiment 10 and 11, mask2 was used. The same approach as in Sections 10.1 and 10.2 were used.

10.3.1 *Experiment 10*

EVALUATING PERFORMANCE Heatmaps showing test scores for models from experiment 10 are presented in Figures A.8 and A.9. When simply comparing with Figure 8.1, performance appear somewhat similar; no AUC values below 50% when features from T2WI only were input, and best performances slightly above 60% for models including the RF, MI or FS selector. Mean values for test std of the six models consisting of the RF, MI and FS selectors, respectively, were 14.1%, 15.2% and 8.5%.

When including features from DWIs (b5), a few test AUC values were below 50% as in experiment 3. Mean values for test std of the six models consisting of the RF, MI and FS selectors, respectively, were 14.2%, 13.7% and 10.9%.

SELECTED FEATURES Selected features with rates > 0.33 are shown in Figure 10.3. Findings are summarized next, in the final section of this chapter.



(a) Predicting PFS based on radiomic data from T2WIs.

(b) Predicting PFS based on radiomic data from T2WIs and DWIs (b5). Gray level variance (b5) and gray level non-uniformity were from the GLSZM.

Figure 10.3: Predicting PFS, all patients. Selection rates from experiment 10, when analysing features from T2WIs only (a), and when including features derived from DWIs (b5) (b). Remaining features had selection rate ≤ 0.33. Information on which matrix each texture feature was derived from can be found in Chapter 4.

10.3.2 *Experiment 11*

Similarly to results from experiment 7 and 9, with $k = 4$ and $C_{LR} = 3$, the model consisting of FS selector and LR classifier performed poorly, with test and train scores of $50.9 \pm 10.6\%$ and $58.9 \pm 3.1\%$, respectively. Selected features were now maximum 2D diameter slice, elongation (b5), run variance (b5) from the GLRLM and ID normalized (b5) from the GLCM.

10.4 FEATURE ROBUSTNESS, A SUMMARY

In Table 10.1, features selected at rates > 0.33 in at least one of experiments 6, 8 and/or 10 in addition to experiment 1 are shown. Table 10.2 shows the same for experiment 3. This was done in an attempt of making a first evaluation of feature robustness.

Experiment:	**1**	**6**	**8**	**10**
Feature:				
Complexity	0.42	0.46	-	0.46
Maximum 2D diameter Row	0.46	-	0.42	-
Maximum 2D diameter Slice	0.42	-	0.42	-
Autocorrelation	0.67	-	0.42	0.50
Cluster Shade	0.42	-	0.42	-
Maximum 3D diameter	0.50	-	0.38	-
Zone variance	0.42	-	0.38	-

Table 10.1: Comparing selection rates for features selected at a frequency > 0.33 across experiments. "-" indicates that the feature was selected at a rate ≤ 0.33. Evaluating feature robustness for experiments in which features derived from T2WIs only were analysed, i.e. $\mathbf{X}_T$ was input.

Experiment:	**3**	**6**	**8**	**10**
Feature:				
Maximum 3D diameter	0.42	0.46	-	
High gray level emphasis (b5), GLDM	0.38	0.46	0.50	-
Gray level non-uniformity (b5), GLDM	0.42	0.42	-	-
Difference variance (b5)	0.38	0.42	-	-
Mean (b5)	0.38	0.38	0.50	0.50
Short run high gray level emphasis (b5)	0.38	-	0.46	0.46

Table 10.2: Comparing selection rates for features selected at a frequency > 0.33 across experiments. "-" indicated that the feature was selected at a rate ≤ 0.33. Evaluating feature robustness for experiments in which features derived from both T2WIs and DWIs (b5) were analysed, i.e. $\mathbf{X}_{Tb5}$ was input.

It is again important to note that several radiomic features may provide overlapping information. This can be seen from feature definitions presented in Chapter 4, and may also occur when the same features are calculated from both T2WIs and DWIs (b5). As such, the degree to which the *exact* feature is selected across experiments as presented in the tables above is just a first, possibly deficient, comparison. The evaluation of feature robustness is continued in Chapter 11 by comparing the information provided by each selected feature.

11 DISCUSSION

In this chapter, model performance and the predictive value of texture features are first discussed on the basis of results outlined in the preceding chapters. Then, a discussion on performing machine learning with small sample sizes and imbalanced datasets is presented in Sections 11.3 and 11.4, respectively. Findings and reviews in support of the overall high performance estimates obtained for the nCRT cohort are included in Section 11.5. Finally, in Sections 11.6 and 11.7 the discussion on correlating features and reproducibility initiated in Chapter 10 is continued.

11.1 EVALUATING MODEL PERFORMANCE

The prediction models obtained in this thesis were evaluated and compared on the basis of their AUC scores. The F1 scoring metric was also used for models predicting ypT in Section 9.2. In all experiments, train and test scores were obtained as mean values from a 5-fold stratified CV scheme [79].

As evident from the previous chapters, several models performed poorly, with high mean values for test std as well as differences between train and test score. The latter, as indicative of an insufficient ability to generalize, may result in low test scores when exposing the model to new data. High values for std obviously render models an inappropriate choice for making trustworthy predictions.

WORSE THAN RANDOM PERFORMANCE Some performance estimates below 50% AUC were obtained. This was surprising considering these models performed worse than random guessing. Overall high values for mean test std were often associated with these experiments. In the following, two articles that may provide some clarification are reviewed.

Parker et al. [80] within the field of genetics show that performing machine learning-based class prediction on random microarray data with 10-fold CV resulted in AUC values as low as 0.30. They refer to AUC values below the value of true performance as being pessimistically biased [80]. The authors argue that methods for reusing samples, like CV, may result in such estimates due to a difference in the number of samples belonging to each class in the train and test datasets, respectively. Although none below 0.50, Parker et al. obtained consistent pessimistic bias for non-random datasets as well, in particular when the number of samples were low (< 25). In conclusion, they found that performing a stratified version of CV that ensures greater balance between the train and test set may improve results.

Microarray data often contains few samples (< 100) and can be difficult to separate into distinct classes [80], similarly to what is often the case for radiomics. On this basis, the arguments made by Parker et al. may be valid for findings presented in this thesis.

As evident from Chapter 7, the 5-fold CV schema used here was in fact stratified. However, when the number of samples is low and if it is not possible to achieve the exact ratio of negative and positive samples in each fold, a small difference in *numbers* may result in a larger change in *percentage*, thereby causing the folds to appear different. With few samples, each sample has a larger say. Furthermore, with small number of samples and large number of features the dataset may appear more or less random. Thus, the small sample size may have caused pessimistic bias in the performance estimates, with values below 50% occurring due to the apparent randomness. Performing machine learning with small sample sizes is further discussed in Section 11.3.

Perlich et al. [81] found that, for datasets with small numbers of positive samples, values below 50% occurred even though stratified CV was performed. This was seen to a greater extent when the number of folds were high. Without going into further detail, this may be of relevance considering all four combinations of samples and response variable evaluated in this thesis were indeed imbalanced, with a majority of negative samples. Dataset imbalance is discussed in Section 11.4.

HYPER-PARAMETER OPTIMIZATION The tree-based classifiers LGBM and DT tended to over-fit, with train scores up to 100% AUC. According to Scikit-learn documentation [76], DT classifiers are prone to over-fitting when the number of features is high. Hyper-parameters controlling tree depth, number of leaves as well as number of samples required to be at a leaf should be carefully tuned in order to avoid the model becoming too complex [75]. If time allowed, a greater effort should have been made to better choose the value ranges from which hyper-parameter configurations were sampled. A greater number of configurations could also have been evaluated.

Selection of relevant features is an important step in establishing prediction models. As evident from Table 7.3, the range corresponding to the number of features the model could select, k, were set to $1-15$ and $1-9$ when training models with data from all 81 patients and just the 35 patients in the nCRT cohort, respectively. Low values were chosen in an attempt to avoid over-fitting. Gillies et al. [5] argue that for binary classification models, about 10 patients are required for each feature.

However, requiring the algorithms to identify such small subsets of features when the training data contains many features relative to the number of samples may be problematic. A discussion on this topic is included in Section 11.3. This issue becomes even more apparent when several features represent overlapping information, as discussed in Sections 11.6 and 11.7.

11.2 THE PREDICTIVE VALUE OF TEXTURE FEATURES

It is known that T-stage of a tumor at diagnosis is predictive of rates like overall and disease-free survival [82] [83]. As evident from Chapter 3, factors like size and invasiveness of the tumor contribute in determining the T-stage. Accordingly, although shape features extracted in a radiomics analysis may provide more precise descriptions of shape-related tumor attributes than what can be obtained with conventional image analysis, many shape features are already more or less known predictors of survival. These may be volume, surface area and various diameters. So in order for the field of radiomics to be of added value in prediction of survival, features not readily obtained from conventional image analysis must prove themselves relevant [84]. Texture and first-order statistical features potentially describing tumor heterogeneity [3], are of interest in this regard.

From Chapters 8 - 10, it appears that several selected features were in fact texture features. Across experiments, several models were based solely on texture features. This is evident from Tables A.1 - A.4. The high representation of texture features is also evident from figures showing selection rates.

To further investigate whether models could perform well based on texture features alone, experiment 5 was performed. Across models, performance was quite similar to that in experiment 1 and 3, with test scores of about 60% AUC and corresponding mean test std values slightly above 10%.

As observed in Chapter 8, the small area high gray level emphasis feature from the GLSZM appeared to be of importance in prediction of PFS.

