Norwegian University
of Life Sciences

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

Searching for Biomarkers of Disease-Free Survival in Head and Neck Cancers Using PET/CT Radiomics

Geir Severin Rakh Elvatun Langberg

Data Science

Abstract

The goals of this thesis were to (1) study methodologies for radiomics data analysis, and (2) apply such methods to identify biomarkers of *disease-free survival* in *head and neck cancers*.

Procedures for radiomics feature extraction and feature exploration in biomarker discovery were implemented with the *Python*^{TM} programming language. The code is available at `https://github.com/gsel9/biorad`.

In a retrospective study of disease-free survival as response to radiotherapy, radiomics features were extracted from PET/CT images of 198 head and neck cancers patients. A total of 513 features were obtained by combining the radiomics features with clinical factors and PET parameters. Combinations of seven feature selection and 10 classification algorithms were evaluated in terms of their ability to predict patient treatment response. By using a combination of *MultiSURF* feature selection and *Extreme Gradient Boosting* classification, subgroup analyses of *HPV negative oropharyngeal* (HPV unrelated) cancers gave 76.4 ± 13.2 % area under the *Receiver Operating Characteristic* curve (AUC). This performance was superior to the baseline of 54 % for disease-free survival outcomes in the patient subgroup.

Four features were identified as prognostic of disease-free survival in the HPV unrelated cohort. Among these were two CT features capturing intratumour heterogeneity. Another feature described tumour shape and was, contrary to the CT features, significantly correlated with the tumour volume. The fourth feature was the median CT intensity. Determining the prognostic value of these features in an independent cohort will elucidate the relevance of tumour volume and intratumour heterogeneity in treatment of HPV unrelated head and neck cancer.

Acknowledgements

First, I would like to thank my main supervisor Prof. Cecilia M. Futsæther for the thesis subject, her enthusiasm and inspiring curiosity.

There are also several others at the Norwegian University of Life Sciences to which I am grateful. Not only have Ass. Profs. Kristian Liland, Oliver Tomic and Ulf Indahl contributed with tips and tricks in fruitful discussions, but their dedication to teaching and lecturing was much appreciated when preparing for this thesis. MSc Aurora Rosvoll Grøndahl has answered many questions and I am very thankful for her assistance.

Moreover, I thank Prof. Eirik Malinen and oncologist Dr Einar Dale from the University of Oslo for granting me access to the data set, which enabled me to carry out my studies.

Last, but never least, I thank my family for supporting me in all my endeavours.

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Sections

List of Abbreviations

CV	Cross-validation
CT	Computed tomography
SVC	C-Support vector classifier
DT	Decision tree
ECOG	Eastern cooperative oncology group
XGB	Extreme gradient boosting
ET	Extremely randomised trees
GLCM	Gray-Level Co-Occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Grey-Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
HPV	Human papillomavirus
ICC	Intraclass correlation coefficient
KNN	K-nearest neighbours
LGBM	Light gradient boosting machine
LR	Logistic regression
MTV	Metabolic tumor volume
NGTDM	Neighbouring Gray Tone Difference Matrix
PET	Positron emission tomography
QDA	Quadratic discriminant analysis
RF	Random forest
ROI	Region of interest (the tumour)

Ridge	Ridge classifier
SMAC	Sequential model-based algorithm configuration
SHAP	Shapley additive explanations
SCC	Spearman’s rank correlation coefficient
SUV	Standardised uptake value
TLG	Total lesion glycolysis

List of Symbols

$\equiv$	*Defined as.*
$\mapsto$	*Maps to.*
$:=$	*Redefined/updated.*
$\vert$	*Such that.*
$\subset$	*Subset.*
$\in$	*An element of.*
$\cup$	The union.
$\lceil \mathbf{x} \rceil$	The ceiling function.
$\odot$	Element-wise multiplication.
$\mathbb{R}$	The set of all real numbers.
$\log x$	The natural logarithm of x with base e.
$\log_2 x$	The binary logarithm of x with base two.
$A \setminus B$	The difference between the elements in the sets A and B.
$\lambda(\cdot, \cdot)$	A learning algorithm.
$\mathbf{\Phi}$	A hyper-parameter domain.
$\mathbb{E}(\cdot)$	The expected value.