Other studies within MRI-based radiomics report of similar findings, i.e. predicting survival from subsets with mainly texture features [85] [86]. Kim et al. [85] calculated a radiomic score based on five selected features from the GLDM, GLSZM or GLCM found to be of relevance in prediction of disease-free survival. Note that features were derived from T2WIs and T1WIs, and that filters were applied prior to extraction. The authors of [86] extracted features from T2WIs and contrast-enhanced T1WIs from patients with non-metastatic nasopharyngeal carcinoma. 20 features were selected as predictive of PFS, with which they performed a statistical survival analysis. One shape feature, elongation, was selected, while the remaining described texture and the first-order intensity histogram. Filters were applied prior to extraction.

11.3 LEARNING WITH SMALL SAMPLE SIZES

In this thesis, prediction models were trained on datasets of different sizes both with respect to the number of samples, n, and number of features, m. When considering all patients and the nCRT cohort, n was equal to 81 and 35, respectively. The number of features per patient varied from $m = 28$ when examining shape features only in experiment 5, to $m = 772$ when deriving features from T2WIs and all seven DWIs in experiment 2.

In radiomics analysis, the number of extracted features and corresponding dimensionality are often high [12]. Accordingly, there is a demand for large samples sizes to obtain predictions with statistical significance [12]. Within the field of medicine however, it is often difficult to obtain datasets with a large number of samples due to high costs and time-consuming work typically associated with the process of data collection [87]. Such high m low n situations may be prone to issues like over-fitting [12], and models misinterpreting noise in the training data as essential patterns [87]. This results in poor generalization and low degree of reproducibility of results [88].

The authors of [88] define a dataset as *wide* if $\frac{m}{n} > 10$. They generated random datasets and showed that predictions measured in accuracy obtained with models trained on wide datasets were largely influenced by chance. Furthermore, considering feature selection, they argue the following. With greater number of features and lower number of samples, the probability of mistaking features that simply correlated with a class by chance as actually relevant, increases [88].

As mentioned in Section 5.3, K-fold CV is often used to increase the reliability of results without the need to fully isolate a test set, something which might be problematic for small datasets. However, the authors of [87] found that performance results from K-fold CV tend to be biased and falsely high due to lack of ability to avoid over-fitting, especially for small datasets but also for larger ones. Nested CV on the other hand ensured more trustworthy evaluations of model performance also for small datasets due to the isolated test set [87]. This finding is supported by documentation provided for the Scikit-learn software [89].

Estimates for model performance obtained in this thesis may be overly optimistic, when viewed in light of the discussion above.

For models predicting PFS based on radiomic data from all 81 patients, overall higher AUC values were obtained when including features from all seven DWIs without significant increase in test std. Evaluation of all $m = 772$ features per patient correspond more or less to a wide dataset situation according to [88], with $\frac{m}{n} \sim 9.5$. For the nCRT cohort with $n = 35$, this fraction would be $\frac{m}{n} \sim 22.1$. Accordingly, results from experiment 2 for the nCRT cohort were not included due to the presumably low reliability of results.

Estimates of model performance obtained for the nCRT cohort in prediction of PFS, TRG and ypT may to some extent be overly optimistic. According to the discussion above, the most reliable results are obtained from experiment 1, in which $n = 107$. As evident from Chapter 8, predictions of PFS for the nCRT cohort were associated with mean std values about 20%. When considering all 81 patients, std values were lower, yet still often above 10%.

Surprisingly, lower train than test scores were observed for some models, both considering all patients and the nCRT cohort. In particular, this occurred for models predicting ypT, where eight combinations of feature selector and classifier in total for both experiments had lower train scores. When this occurred, associated mean values for std were often high, thereby adding uncertainty to the obtained scores. Moreover, across models and experiments, the difference between train and test scores were consistently within the mean train std value for the particular model.

The overall high test std values, as well as lower train than test scores, may again reflect the fact that when establishing CV folds from relatively few samples, each sample obtains a greater impact on the final score.

11.4 DATASET IMBALANCE

In this thesis, all four combinations of samples and endpoints were imbalanced with a majority in the negatively labeled class, as evident from Table 7.1. This may contribute further to causing overly optimistic performance estimates when the metric is AUC.

As evident from Section 5.3, the AUC score is based on the true positive rate (TPR) and the false positive rate (FPR). While the former focuses on the positive, i.e. minority class, the FPR evaluates the models ability to classify negative samples correctly. In contrast, the F1 score is based on recall, equal to the TPR, and precision. The latter also depends on the number of true positives, thereby rendering the F1 score a more reliable measure for models trained with datasets containing smaller number of positive samples [66].

Considering the nCRT cohort, higher AUC values were obtained for models predicting TRG and ypT than PFS. Moreover, mean test std values for models predicting TRG were overall lower than for models predicting PFS. Results obtained from models predicting ypT included several models with scores above 90% AUC and low associated std, but also models with values for std above 20%. In short, the larger the dataset imbalance, the more optimistic the performance estimates tended to get.

Repeating experiments 1 and 3 on the nCRT dataset with ypT as endpoint using the F1 metric may have provided more trustworthy results. However, with mean values for test std now overall very high, no reliable insight can be drawn from these models. Note that ranges from which hyper-parameter configurations were samples could have been optimized further. In any case, the dataset consisting of the nCRT cohort with ypT as endpoint is both small and severely imbalanced, i.e. not ideal for training prediction models.

When trained with imbalanced datasets, classification-based prediction models tend to have higher degree of misclassification for the minority, or positive, class [90]. This may again contribute to explaining performance estimates obtained for the nCRT cohort: when the number of positive samples decreases, from 14 when predicting TRG to 8 when predicting ypT, the number of misclassifications (as apparent from higher performance) decreased accordingly. It would here be of interest to display the confusion matrix for each model with numbers of true/false negatives and positives, as explained in Section 5.3.

Methods for decreasing class imbalance exist, some focused on undersampling the majority class, while others, like the Synthetic minority oversampling technique (SMOTE), oversample the minority class [90]. However, when the goal is to establish a model with the ability to make accurate predictions for new, unlabeled data, it is essential that the data used for training were representative of the actual situation [65]. If the problem is in fact imbalanced with respect to class, then so should the datasets used for both training and testing of the prediction model [65].

Still, in many problems for which classification is used, one type of misclassification (false positive or false negative) is often worse than the other [65] [91], like classifying a cancerous tumor as benign [65]. Then, a cost or penalty for misclassification can be assigned to this class [91]. This may be a good strategy for imbalanced datasets in which correct classification of the minority class is most important [91].

The choice of metric might also contribute to obtaining performance estimates in accordance with what is most important, as evident from the discussion above. If misclassifying poor response to nCRT (i.e. $y = 0$) is worse than misclassifying good response, and if poor response being the majority class in fact represents reality, then AUC may be a good metric. Avoiding nCRT due to a prediction of no response when the patient would indeed benefit from such treatment would be problematic. Furthermore, for patients predicted to respond well or even achieve complete response a "wait-and-see" approach may be chosen instead of performing surgery right away [9] [26]. Still, considering CRT being invasive [92], correctly classifying good response should also be given priority.

While it is obvious that any ideal binary prediction model would perform well considering both classes, matters like those mentioned should be taken into account when establishing, evaluating and improving algorithms for use in prediction of both PFS and response to nCRT.

11.5 IN SUPPORT OF HIGH PERFORMANCE ESTIMATES

In the preceding sections it was argued that the overall higher AUC scores obtained for models considering the nCRT cohort were overly optimistic and potentially misleading. In the current section, aspects and findings that may add support to such high performance estimates are outlined. Still, the overall high values for test std remain problematic.

According to the therapeutic guidelines mentioned in Chapter 3 and the clinical data presented in Chapter 6.1, patients selected to undergo CRT prior to surgery typically have tumors staged as T4 (i.e. locally advanced) or T3 with positive lymph nodes. The subset of patients with such high-risk disease may then form a more homogeneous group than all 81 patients. Accordingly, it is not unlikely that the radiomic features of *actual* relevance are more readily identified by the machine learning algorithms. Relevant features are those that are found to differ more or less consistently between patients belonging to each respective class, and these may manifest themselves better when other tumor attributes are more similar between classes.

Several studies predicting response to nCRT for LARC patients report of high performance estimates [9] [26] [27]. The authors of [9] obtained AUC values of 90.17% and 89.72% for the train and validation dataset, respectively, when predicting good response to nCRT based on radiomic features derived from T2WIs for 134 patients. Note that the Dowrak/Rödel TRG system was used, in which TRG 0-2 and 3-4 correspond to poor and good tumor response to nCRT, respectively [25] [9].

The authors of [26] obtained AUC values of 0.89 when predicting good response to nCRT, with good and no/poor tumor response defined as in this thesis. The patient cohort consisted of 48 patients, 31 responding well and 17 responding poorly. Features were extracted from T1WIs, T2WIs, DWIs and dynamic contrast-enhanced images, and the AUC value found to be higher when including all features compared to when considering features from each scan type alone.

Horvat et al. [27] predicted pathological complete response (pCR) to nCRT with AUC value of 0.93 based on radiomic features derived from T2WIs. 114 patients with rectal cancer were evaluated, of which 21 obtained pCR. Note that pCR corresponds to TRG = 0 according to the AJCC 7th. edition [24].

In order for predictive models established from radiomic data to be of use in clinical decision making, validation with unseen datasets must be performed [12]. Zhenwei et al. [93] compared performance of a radiomic signature developed for non-small cell lung cancer at one institution, with the performance of this signature obtained at a new, independent institution. Performance, measured as ability to predict overall survival, across institutions was found to be similar.

11.6 CORRELATED FEATURES

When revisiting the definitions of radiomic features outlined in Chapter 4 several features appear to provide overlapping information. In particular, features derived from texture matrices may correlate.

Some features are calculated from more than one matrix, like gray level non-uniformity from the GLSZM, GLRLM and GLDM. Furthermore, features like small/large area emphasis, short/long run emphasis and small/large dependence emphasis all quantify texture coarseness. Accordingly, these features would be positively correlated. This applies to several other features calculated from the GLSZM, GLRLM and GLDM, as evident from Tables 4.4, 4.5 and 4.7.

Furthermore, features calculated from the same matrix may be descriptive of the same image attributes and thereby correlate. Examples are gray level non-uniformity and gray level variance, as well as busyness and complexity from the NGTDM. Compared to the latter two, coarseness from the NGTDM may to some extent be negatively correlated. Various other similar examples exists. This demonstrates the need for removal of redundant features.

REMOVAL OF REDUNDANT FEATURES In many radiomic studies, feature selection is performed in more than one step [74] [94] [5] [29]. The initial feature selection step often involves dimensionality reduction with corresponding removal of correlated and redundant features, often achieved by use of a cluster analysis [94] [5] [29]. This is an unsupervised approach in which highly correlated features are grouped together, thereby yielding low similarity between clusters [29]. Each cluster can then be averaged into a single value [94], or represented by selecting one [29] or a few [5] features. Leger et al. [94] extracted 1610 features per patient, from which they established 229 clusters.

A second method for excluding redundant features is the principal component analysis (PCA) [29]. It involves selection of a minimum number of features, i.e. the principal components, that capture the variation in the dataset and eliminate all others [29]. Huynh et al. [95] reduced the number of radiomic features from 855 to 12 in a process involving PCA.

Although CT- and not MRI-based radiomics were performed in both [94] and [95], the significant decrease in numbers illustrate the high degree of correlation between radiomic features.

The authors of [96] argue that when feature selection methods like Lasso [44] are used with datasets containing correlated features, a more or less random selection of features from the groups or clusters found to be relevant occurs. Accordingly, if slight changes are introduced in the training data, the risk of obtaining a completely different set of selected features exists, i.e. the model is unstable [96]. These arguments were made within the field of genomics. However, as mentioned, the high number of potentially correlated features relative to a small number of patients as typical for this field [96], also applies in radiomics, as evident from the discussion in previous sections.

Several findings in this thesis are suggestive of feature correlation, and will be the focus of the ensuing paragraphs.

SELECTION RATES In most experiments, when considering features with rates higher than 0.33, the majority were selected at rates lower than 0.50. This low degree of agreement across models regarding specific, relevant features could indicate that groups of correlated features exist. Accordingly, as explained above, from a group found to be relevant for the endpoint in question (PFS, TRG or ypT), several features would qualify as the representative.

From Figures 8.2 and 8.6 in Chapter 8, higher selection rates are associated with experiment 1 than 3 for all patients as well as the nCRT cohort when predicting PFS. The number of input features is doubled from experiment 1 to experiment 3, with the same 107 features derived from both T2WIs and DWIs (b5) in the latter. This may further contribute to introducing redundant features.

The overall correlation between features derived from the different scan types could potentially be decreased by extracting features from ADC maps rather than from the DWIs themselves. This is done in a number of studies [97] [26].

Selection rates are more similar between experiment 1 and 3 for models predicting TRG or ypT, as evident from Figures 9.2 and 9.5 in Chapter 9. For these models however, but also for the ones predicting PFS in the nCRT cohort, the high number of features and few patients could make it difficult for the machine learning algorithms to identify consistent patterns in the datasets.

The overall high selection rates obtained in experiment 4 as evident from Figure 8.3 are likely influenced by the mere fact that the dataset consisted of 81 samples with only 28 features per patient.

The discussion on feature correlation is continued in the next section, now with the focus being the seemingly low reproducibility.

11.7 REPRODUCIBILITY AND FEATURE ROBUSTNESS

In the field of radiomics, many variables with potentially large impact on obtained results exist [12]. Important areas from which such variables may emerge are image acquisition and reconstruction, intensity binning, and segmentation [12] [14] [98].

For a prediction model to be of clinical use it must provide reliable and stable results, i.e. results must be reproducible [12]. By evaluating reproducibility in view of different aspects, robust features may be identified [14]. It is important to note, however, that performing feature selection strictly according to robustness may result in removal of features with predictive value [14].

Among studies in which evaluation of reproducibility of radiomic results are performed, metrics used to quantify this reproducibility are often intraclass-correlation coefficient (ICC) or concordance correlation coefficient (CCC) [12] [98]. In this thesis, reproducibility was not quantified as such.

An attempt was made to perform a preliminary evaluation of reproducibility. The apparent high degree of feature correlation as discussed in the previous section is likely to have affected the obtained results.

As evident from Chapter 10, reproducibility of the results obtained in experiments 1 and 3 was overall low with respect to voxel resampling, intensity binning and VOI delineation.

In experiments 6, 8 and 10, values for test std were about the same as in experiments 1 and 3, likely influenced by the ratio between samples and features being equal. Comparing performance and selected features by each individual model across experiments however, reproducibility appeared to be low. As evident from Tables 10.1 and 10.2, few features are selected at high rates across experiments, i.e. few features emerge as robust. As argued by Tolosi et al. [96] in Section 11.6, these models may be unstable due to the presence of correlated features.

FIXED HYPER-PARAMETERS In experiments 7, 9 and 11, hyper-parameters for the model consisting of FS selector and the LR classifier were equivalent to as in experiment 3. Comparing each of the three former experiments with the latter, performance was consistently lower (slightly above 50% AUC) and none of the same features were selected. The latter is evident from Table 11.1.

Consider the features selected in experiment 3. As explained in Chapter 4, both run percentage and short run emphasis from the GLRLM quantify texture coarseness. Gray level variance and dependence non-uniformity evaluates variability in intensities and dependencies, respectively.

Consider features selected in experiment 7. Short run low gray level emphasis from the GLRLM represents the fraction of short, low-intensity runs. As evident from Table 4.5, this feature is obtained from the expression for short run emphasis simply by including intensity in the sum as $\frac{1}{i}$. Furthermore, gray level non-uniformity (b5) quantifies variability in intensity values, just like gray level variance.

Similar comparisons could be made between selected features from the remaining experiments.

Experiment 3
Gray level variance, GLRLM
Run percentage (b5), GLRLM
Short run emphasis (b5), GLRLM
Dependence non-uniformity normalized (b5), GLDM
Experiment 7
Small area low gray level emphasis, GLSZM
Short run low gray level emphasis (b5), GLRLM
Busyness (b5), NGTDM
Gray level non-uniformity (b5), GLDM
Experiment 9
Zone entropy, GLSZM
Low gray level run emphasis (b5), GLRLM
Run length non-uniformity (b5), GLRLM
Complexity (b5), NGTDM
Experiment 11
Maximum 2D diameter slice
Elongation (b5)
Run variance (b5), GLRLM
ID normalized (b5), GLCM

Table 11.1: Features selected by the FS selector in combination with LR classifier in experiments 3, 7, 9 and 11.

11.8 TOPICS FOR FURTHER INVESTIGATION

From the results and discussion as presented here and in previous chapters, certain topics have emerged as relevant for further investigation.

First, it would be of interest to repeat the experiments outlined in this thesis after performing an evaluation of feature correlation and corresponding removal of redundant features. Moreover, features could be extracted from ADC-maps rather than from the DWIs themselves. A nested CV scheme [89] could be utilized to evaluate performance. Even with the given relatively small, imbalanced dataset, these modifications may decrease std and potential over-fitting, i.e. render results more reliable. Finally, plotting the confusion matrix would enable a more precise evaluation of performance.

Second, dimensionality reduction and removal of redundant features are likely to facilitate a better evaluation of reproducibility and feature robustness. Implementing software that allows for fixation of hyper-parameters would ease this investigation. Consistently using dataset2, with voxels resampled to a $1 \times 1 \times 1\,\text{mm}^3$ dimension could better promote reproducibility across studies. Due to the many factors with potential impact on results [14] [98], standardization is of importance [5]. In summary, completion of the following three objectives should be part of establishing a radiomics-based model with ability to assist in clinical decision-making: perform an exhaustive evaluation of effects posed on the results by different factors [14], choose approaches for image acquisition, reconstruction and processing that optimally facilitate reproducibility, and report these [5], and finally, evaluate performance with new datasets not used during training.

Lastly, to evaluate the potential prognostic value of specific features, a statistical survival analysis may be performed. Such analysis with radiomic features have previously been done within the field of rectal cancer [99] [100]. Kaplan-Meier and Cox regression are two methods frequently used for this purpose [101]. The former involves estimation of survival curves [102], while the latter allows for a more in-depth analysis of the relationship between features and outcome by taking confounding variables into account [101] [102].

12 CONCLUSION

In this thesis, it was investigated how a radiomics analysis based on T2WIs and DWIs may provide added predictive and prognostic value within the field of rectal cancer. More specifically, an attempt was made to achieve tree main goals. Namely, (1) establish binary predictions models based on radiomic data, (2) investigate whether texture and first-order features in particular proved valuable, and (3) evaluate reproducibility of results.

To this end, eleven experiments were designed and performed. A machine learning approach was chosen, realized by the Python-based Biorad program as available from https://github.com/ahmedalbuni/biorad. Performance of four feature selector algorithms in combination with six classifiers were evaluated. Three endpoints were used: PFS, TRG and ypT, the latter two representing response to nCRT.

A few trends were apparent across experiments. Values for test std were high, likely influenced by the relatively small sample sizes (81 and 35). The presence of redundant features and the need for investigation of feature correlation prior to performing selection and classification using Biorad was apparent. Texture features were well represented among those selected. Several models predicted PFS, TRG or ypT purely from texture and first-order features, some of which achieved test AUC scores of 62.2 ± 5.9% (FS and DT, all patients, RV = PFS), 73.0 ± 10.8% (FS and DT, nCRT cohort, RV = PFS), and 76.7 ± 4.6% (MI and DT, nCRT cohort, RV = TRG). Results from experiment 5, in which only texture features were analysed, added further support to the finding that features describing image texture were indeed of predictive value.