Chapter 1

Introduction

Over 9.5 million people died from cancer in 2018 [1]. Head and neck cancers accounted for more than 300,000 deaths, as the seventh most common type of cancer worldwide [2].

Treatment selection in head and neck cancers relies primarily on the location and stage of the primary tumour at diagnosis [3]. One of the conventional treatments for head and neck cancers is radiotherapy [4]. Radiotherapy involves using ionising radiation to destroy or damage cancer cells and is a significant contribution to cancer treatment [5].

However, radiation damage to healthy tissue may considerably reduce the life quality of patients. High-precision techniques, such as *intensity modulated radiation therapy* (IMRT), adapts the radiation doses to avoid critical organs while conforming to the tumour [5]. Techniques, such as IMRT, have been used to reduce complications and side effects compared to conventional therapy.

Still, different treatment outcomes have been observed despite patients having seemingly identical disease characteristics [4]. To further adapt radiotherapy treatment to each patient, Caudell et al. (2017) suggested to replace fractionation and empirical dosing by precision medicine tools, such as *genomics* and *radiomics* [4].

Genomic approaches to molecular tumour characterisation typically require invasive tissue extraction [6]. However, methods such as *biopsy* are prone to sampling errors. These errors occur due to the spatial differences in the tumour, which is referred to as intratumour heterogeneity [7]. Intratumour heterogeneity describes genomic differences between clusters of cells within the same tumour [8], and is one of the main challenges for precision medicine, according to Caudell et al. (2017).

Medical imaging technology, on the other hand, enables non-invasive visualisation of patient biology and internal structures [9]. For instance, a *18F-fluorodeoxyglucose Positron Emission Tomography/Computed Tomography* (PET/CT) scanner combines the PET and CT technologies to image biological function and anatomy [10]. With human cancers exhibiting phenotypic differences, medical imaging can be used to characterise intratumour heterogeneity [11].

1.1 Motivation

Radiomics is a field of medical study where quantification of disease characteristics is based on radiographic phenotyping [12]. The term *radiomics* was first used by Gillies et al. (2010) [13] to describe gene expression in terms of image descriptors.

In radiomics, medical images are transformed into high-dimensional descriptors, or *features*, assumed to encapsulate the underlying cancer pathophysiology [14], [15]. These features does not only quantify intratumour heterogeneity but also the shape and size of the tumour, as well as image intensity characteristics [16].

Studies have demonstrated the potential of radiomics features to predict clinical outcomes across different types of tumours and modalities [14], [17], [18], [19]. Thus, these features may be used as objective indicators of medical state, referred to as *biomarkers* [20]. Derivation of disease-specific biomarkers can contribute to elucidate the relevance of intratumour heterogeneity in treatment of head and neck cancers. Moreover, such biomarkers can be used to adapt therapies to individuals or subgroups of patients.

1.2 Subjects and Goals

The initial goal of this thesis was to develop methods for radiomics data analysis. These methodologies were to include radiomics feature extraction, and assessment of the prognostic value of such features given a clinical outcome.

The second goal of this thesis was to identify potential biomarkers for prediction of *disease-free survival* [21] as a response to radiotherapy in head and neck cancers.

1.3 Method

In previous radiomics studies, biomarkers have been determined by using a predictive model to infer the relevance of features with respect to a clinical outcome [14], [17], [22]. According to the *No Free Lunch* theorems [23], no single algorithm will be superior in all applications. Therefore, several algorithms should be compared in terms of their ability to predict patient treatment response.

In general, the steps in biomarker discovery comprised:

1. A data-driven approach to identify the classification model superior in predicting disease-free survival.
2. Inference on the relevancy of each feature, based on the selected model, to identify predictors as potential biomarkers.

Note that this thesis was not dedicated only to the study of radiomics features but also included clinical factors and PET parameters.

Preparation of an independent test set was not completed during this thesis. External validation of results is therefore left to future studies.

1.4 Organisation

This thesis is structured according to the IMRaD format, which is an acronym of Introduction, Method, Results, and Discussion [24].