PREDICTING PFS FOR ALL PATIENTS The FS selector and DT classifier performed relatively well, predicting PFS with AUC test scores of 62.2 ± 5.9%, 68.6 ± 8.9% and 60.6 ± 11.0% in experiments 1, 2 and 3, respectively. Overall, without significant increase in std, test scores improved from experiment 1 to experiment 3, and again in experiment 2. Considering that dimensionality increased correspondingly, over-fitting is likely to be part of the explanation. Reproducibility with respect to voxel resampling, intensity binning and VOI delineation was poor, in particular considering selected features. It was argued that this may be partly due to the presence of correlated features.

Small area high gray level emphasis from the GLSZM appeared to be of relevance in prediction of PFS both for all patients and the nCRT cohort. Note that this may be subject to change if correlation is evaluated and redundant features removed.

PREDICTING PFS FOR THE NCRT COHORT Values for mean test std were higher than when predicting PFS for all patients. Considering selection rates, the same trend as when predicting PFS for all patients was apparent, with lower and more similar rates in experiment 3. Feature correlation was proposed as a feasible explanation.

The number of samples relative to the number of features were low across experiments, which may give rise to over-fitting and difficulties in identifying relevant features. This also applies to models predicting TRG and ypT.

PREDICTING TRG AND YPT Models predicting TRG and ypT typically achieved test scores of about 80% and 90% AUC, respectively. This was more or less in accordance with reported findings from literature. Still, some models clearly over-fit. In prediction of ypT, performance measured in terms of the F1 score was lower, with high associated test std values. Given the dataset and software used in this thesis, performance estimates in prediction of response to nCRT may be most reliably obtained using TRG as endpoint, due to a more severe class imbalance associated with ypT.

BIBLIOGRAPHY

[1] International Agency for Research on Cancer. World Health Organization. *Global Cancer Observatory*. 2018. URL: `http://gco.iarc.fr/` (visited on 02/20/2020).

[2] Natally Horvat, Camilla C. T. Rocha, Brunna C. Oliveira, Iva Petkovska, and Marc J. Gollub. "MRI of Rectal Cancer: Tumor Staging, Imaging Techniques, and Management." In: *RadioGraphics* 39.2 (2019), pp. 367 –387.

[3] S. Alobaidli, S. McQuaid, C. South, V. Prakash, P. Evans, and A. Nisbet. "The Role of Texture Analysis in Imaging as an Outcome Predictor and Potential Tool in Radiotherapy Treatment Planning." In: *The British Journal of Radiology* 87.1042 (2014).

[4] N. Just. "Improving Tumour Heterogeneity MRI Assessment with Histograms." In: *British Journal of Cancer* 111.12 (2014), pp. 2205 –2213.

[5] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. "Radiomics: Images are More Than Pictures, They are Data." In: *Radiology* 278.2 (2016), pp. 563 –577.

[6] A. Zwanenburg, S. Leger, M. Valliéres, and S. Löck. "Image Biomarker Standardisation Initiative." In: *arXiv preprint arXiv:1612.07003* (2016).

[7] J. J. M. van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype." In: *Cancer Research* 77.21 (2017), e104 –e107.

[8] Pyradiomics Community. *Pyradiomics Documentation*. Release v3.0.post2+g896682d. 2020.

[9] Xiaoping Yi et al. "MRI-based Radiomics Predicts Tumor Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer." In: *Frontiers in Oncology* 9.552 (2019).

[10] Xiaolu Ma, Fu Shen, Yan Jia, Yuwei Xia, Qiuha Li, and Jianping Lu. "MRI-based Radiomics of Rectal Cancer: Preoperative Assessment of the Pathological Features." In: *BMC Medical Imaging* 19.86 (2019).

[11] Xuezhi Zhou, Yongju Yi, Zhenyu Liu, Wuteng Cao, Bingjia Lai, Kai Sun, Longfei Li, Zhiyang Zho, Yanqiu Feng, and Jie Tian. "Radiomics-Based Pretherapeutic Prediction of Non-response to Neoadjuvant Therapy in Locally Advanced Rectal Cancer." In: *Annals of Surgical Oncology* 26 (2019), pp. 1676 –1684.

[12] Ji Eun Park, Seo Young Park, Hwa Jung Kim, and Kim Ho Sung. "Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives." In: *Korean Journal of Radiology* 20.7 (2019), pp. 1124 –1137.

[13] Huan Liu and Hiroshi Motoda. "Data Processing and Knowledge Discovery in Databases." In: *Feature Selection for Knowledge Discovery and Data Mining*. Springer, Boston, MA, 1998, pp. 1 –15.

[14] Alex Zwanenburg. "Radiomics in Nuclear Medicine: Robustness, Reproducibility, Standarization, and How to Avoid Data Analysis Traps and Replication Crisis." In: *European Journal of Nuclear Medicine and Molecular Imaging* 46 (2019), pp. 2638 –2655.

[15] Andreas Abildgaard. *MR for Radiografer og Radiologer, Fysikk og Fysiologi*. 1st ed. Oslo, Norway: Universitetsforlaget, 2016.

[16] Donald W. McRobbie, Elizabeth A. Moore, Martin J. Graves, and Martin R. Prince. *MRI from Picture to Proton*. 2nd ed. Cambridge, UK: Cambridge University Press, 2006.

[17] Gabriela Castellano, Leonardo F. Bonilha, Limin Li, and Fernando Cendes. "Texture Analysis of Medical Images." In: *Clinical Radiology* 59.12 (2004), pp. 1061 –1069.

[18] Roland Bammer. "Basic principles of diffusion-weighted imaging." In: *European Journal of Radiology* 45.3 (2003), pp. 169 – 184.

[19] Bachir Taouli and Dow-Mu Koh. "Diffusion-Weighted MR Imaging of the Liver." In: *Radiology* 254.1 (2010), pp. 47 –66.

[20] Patric Hagmann, Paeder Jonasson Lisa ad Maeder, Jean-Philippe Thiran Thiran, Van J. Wedeen, and Reto Meuli. "Understanding Diffusion MR Imaging Techniques: From Scalar Diffusion-weighted Imaging to Diffusion Tensor Imaging and Beyond." In: *RadioGraphics* 26 (2006), pp. 205 –223.

[21] Kathrine R. Redalen. *Functional MRI of Hypoxia-mediated Rectal Cancer Aggressiveness (OxyTarget)*. 2013. URL: https://clinicaltrials.gov/ct2/show/NCT01816607 (visited on 04/20/2020).

[22] Slawomir Marecik, John Park, and Leela M. Prasad. "Rectal Anatomy: Clinical Perspective." In: *Rectal Cancer*. Ed. by George J. Chang. Springer, Cham, 2018, pp. 1–23.

[23] National Cancer Institute. National Institutes of Health. *Rectal Cancer Treatment - Patient Version*. 2020. URL: https://www.cancer.gov/types/colorectal/patient/rectal-treatment-pdq#_111 (visited on 02/20/2020).

[24] American Joint Committee on Cancer. *AJCC Cancer Staging Manual*. 7th Edition. 2015.

[25] Soo H Kim, Hee J Chang, Dae Y. Kim, Ji W. Park, Ji Y. Baerk, Sun Y. Kim, Sung C. Park, Jae H. Oh, and Byung-Ho Nam. "What is the Ideal Tumor Regression Grading System in Rectal Cancer Patients After Preoperative Chemoradiotherapy?" In: *Cancer Research and Treatment* 48.3 (2016), pp. 998 –1009.

[26] Ke Nie, Liming Shi, Qin Chen, Xi Hu, Salma K. Jabbour, Ning Yue, Tianye Niu, and Xiaonan Sun. "Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI." In: *Clinical Cancer Research* 22.21 (2016), pp. 5256 –5264.

[27] Natally Horvat, Harini Veeraraghavan, Monika Khan, Ivana Blazic, Junting Zheng, Marinela Capanu, Evis Sala, Julio Garcis-Aguilar, Marc J. Gollub, and Iva Petkovska. "MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy." In: *Radiology* 287.3 (2018), pp. 833 –843.

[28] Maria Petrou. "Texture in Biomedical Images." In: *Biomedical Image Processing*. Ed. by Thomas Martin Deserno. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 157–176.

[29] Stefania Rizzo, Francesca Botta, Sara Raimondi, Daniela Origgi, Christina Fanciullo, Alessio G. Morganti, and Massimo Bellomi. "Radiomics: the Facts and the Challenges of Image Analysis." In: *European Radiology Experimental* 2.36 (2018).

[30] Gopinath Rebala, Ajay Ravi, and Sanjay Churiwala. "Regressions." In: *An Introduction to Machine Learning*. Springer Nature Switzerland, 2019, pp. 25 –40.

[31] Gopinath Rebala, Ajay Ravi, and Sanjay Churiwala. "Learning Models." In: *An Introduction to Machine Learning*. Springer Nature Switzerland, 2019, pp. 19 –23.

[32] Max Kuhn and Kjell Johnson. "Over-Fitting and Model Tuning." In: *Applied Predictive Modeling*. Springer, New York, NY, 2013, pp. 61 –92.

[33] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Overview of Supervised Learning." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 9 –41.

[34] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Model Assessment and Selection." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 219 –259.

[35] Verónica Bolón-Canedo, Noelia Sánches-Marino, and Ampero Alonso-Betanzos. "Foundations of Feature Selection." In: *Feature Selection for High-Dimensional Data*. Springer, 2015, pp. 13 –28.

[36] Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. "Relief-based feature selection: Introduction and review." In: *Journal of Biomedical Informatics* 85 (2018), pp. 189 –203.