Relevant definitions and notation are described in the introduction to each chapter.

Chapter 2 outlines the theory behind the experiments described in Chapter 3. Chapter 2 aims to elaborate on the methodology leading up to classification experiments.

These experiments are described in Chapter 3 and the results are given in Chapter 4. Experimental results that were not included in Chapter 4 are given in Appendix B. A discussion of the experimental framework, results and observations is given in Chapter 5. Suggestions to future work in radiomics are given in the last two sections of Chapter 5. Chapter 6 is structured such that conclusions for each of the two thesis goals are given in separate sections.

All code material produced in this thesis is publicly available via the GitHub© online hosting service [25].

Chapter 2

Theory

The notation used in this chapter is as follows. Scalars, such as $b \in \mathbb{R}$, are not given in bold typeface. A feature of n elements is denoted $\mathbf{x} \in \mathbb{R}^n$. The number of elements in $\mathbf{x}$ is expressed as $|\mathbf{x}|$. A feature norm is defined as

$$\|\mathbf{x}\| \equiv \sqrt{\mathbf{x}_1^2 + \cdots + \mathbf{x}_n^2}$$

A set of p features organised into a feature matrix is indicated by $\mathbf{X} \in \mathbb{R}^{n \times p}$. Column j of this matrix refers to feature $\mathbf{x}^{(j)}$, while row i refers to an observation, or sample, $\mathbf{x}_i$.

In a classification setting, each observation in $\mathbf{X}$ belongs to a class $c \in \Omega$, which is also expressed as $\mathbf{y} \in \Omega$. Moreover, each element $\mathbf{y}_i \in \mathbf{y}$ corresponds to exactly one observation $\mathbf{x}_i \in \mathbf{x}$.

The mean and variance of a set of elements are indicated by $\mu(\cdot)$ and $\sigma(\cdot)$, respectively.

Given a learning algorithm, $\lambda(\cdot, \cdot)$, and a parameter configuration, ϕ, a model is represented as $\lambda(\phi, \cdot)$. That is, a model is given as a particular configuration of a learning algorithm. Note that these parameters, termed *hyperparameters*, are not learned during model training, but specified before training the model. Model predictions, $\hat{\mathbf{y}}$, are obtained by applying a trained model to observations.

2.1 Statistical Significance Tests

Statistical significance tests are used to assess the probability of making false assumptions about the data. Assumptions can, for instance, be made about the distribution of samples or the relationship between features. These statistical tests can be separated into two classes referred to as *parametric* and *non-parametric*. The *parametric* tests are based on assumptions about the distribution of the data, while *non-parametric* tests does not require any such condition [26].

2.1.1 Wilcoxon Signed-Rank Test

The *Wilcoxon Signed-Rank* (WSR) [27] test evaluates if the difference between paired samples are likely to follow a normal probability distribution [28]. This test constitutes a non-parametric alternative to the *paired t-test* [26]. The main difference between the WSR and t-test is that the latter compares the means of samples, whereas WSR considers the ordering of the data [26].

Given two features, $\mathbf{x}^{(j)}$ and $\mathbf{x}^{(l)}$, the WSR null hypothesis states that the mean ranks of the $\mathbf{x}^{(j)}$ and $\mathbf{x}^{(l)}$ populations differ [27]. The test statistic, W, is calculated from a reduced set of n_r paired samples where all samples satisfying

$$\sqrt{\left(\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)}\right)^2} = 0$$

have been excluded. Then, W, is calculated as

$$W = \sum_{i=1}^{n_r} z\left(\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)}\right) \cdot \text{rank}\left(\sqrt{\left(\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)}\right)^2}\right)$$

where $z(\cdot)$ is the sign function.

2.1.2 Shapiro-Wilk W-Test

The *Shapiro-Wilk* (SW) test evaluates if a sample is normally distributed by considering the skewness and the kurtosis of the data [29].