[37] Scikit-learn. *Feature Selection*. 2007. URL: https://scikit-learn.org/stable/modules/feature_selection.html (visited on 03/30/2020).

[38] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825 –2830.

[39] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating Mutual Information." In: *Physical Review E.* 69.6 (2004).

[40] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy." In: *IIEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1226 –1238.

[41] Xiaofei He, Deng Cai, and Partha Niyogi. "Laplacian Score for Feature Selection." In: *NIPS* (2006).

[42] Huan Liu and Hiroshi Motoda. "Feature Selection Methods." In: *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Springer, Boston, MA, 1998, pp. 73 –95.

[43] A. J. Van der Kooij and J. J. Meulman. "Regularization with Ridge penalties, the Lasso, and the Elastic Net for Regression with Optimal Scaling Transformations." In: *Prediction Accuracy and Stability of Regression with Optimal Scaling Transformations (Doctoral Thesis)*. 2006, pp. 65 –90.

[44] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Linear Methods for Regression." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 43 –99.

[45] Juliana Tolles and William J. Meurer. "Logistic Regression, Relating Patient Characteristics to Outcome." In: *JAMA Guide to Statistics and Methods* 316.5 (2016), pp. 533 –534.

[46] Yan-Qin Bai and Kai-Ji Shen. "Alternating Direction Method of Multipliers for l1-l2-Regularized Logistic Regression Methods." In: *Journal of the Operations Research Society of China* 4 (2016), pp. 243 –253.

[47] Scikit-learn. *Linear Models*. 2007. URL: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares (visited on 04/05/2020).

[48] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Linear Methods for Classification." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 103 –137.

[49] Lance M. Kaplan, Yaakov Bar-Shalom, and William D. Blair. "Assignment Costs for Multiple Sensor Track-to-Track Association." In: *IEEE Transactions on Aerospace and Electronic Systems* 44.2 (2008), pp. 655 –677.

[50] Shan Suthaharan. "Support Vector Machine." In: *Machine Learning Models and Algorithms for Big Data Classification, Thinking with Examples for Effective Learning*. Ed. by Ramesh Sharda and Stefan Voss. Springer, 2016, pp. 207 –235.

[51] Catalin Stoean and Ruxandra Stoean. "Support Vector Learning and Optimization." In: *Support Vector Machines and Evolutionary Algorithms for Classification*. Ed. by J. Kacprzyk and L. Jain. Springer, 2014, pp. 7 –25.

[52] Scikit-learn. *Support Vector Machines*. 2007. URL: https://scikit-learn.org/stable/modules/svm.html#svm-kernels (visited on 05/21/2020).

[53] Gopinath Rebala, Ajay Ravi, and Sanjay Churiwala. "Random Forests." In: *An Introduction to Machine Learning*. Springer Nature Switzerland, 2019, pp. 77 –94.

[54] V Kishore Ayyadevara. "Decision Tree." In: *Pro Machine Learning Agorithms*. Ed. by Celestine Suresh, Matthew Moodie, and Divya Modi. Apress, 2018, pp. 71 –103.

[55] Trevr Hastie, Robert Tibshirani, and Jerome Friedman. "Additive Models, Trees, and Related Methods." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 295 –336.

[56] Benjawan Srisura. "Impurity Measurement in Selecting Decision Node Tree that Tolerate Noisy Cases." In: *Recent Advances in Information and Communication Technology 2017, Proceedings of the 13th International Conference on Computing and Information Technology*. Ed. by Phayung Meesad, Sunantha Sodsee, and Herwig Unger. Springer, 2017, pp. 13 –21.

[57] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Random Forests." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 587 –604.

[58] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Model Inference and Averaging." In: *The Elements of Statistical Learning*. Springer, 2017, pp. 261 –294.

[59] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." In: *Machine Learning* 63.1 (2006), pp. 3 –42.

[60] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “Boosting and Additive Trees.” In: *The Elements of Statistical Learning*. Springer, 2017, pp. 337 –387.

[61] K. Ayyadevara. “Gradient Boosting Machine.” In: *Pro Machine Learning Agorithms*. Ed. by C. Suresh, M. Moodie, and D. Modi. Apress, 2018, pp. 117 –134.

[62] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” In: *31st Conference on Neural Information Processing Systems (NIPS)* (2017).

[63] James Bergstra and Yoshua Bengio. “Random Search for Hyperparameter Optimization.” In: *Journal of Machine Learning Research* 13 (2012), pp. 281 –205.

[64] Tom Fawcett. “An introduction to ROC analysis.” In: *Pattern Recognition Letters* 27.8 (2006), pp. 861 –874.

[65] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. “Foundations on Imbalanced Classification.” In: *Learning from Imbalanced Data Sets*. Springer, 2018, pp. 19 –46.

[66] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. “Performance Measures.” In: *Learning from Imbalanced Data Sets*. Springer, 2018, pp. 47 –61.

[67] Acredit. *The OxyTarget study*. 2013. URL: https://www.acredit.no/the-oxytarget-study/ (visited on 04/20/2020).

[68] Kine M. Bakke et al. “Comparison of Intravoxel Incoherent Motion Imaging and Multiecho Dynamic Contrast-Based MRI in Rectal Cancer.” In: *Journal of Magnetic Resonance Imaging* 50.4 (2019), pp. 1114 –1124.

[69] Helsedirektoratet. *Nasjonalt Handlingsprogram med Retningslinjer for Diagnostikk, Behandling og Oppfølging av Kreft i Tykktarm og Endetarm*. 2019.

[70] Klaus D. Toennies. “Registration and Normalizaton.” In: *Guide to Medical Image Analysis, Advances in Computer Vision and Pattern Recognition*. Springer, 2017, pp. 361 –404.

[71] Kasper Marstal. *SimpleElastix Documentation*. Release 0.1. 2018.

[72] Michael Schwier, Joost van Griethuysen, Mark G. Vangel, Steve Pieper, Sharon Peled, Clare Tempany, Hugo J. W. L. Aerts, Ron Kikinis, Fiona M. Fennessy, and Andriy Fedorov. “Repeatability of Multiparametric Prostate MRI Radiomics Features.” In: *Scientific Reports* 9 (2019).

[73] Florent Tixier, Mathieu Hatt, Cathrine C. Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET." In: *Journal of Nuclear Medicine* 53.5 (2012), pp. 693 –700.

[74] Geir Severin Rakh Elvatun Langberg. "Searching for Biomarkers of Disease-Free Survival in Head and Neck Cancers using PET/CT Radiomics (Master thesis)." In: (2019).

[75] Microsoft Corporation. *LightGBM*. Release v. 2.3.2. 2020.

[76] Scikit-learn. *Decision Trees*. 2007. URL: https://scikit-learn.org/stable/modules/tree.html#tree (visited on 05/21/2020).

[77] Scikit-learn. *Ensemble Methods*. 2007. URL: https://scikit-learn.org/stable/modules/ensemble.html#forest (visited on 06/20/2020).

[78] Scikit-learn. *Tuning the hyper-parameters of an estimator*. 2007. URL: https://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-search (visited on 05/19/2020).

[79] Scikit-learn. *Cross-validation: Evaluating Estimator Performance*. 2007. URL: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation (visited on 05/12/2020).

[80] Brian Parker, Simon Günter, and Justin Bedo. "Stratification Bias in Low Signal Microarray Studies." In: *BMC Bioinformatics* 8.326 (2007).

[81] Claudia Perlich and Grzegorz Swirszcz. "On Cross-Validation and Stacking: Building Seemingly Predictive Models on Random Data." In: *ACM SIGKDD Explorations Newsletter* 12.2 (2010), pp. 11 –15.

[82] S. McPhail, S. Johnson, D. Greenberg, M. Peake, and B. Rous. "Stage at Diagnosis and Early Mortality from Cancer in England." In: *British Journal of Cancer* 112 (2015), pp. 108 –115.

[83] Leonard L. Gunderson, Matthew Callister, Robert Marschke, Tonia Young-Fadok, Jacques Heppell, and Jonathan Efron. "Stratification of Rectal Cancer Stage for Selection of Postoperative Chemoradiotherapy: Current status." In: *Gastrointerstinal Cancer Research* 2.1 (2008), pp. 25 –33.

[84] Philippe Lambin et al. "Radiomics: Extracting more Information from Medical Images using Advanced Feature Analysis." In: *European Journal of Cancer* 48.4 (2012), pp. 441–446.

[85] Sungwon Kim, Min Jung Kim, Eun-Kyung Kim, Jung Hyun Yoon, and Vivial Youngjean Park. "MRI Radiomic Features: Association with Disease-Free Survival in patients with Triple-Negative Breast Cancer." In: *Scientific Reports* 10.3750 (2020).

[86] Hesong Shen et al. “Predicting Progression-Free Survival Using MRI-Based Radiomics for Patients With Nonmetastatic Nasopharyngeal Carcinoma.” In: *Frontiers in Oncology* 10.618 (2020).

[87] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. “Machine Learning Algorithm Validation with a Limited Samples Size.” In: *PLoS ONE* 14.11 (2019).

[88] Abdel Aziz Taha, Alexandros Bampoulidis, and Mihai Lupu. “Chance Influence in Datasets with a Large Number of Features.” In: *Data Science - Analytics and Applications, Proceedings of the 2nd International Data Science Conference - iDSC2019*. Ed. by Peter Haber, Thomas Lampoltshammer, and Manfred Mayr. Springer, 2019, pp. 21 –26.

[89] Scikit-learn. *Nested versus non-nested cross-validation*. 2007. URL: https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html (visited on 05/10/2020).

[90] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. “Data Level Preprocssing Methods.” In: *Learning from Imbalanced Data Sets*. Springer, 2018, pp. 79 –121.

[91] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. “Cost-Sensitive Learning.” In: *Learning from Imbalanced Data Sets*. Springer, 2018, pp. 63 –78.

[92] Cancer Research UK. *Side-effects of Chemoradiotherapy*. 2018. URL: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/treatment/treatment-rectal/chemoradiotherapy/side-effects-chemoradiotherapy (visited on 06/14/2020).

[93] Zhenwei Shi et al. “Distributed Radiomics as a Signature Validation Study Using the Personal Health Train Infrastructure.” In: *Scientific Data* 6.218 (2019).

[94] Stefan Leger et al. “A Comparative Study of Machine Learning Methods for Time-to-Even Survival Data for Radiomics Risk Modelling.” In: *Scientific Reports* 7.13206 (2017).

[95] Elizabeth Huynh, Thibaud Coroller, Vivek Narayan, Vishesh Agraqal, Ying Hou, John Romano, Idalid Franco, Raymond H. Mak, and Hugo J.W.L. Aerts. “CT-based Radiomic Analysis of Stereotactic Body Radiation Therapy Patients with Lung Cancer.” In: *Radiotherapy and Oncology* 120.2 (2015), pp. 258 –266.

[96] Laura Tolosi and Thomas Lengauer. “Classification with Correlated Features: Unreliability of Feature Ranking and Solutions.” In: *Bioinformatics* 27.14 (2011), pp. 1986 –1994.

[97] Alberto Traverso, Michal Kazmierski, Zhenwei Shi, Petros Kalendralis, Mattea Welch, Henrik D. Nissen, David Jaffray, Andre Dekker, and Leonard Wee. "Stability of Radiomic Features of Apparent Diffusion Coefficient (ADC) Maps for Locally Advanced Rectal Cancer in Response to Image Pre-Processing." In: *Physica Medica* 61 (2019), pp. 44 –51.

[98] Qihua Li, Hongmin Bai, Yinsheng Chen, Qiuchang Sun, Lei Liu, Sijie Zhou, Guoliang Wang, Chaofeng Liang, and Zhi-Cheng Li. "A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme." In: *Scientific Reports* 7 (2017).

[99] Seung Hyuck Jeon et al. "Delta-Radiomics Signature Predicts Treatment Outcomes After Preoperative Chemoradiotherapy and Surgery in Rectal Cancer." In: *Radiation Oncology* 14.43 (2019).

[100] Nicola Dinapoli et al. "Radiomics for Rectal Cancer." In: *Translational Cancer Research* 5.4 (2016), pp. 424 –431.

[101] Samar Abd ElHafeez, Claudia Torino, Graziella D'Arrigo, Davide Bolignano, Fabio Provenzano, Francesco Mattace-Raso, Carmine Zoccali, and Giovanni Tripepi. "An Overview on Standard Statistical Methods for Assessing Exposure-Outcome Link in Survival Analysis (Part 2): the Kaplan-Meier Analysis and the Cox Regression Methods." In: *Aging Clinical and Experimental research* 24.3 (2012), pp. 203 –206.

[102] Viv Bewick, Liz Cheek, and Jonathan Ball. "Statistics Review 12: Survival Analysis." In: *Critical care* 8.5 (2004), pp. 389 –394.

Part IV

APPENDIX

A APPENDIX

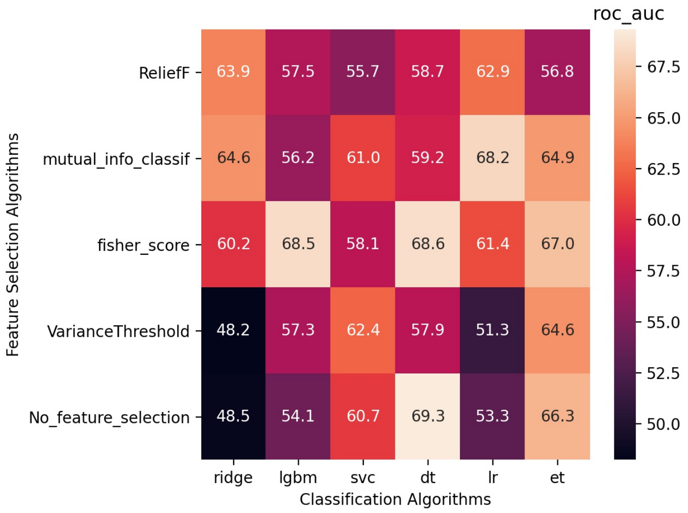


Figure A.1: Experiment 2, i.e. evaluating features derived from T2WIs and all DWIs. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Table 8.2 for details on a selection of models, and Figure 8.2b for feature selection rates.

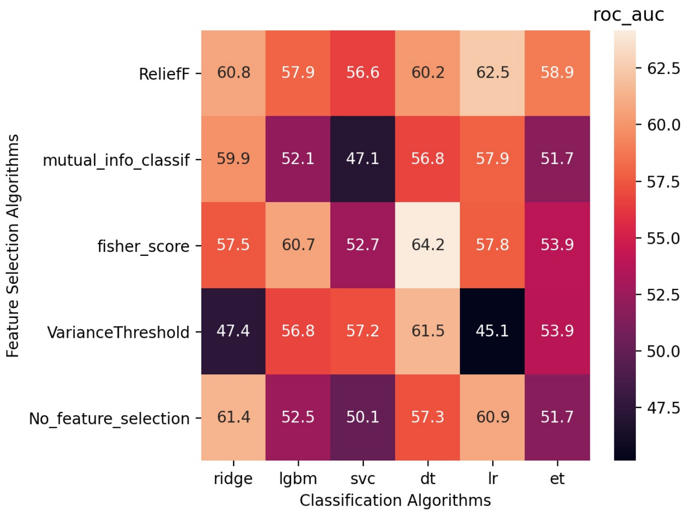


Figure A.2: Experiment 4, i.e. evaluating shape features derived from T2WIs and DWIs (b5). Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 8.3 for feature selection rates.

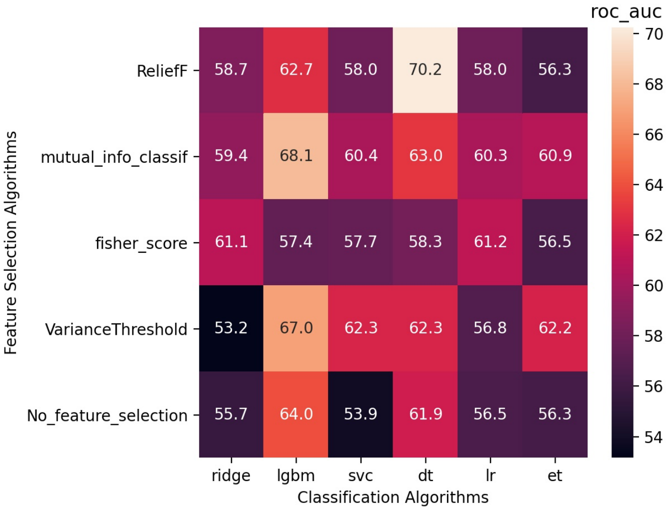


Figure A.3: Experiment 5, i.e. evaluating texture features derived from T2WIs and DWIs (b5). Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 8.4 for feature selection rates.

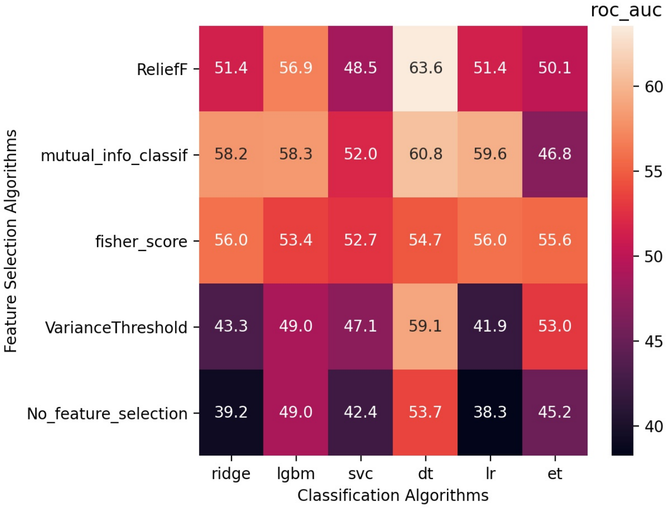


Figure A.4: Reproducibility of radiomic results: Experiment 6, repeating experiment 1 with Dataset2. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort.

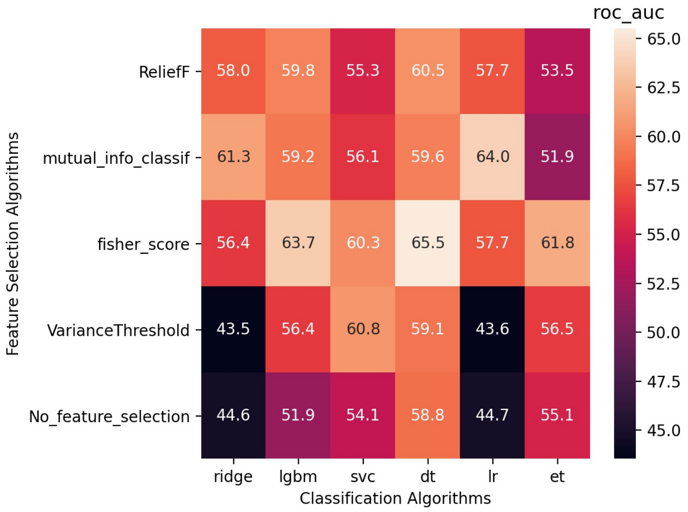


Figure A.5: Reproducibility of radiomic results: Experiment 6, repeating experiment 3 with Dataset2. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 10.1 for feature selection rates.