The SW test statistic, W, is given as [29]

$$W = \frac{\left(\sum_{i=1}^{n} a_i \mathbf{x}'_i\right)^2}{\sum_{i=1}^{n} \left(\mathbf{x_i} - \mu(x)\right)^2}$$

where $\mathbf{x}'_i$ denotes the *i*th smallest element of $\mathbf{x}$, or the *i*th order statistic. The coefficients $a_i \in \mathbf{a}$ are given as

$$\mathbf{a} = \frac{\mathbf{m}^T \mathbf{C}^{-1}}{\|\mathbf{C}^{-1}\mathbf{m}\|}$$

with $\mathbf{m}$ as the expected order statistics obtained by sampling from a standard normal distribution. Moreover, $\mathbf{C}$ denotes the co-variance matrix of the normal order statistics.

2.1.3 D'Agostino's K-Squared Test

Similar to the Shapiro-Wilk test, the *D'Agostino's K^2* (K^2) test is also based on skewness and kurtosis to determine if a sample originates from a normally distributed population [30].

Let k and s denote the kurtosis and skewness of feature $\mathbf{x}$. The K^2 statistic is given as [30]

$$K^2 = Z_1(s) + Z_2(k)$$

where Z_1 and Z_2 are the transformed versions of the skewness and kurtosis, respectively. Details on these transformations are available in the literature [30].

2.2 Measures of Feature Correlation

Correlation metrics describe the relationship between features. If two features are highly correlated then these features represent the same information, which renders one of them redundant.

2.2.1 Intraclass Correlation Coefficient

The *Intraclass Correlation Coefficient* (ICC) describes the relationship between features in a group [31]. A two-way mixed effects ICC score for a single measurement is given by [32]

$$\mathrm{ICC} = \frac{\mathrm{MS}_R - \mathrm{MS}_E}{\mathrm{MS}_R + (k-1)\mathrm{MS}_E} \tag{2.1}$$

where MS_R indicates the mean square of rows, MS_E is the mean square error and k is the number of features in the group. The MS_R and MS_E quantities can be obtained from a two-way *Analysis of Variance* [33]. The ICC score ranges from zero to one, where perfect correlation between the group members is indicated by an ICC equal to one [31].

2.2.2 Spearman's Rank Correlation Coefficient

The *Spearman's Rank Correlation* coefficient (SCC) measures statistical dependence between two features by comparing their ranks [34]. The SCC captures all monotonic relationships.

Let $\mathbf{x}_r = \text{rank } \mathbf{x}$ and $\mathbf{y}_r = \text{rank } \mathbf{y}$ denote the order statistics of two features $\mathbf{x}$ and $\mathbf{y}$, respectively. The SCC, ρ, is calculated from [34]

$$\rho = \frac{\mathbf{C}(\mathbf{x}_r, \mathbf{y}_r)}{\sigma(\mathbf{x}_r)\sigma(\mathbf{y}_r)}$$

where $\mathbf{C}$ is the co-variance matrix, and σ is the standard deviation. The SCC ranges from negative to positive one, whereas both endpoints represent perfect correlation. Hence, SCC equal to zero signifies no correlation.

2.3 Clustering

Clustering is referred to as *unsupervised learning* methods that determines groups, or *clusters*, based on sample characteristics [35]. Clustering is an approach to express underlying patterns in data.

2.3.1 K-Means

The *K-means* algorithm partitions n samples into K clusters based on the squared Euclidean distance between these data points [36]. In each cluster, the distance between samples are measured relative to a set of cluster centers, termed *centroids*. The K-means objective is to minimise the within-cluster sum of squares given as

$$\arg\min_{\mathbf{S}} \sum_{k=1}^{K} \sum_{\mathbf{x}\in\mathbf{S}_k} \|\mathbf{x} - \widehat{\mathbf{x}}_k\|^2 \tag{2.2}$$

for a feature $\mathbf{x}$, and $\mathbf{S} = \{\mathbf{S}_k\}_{k=1}^{K}$ as the K clusters with centroids $\{\widehat{\mathbf{x}}_k\}_{k=1}^{K}$. In practice, K-means approximates this objective over a budget of iterations. Initially, samples are randomly selected from the data to serve as centroids. However, at iteration t, centroid $\widehat{\mathbf{x}}_k^{(t-1)}$ of cluster $\mathbf{S}_k^{(t-1)}$ is updated according to

$$\widehat{\mathbf{x}}_k^{(t)} = \frac{1}{\left|\mathbf{S}_k^{(t-1)}\right|} \sum_{\mathbf{x}_i\in\mathbf{S}_k^{(t-1)}} \mathbf{x}_i$$

The K-means algorithm converges when the centroid at step t is equal to the centroid at step $t-1$.