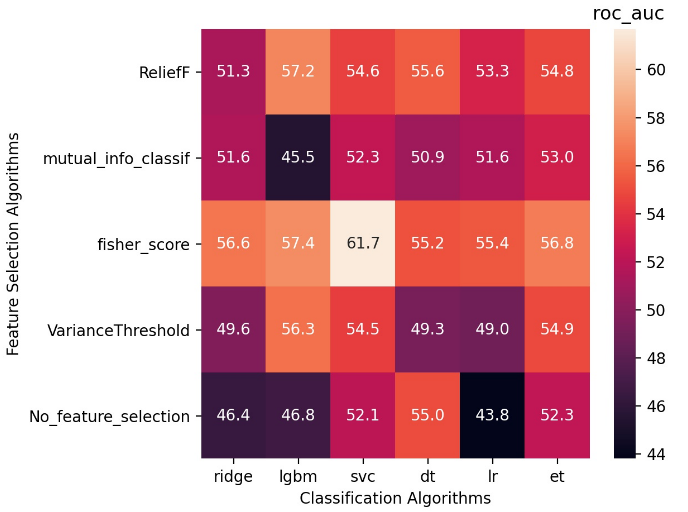


Figure A.6: Reproducibility of radiomic results: Experiment 8, repeating experiment 1 with images with bin width equal to 35. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 10.2a for feature selection rates.

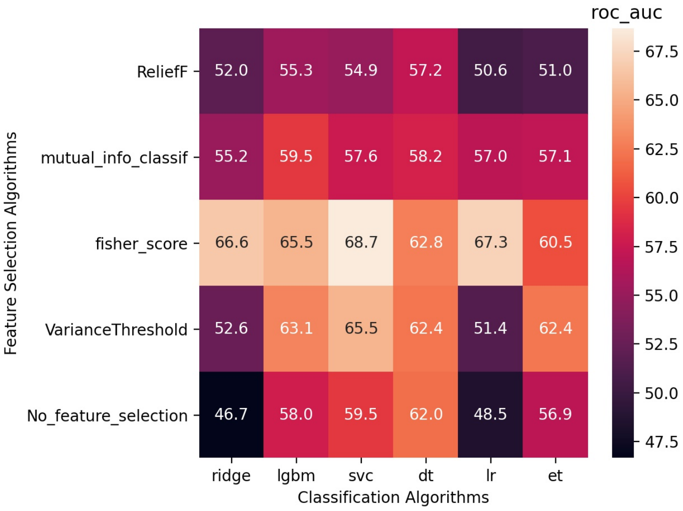


Figure A.7: Reproducibility of radiomic results: Experiment 8, repeating experiment 3 with with images with bin width equal to 35. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 10.2b for feature selection rates.

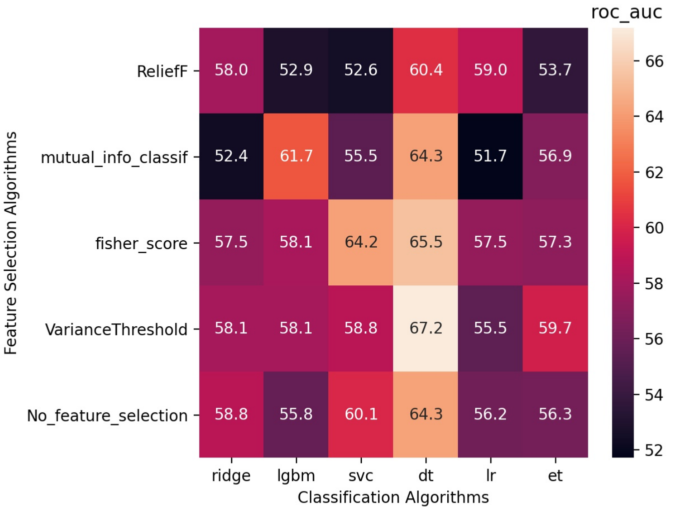


Figure A.8: Reproducibility of radiomic results: Experiment 10, i.e. repeating experiment 1 with mask2 defining the ROI. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 10.3a for feature selection rates.

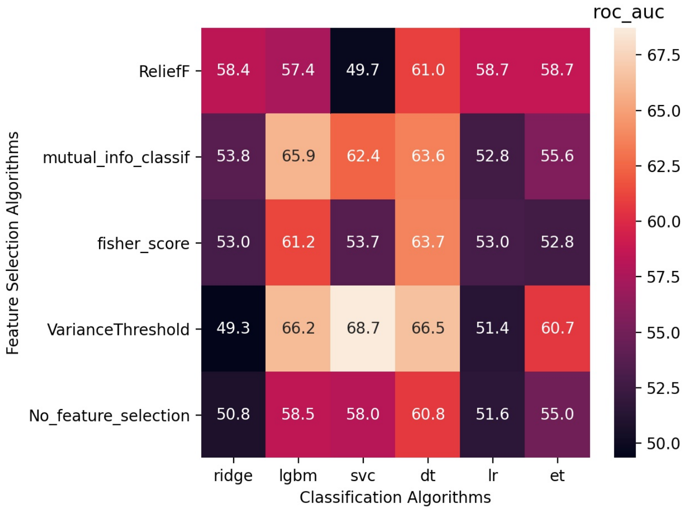


Figure A.9: Reproducibility of radiomic results: Experiment 10, i.e. repeating experiment 3 with mask2 defining the ROI. Performance, measured in AUC, for combinations of feature selector and classifiers predicting PFS in the all patient cohort. See Figure 10.3b for feature selection rates.

Experiment 1, All patients, Selected hyper-parameters and features
FS and SVC
$C_{SVC} = 2$
Selected features:
Maximum, first-order
Autocorrelation, GLCM
FS and DT
Impurity measure: gini index
Maximum tree depth = 30
Minimum number of samples at each leaf = 10
Selected features:
Maximum, first-order
Autocorrelation, GLCM
Small dependence low gray level emphasis, GLDM
Correlation, GLCM
MCC, GLCM
FS and ET
Impurity measure: entropy
Minimum number of samples at each leaf = 11
Selected features:
Maximum, first-order
Autocorrelation, GLCM
Small dependence low gray level emphasis, GLDM
Correlation, GLCM
MCC, GLCM

Table A.1: Experiment 1, i.e. evaluating features derived from T2WIs. Selected hyper-parameters and features for models presented in Table 8.1. Predicting PFS in the all patients cohort.

Experiment 2, All patients, Selected hyper-parameters and features
FS and DT
Impurity measure: entropy
Maximum tree depth = 10
Minimum number of samples at each leaf = 9
Selected features:
IMC 2 (b0), GLCM
Median (b1), first-order
High gray level run emphasis (b5), GLRLM
Long run high gray level emphasis (b5), GLRLM
Correlation (b5), GLCM
Difference entropy (b5), GLCM
Minimum (b6), first-order
Variance (b6), first-order
Low gray level zone emphasis (b6), GLSZM
Dependence non-uniformity normalized (b6), GLDM
High gray level emphasis (b6), GLDM
Large dependence low gray level emphasis (b6), GLDM
ID (b6), GLCM
Joint average (b6), GLCM
FS and ET
Impurity measure: gini index
Minimum number of samples at each leaf = 11
Selected features:
IMC 2 (b0), GLCM
Correlation (b5), GLCM
Difference entropy (b5), GLCM
Minimum (b6), first-order
Joint average (b6), GLCM

Table A.2: Experiment 2, i.e. evaluating features derived from T2WIs and all DWIs. Selected hyper-parameters and features for models presented in Table 8.2. Predicting PFS in the all patients cohort.

Experiment 2, All patients, Selected hyper-parameters and features
MI and LR
$C_{LR} = 3$
Selected features:
Small area high gray level emphasis, GLSZM
Joint energy (b1), GLCM
MI and ET
Impurity measure: entropy
Minimum number of samples at a leaf = 13
Selected features:
Small area high gray level emphasis, GLSZM
Joint energy (b1), GLCM
IDMN (b2), GLCM

Table A.3: Continued: Experiment 2, i.e. evaluating features derived from T2WIs and all DWIs. Selected hyper-parameters and features for models presented in Table 8.2. Predicting PFS in the all patients cohort.

Experiment 3, All patients, Selected hyper-parameters and features
MI and ET
Impurity measure: gini index
Minimum number of samples at each leaf = 8
Selected features:
Small area high gray level emphasis, GLSZM
Dependence non-uniformity (b5), GLDM
FS and ET
Impurity measure: entropy
Minimum number of samples at each leaf = 5
Selected features:
Gray level variance, GLRLM
Zone percentage (b5), GLSZM
Run percentage (b5), GLRLM
Short run emphasis (b5), GLRLM
Dependence non-uniformity normalized (b5), GLDM
Gray level non-uniformity (b5), GLDM
Large dependence high gray level emphasis (b5), GLDM
Contrast (b5), GLCM
Difference average (b5), GLCM
Difference variance (b5), GLCM
Inverse variance (b5), GLCM
Joint energy (b5), GLCM
MI and SVC
$C_{SVC} = 1$
Selected features:
Small area high gray level emphasis, GLSZM
Dependence non-uniformity (b5), GLDM
MI and DT
Impurity measure: gini index
Max tree depth: 30
Minimum number of samples at each leaf = 7
Selected features:
Small area high gray level emphasis, GLSZM

Table A.4: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 8.3. Predicting PFS in the all patients cohort.