K-Means++

An alternative approach to initialise centroids, other than the random selection of samples, was introduced by Arthur and Vassilvitskii (2007) [37]. Arthur and Vassilvitskii (2007) named this the K-Means++ algorithm, which has shown to be an improvement considering cluster quality and convergence [37]. In K-Means++, only the first centroid, $\widehat{\mathbf{x}}_1$, is selected uniformly at random from the data. The following $K-1$ centroids are selected from the remaining samples with probability.

$$P(\widehat{\mathbf{x}}_k = \mathbf{x}_i \mid C) = \frac{d(\mathbf{x}_i, \widehat{\mathbf{x}})^2}{\sum_{i=1}^n d(\mathbf{x}_i, \widehat{\mathbf{x}})^2}$$

where $d(\mathbf{x}_i, \widehat{\mathbf{x}})$ represents the shortest Euclidean distance from a sample, $\mathbf{x}_i$, to any of the already selected centroids, $\widehat{\mathbf{x}} = (\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_{k-1})$. Once all the K centroids have been selected, the algorithm proceeds according to K-means, as described in the previous section [37].

Cluster Distortion

The quality of the clusters obtained with K-means [36] or K-means++ [37] can be quantified in terms of the *Sum of Squared Errors* (SSE). The SSE is also refereed to as *cluster distortion* [36], and is given as

$$\text{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{n} \|\mathbf{x}_i - \widehat{\mathbf{x}}_k\| \tag{2.3}$$

The value of K giving the smallest SEE corresponds to the optimal number of clusters.

2.3.2 Spectral Co-Clustering

The *Spectral Co-Clustering* algorithm [38] belongs to the category of *biclustering* algorithms [39]. Biclustering involves simultaneous clustering of rows and columns by dividing the original data into subsets of samples and features [39]. These subsets are referred to as a *biclusters*. The Spectral Co-Clustering algorithm assumes that each row and column of the original data matrix belongs to exactly one such bicluster [38].

Initially, the data is processed to give a matrix $\widehat{\mathbf{X}}$ with constant row and column sums. This processing is performed over t iterations according to

$$\widehat{\mathbf{X}}_{t+1} = \mathbf{R}_t^{-1/2}\widehat{\mathbf{X}}_t\mathbf{M}_t^{-1/2}$$

where $\mathbf{R}$ and $\mathbf{M}$ are diagonal matrices. The diagonal elements of $\mathbf{R}$ and $\mathbf{M}$ are given as

$$\mathbf{R}_i^{(i)} = \sum_{j=1}^{p} \widehat{\mathbf{X}}_i^{(j)}$$

$$\mathbf{M}_j^{(j)} = \sum_{i=1}^{n} \widehat{\mathbf{X}}_i^{(j)}$$

That is, entry (i, i) in $\mathbf{R}$ holds the sum across all p columns, at row i of $\widehat{\mathbf{X}}$. Moreover, $\mathbf{M}$ holds the sum across all n rows of column j in $\widehat{\mathbf{X}}$ in each diagonal entry (j, j).

After pre-processing, the *Singular Value Decomposition* [40] gives

$$\widehat{\mathbf{X}} = \mathbf{USV}^T$$

where subsets, $\mathbf{U}'$ and $\mathbf{V}'$, of l vectors from $\mathbf{U}$ and $\mathbf{V}$ represents the bicluster row and column partitions. For K number of biclusters to detect, the size of these subsets is determined by

$$l = \lceil \log_2 K \rceil + 1$$

A matrix Z is constructed according to

$$\mathbf{Z} = \begin{bmatrix} \mathbf{R}^{-1/2}\mathbf{U}' \\ \\ \mathbf{M}^{-1/2}\mathbf{V}' \end{bmatrix}$$

By a K-means++ clustering of $\mathbf{Z}$, the biclusters are obtained from the resulting row and column partitions [38].