Experiment 3, All patients, Selected hyper-parameters and features
FS and DT
Impurity measure: gini index
Max tree depth: 30
Minimum number of samples at each leaf = 14
Selected features:
Gray level variance, GLRLM
Run percentage (b5), GLRLM
Short run emphasis (b5), GLRLM
Dependence non-uniformity normalized (b5), GLDM
Gray level non-uniformity (b5), GLDM
Joint energy (b5), GLCM
FS and LR
$C_{LR} = 3$
Selected features:
Gray level variance, GLRLM
Run percentage (b5), GLRLM
Short run emphasis (b5), GLRLM
Dependence non-uniformity normalized (b5), GLDM

Table A.5: Continued: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 8.3. Predicting PFS in the all patients cohort.

Experiment 1, nCRT cohort, Selected hyper-parameters and features
MI and ET
Impurity measure: entropy
Minimum number of samples at each leaf = 7
Selected features:
Skewness, first-order
Small area high gray level emphasis, GLSZM
Zone entropy, GLSZM
Zone variance, GLSZM
High gray level run emphasis, GLRLM
Autocorrelation, GLCM
MI and DT
Impurity measure: gini index
Max tree depth: 20
Minimum number of samples at each leaf = 12
Selected features:
Median, first-order
Skewness, first-order
Small area high gray level emphasis, GLSZM
Zone entropy, GLSZM
Zone variance, GLSZM
High gray level run emphasis, GLRLM
Autocorrelation, GLCM
MI and RR
$\alpha_{RR} = 1$
Selected features:
Mean, first-order
Skewness, first-order
Large area emphasis, GLSZM
Small area high gray level emphasis, GLSZM
Zone entropy, GLSZM
Zone variance, GLSZM
High gray level run emphasis, GLRLM
Autocorrelation, GLCM

Table A.6: Experiment 1, i.e. evaluating features derived from T2WIs. Selected hyper-parameters and features for models presented in Table 8.4. Predicting PFS in the nCRT cohort.

Experiment 3, nCRT cohort, Selected hyper-parameters and features
FS and ET
Impurity measure: entropy
Minimum number of samples at each leaf = 5
Selected features:
Size zone non-uniformity normalized (b5), GLSZM
Zone variance (b5), GLSZM
Dependence non-uniformity normalized (b5), GLDM
Gray level non-uniformity (b5), GLDM
Gray level variance (b5), GLDM
ID (b5), GLCM
FS and DT
Impurity measure: gini index
Max tree depth: 20
Minimum number of samples at each leaf = 5
Selected features:
Size zone non-uniformity normalized (b5), GLSZM
Gray level variance (b5), GLDM

Table A.7: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 8.5. Predicting PFS in the nCRT cohort.

Experiment 1, nCRT cohort, Selected hyper-parameters and features
FS and LR
$C_{LR} = 3$
Selected features:
Small area emphasis, GLSZM
Autocorrelation, GLCM
IDM, GLCM
IDMN, GLCM
ReF and SVC
$C_{SVC} = 1$
$n = 1$
Selected features:
Flatness, shape
IMC1, GLCM
Elongation, shape
Correlation, GLCM
ReF and ET
Impurity measure: gini index
Minimum number of samples at each leaf = 7
Selected features:
Flatness, shape
IMC1, GLCM
Elongation, shape

Table A.8: Experiment 1, i.e. evaluating features derived from T2WIs. Selected hyper-parameters and features for models presented in Table 9.1. Predicting TRG in the nCRT cohort.

Experiment 1, nCRT cohort, Selected hyper-parameters and features
FS and ET
Impurity measure: gini index
Minimum number of samples at each leaf = 5
Selected features:
Small area emphasis, GLSZM
Autocorrelation, GLCM
Cluster tendency, GLCM
Difference average, GLCM
IDM, GLCM
IDMN, GLCM
MI and DT
Impurity measure: entropy
Max tree depth: 10
Minimum number of samples at each leaf = 7
Selected features:
Zone entropy, GLSZM
Contrast, NGTDM
Dependence variance, GLDM
Cluster prominence, GLCM
Cluster tendency, GLCM
IMC1, GLCM
Sum entropy, GLCM

Table A.9: Continued: Experiment 1, i.e. evaluating features derived from T2WIs. Selected hyper-parameters and features for models presented in Table 9.1. Predicting TRG in the nCRT cohort.

Experiment 3, nCRT cohort, Selected hyper-parameters and features
MI and LR
$C_{LR} = 1$ Selected features:
Dependence variance, GLDM
Cluster prominence, GLCM
Cluster tendency, GLCM
IMC1, GLCM
Sum entropy, GLCM
Low gray level zone emphasis (b5), GLSZM
Zone entropy (b5), GLSZM
Busyness (b5), NGTDM
ReF and LGBM
Max tree depth: 28
Minimum number of samples at each leaf = 8
Minimum number of leaves = 13
$n = 1$
Selected features:
Flatness, shape
Short run low gray level emphasis (b5), GLRLM
Median (b5), first-order
IMC1, GLCM
Low gray level run emphasis (b5), GLRLM
Run entropy (b5), GLRLM
Difference entropy, GLCM
Cluster prominence (b5), GLCM
MI and RR
$\alpha_{RR} = 3$
Selected features:
Dependence variance, GLDM
Cluster prominence, GLCM
Cluster tendency, GLCM
IMC1, GLCM
Sum entropy, GLCM
Low gray level zone emphasis (b5), GLSZM
Zone entropy (b5), GLSZM

Table A.10: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 9.2. Predicting TRG in the nCRT cohort.

Experiment 1, nCRT cohort, Selected hyper-parameters and features
MI and ET
Impurity measure: entropy
Minimum number of samples at each leaf = 9
Selected features:
Dependence variance, GLDM
Cluster prominence, GLCM
IMC1, GLCM
Sum entropy, GLCM
Low gray level zone emphasis (b5), GLSZM
Zone entropy (b5), GLSZM
MI and DT
Impurity measure: entropy
Max tree depth: 30
Minimum number of samples at a leaf = 12
Selected features:
Dependence variance, GLDM
Cluster prominence, GLCM
Cluster tendency, GLCM
IMC1, GLCM
Sum entropy, GLCM
Low gray level zone emphasis (b5), GLSZM
Zone entropy (b5), GLSZM

Table A.11: Continued: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 9.2. Predicting TRG in the nCRT cohort.

Experiment 1, nCRT cohort, Selected hyper-parameters and features
ReF and ET
Impurity measure: gini index
Minimum number of samples at each leaf = 7
$n = 1$
Selected features:
Sum entropy, GLCM
Size zone non-uniformity normalized, GLSZM
Low gray level zone emphasis, GLSZM
High gray level zone emphasis, GLSZM
Small area high gray level emphasis, GLSZM
Small area emphasis, GLSZM
ReF and SVC
$C_{SVC} = 2$
$n = 1$
Selected features:
Sum entropy, GLCM
Size zone non-uniformity normalized, GLSZM
Low gray level zone emphasis, GLSZM
High gray level zone emphasis, GLSZM
Small area high gray level emphasis, GLSZM

Table A.12: Experiment 1, i.e. evaluating features derived from T2WIs. Selected hyper-parameters and features for models presented in Table 9.3. Predicting ypT in the nCRT cohort.

Experiment 1, nCRT cohort, Selected hyper-parameters and features
MI and DT
Impurity measure: gini index
Max tree depth: 30
Minimum number of samples at each leaf = 4
Selected features:
Low gray level zone emphasis, GLSZM
Small area low gray level emphasis, GLSZM
Sum entropy, GLCM
ReF and LGBM
Max tree depth: 18
Minimum number of samples at each leaf = 6
Minimum number of leaves = 5
$n = 2$
Selected features:
Sum entropy, GLCM
Low gray level zone emphasis, GLSZM
Small area high gray level emphasis, GLSZM

Table A.13: Continued: Experiment 1, i.e. evaluating features derived from T2WIs. Selected hyper-parameters and features for models presented in Table 9.3. Predicting ypT in the nCRT cohort.

Experiment 3, nCRT cohort, Selected hyper-parameters and features
MI and DT
Impurity measure: entropy
Max tree depth: 30
Minimum number of samples at each leaf = 3
Selected features:
Large area emphasis, GLSZM
Low gray level zone emphasis, GLSZM
Small area low gray level emphasis, GLSZM
Cluster tendency, GLCM
Sum entropy, GLCM
Entropy (b5), first-order
Uniformity (b5), first-order
Run entropy (b5), GLRLM
MI and ET
Impurity measure: entropy
Minimum number of samples at each leaf = 10
Selected features:
Large area emphasis, GLSZM
Low gray level zone emphasis, GLSZM
Small area low gray level emphasis, GLSZM
Cluster tendency, GLCM
Sum entropy, GLCM
Entropy (b5), first-order
Uniformity (b5), first-order
Run entropy (b5), GLRLM

Table A.14: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 9.3. Predicting ypT in the nCRT cohort.

Experiment 3**, Selected hyper-parameters and features**
MI and RR
$\alpha_{RR} = 1$
Selected features:
Low gray level zone emphasis, GLSZM
Sum entropy, GLCM
Entropy (b5), first-order
FS and DT
Impurity measure: entropy
Max tree depth: 30
Minimum number of samples at each leaf = 8
Selected features:
Sum entropy, GLCM
Maximum 2D diameter slice (b5), shape
Mean absolute deviation (b5), first-order
Gray level non-uniformity (b5), GLRLM
Low gray level run emphasis (b5), GLRLM
Run entropy (b5), GLRLM
Short run high gray level emphasis (b5), GLRLM
Maximum probability (b5), GLCM

Table A.15: Continued: Experiment 3, i.e. evaluating features derived from T2WIs and DWIs (b5). Selected hyper-parameters and features for models presented in Table 9.3. Predicting ypT in the nCRT cohort.

Aase Mellingen Langan

MRI-Based Radiomics Analysis for Predicting Treatment Outcome in Rectal Cancer

NTNU
Norwegian University of
Science and Technology