2.3.3 Measuring the Quality of a Bicluster

Biclustering algorithms were originally applied to gene expression data [41]. Consequently, measures to quantify the quality of clusters were developed based on behavioural patterns in gene expression. Two types of such patterns are *scaling* and *shifting*. Scaling describes a multiplicative relation between samples, while shifting describes an additive relation. The *Transposed Virtual Error* is a metric that detects both shifting and scaling patterns [41].

Let $B \in \mathbb{R}^{n \times m}$ be a bicluster with n rows and m columns. Each entry in B at row i and column j is indicated by b_i, j. The Transposed Virtual Error metric includes a quantity referred to as *Virtual Condition* given as [41]

$$\rho_j = \frac{\sum_{i \in n} b_{i,j}}{n}$$

The Virtual Condition value, ρ_j, represents the mean of column j in a bicluster. Thereby, the Transposed Virtual Error is calculated for a bicluster as

$$\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \sqrt{(b_{i,j} - \rho_j)^2} \tag{2.4}$$

In Equation 2.3.3, the values in a bicluster are compared to the Virtual Condition of that bicluster. That is, the quality of the cluster is represented as the distance between cluster members and the Virtual Condition.

2.4 Feature Selection

Feature selection methods seek a subset of features under the assumption that amongst the original features is irrelevant or redundant information [42], [43]. Feature selection may contribute to reducing over-fitting and improve the performance of predictive models [43].

Given a criterion $J(\cdot)$ to measure feature relevance with respect to some objective. A subset, $\widetilde{X} \subset X$, of the most relevant features according to J can be selected by [43], [44]

$$\widetilde{X} = \arg\max_{X' \subset X} J(X')$$

Feature selection algorithms can be categorised as (1) embedded, (2) filter or (3) wrapper methods. Filter methods are described in the following sections. Embedded feature selection is performed inherently by mechanisms of some algorithm. For instance, regularisation of the algorithm optimisation objective [45]. Wrappers use a specific model to select features according to the performance of the model. Thus, wrapper methods are more prone to over-fitting and more computationally expensive compared to embedded and filter methods.

2.4.1 Univariate Filter Methods

Univariate methods consider only the relationship between individual features, and are typically computationally efficient even in high-dimensional problems [44]. However, these methods are incapable of capturing information from multiple interacting features.

Chi-Squared

The *Chi-Squared*, or χ^2, method uses the χ^2 statistic to rank features [46], [17].

Let $|\mathbf{x}|$ denote the cardinality of feature $\mathbf{x}$ with n samples belonging to distinct classes, Ω. The χ^2 score function, $J(\cdot)$, for feature selection is given as [17]

$$J(\mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \sum_{c \in \Omega} \frac{(n_{i,c} - \mu_{i,c})^2}{\mu_{i,c}}$$

for $n_{i,c}$ as the number of samples equal to value i and belonging to class c. Moreover,

$$\mu_{i,c} = \frac{n_i \cdot n_c}{n}$$

where n_i is the number of samples of value i, while n_c the number of samples in class c.

Mutual Information

The *Mutual Information* (MI) from information theory can be used in feature selection to quantify the dependence between two features [47]. Independent features correspond to zero MI, but as features are more related, the MI increases [48].

Let $\mathbf{x}_c$ represent the observations in feature $\mathbf{x}$ that belongs to class c. For each sample, $\mathbf{x}_i$, the quantity I_i is calculated as [47]

$$I_i = \psi(n) - \psi(\mathbf{x}_c) + \psi(k) - \psi(m_{i,k}) \tag{2.5}$$

where $m_{i,k}$ is the k nearest neighbours to $\mathbf{x}_i$ selected from $\mathbf{x}_c$. In Equation 2.5, $\psi(\cdot)$ represents the digamma function defined as the logarithmic derivative of the gamma function [49]

$$\psi(n) = \frac{d}{dn} \log \Gamma(n)$$

The MI score is estimated by averaging I_i across all observations [47]

$$J(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} I_i$$

The number of neighbours parameter, k, can be optimised to each problem.

Wilcoxon Rank Sum

The *Wilcoxon Rank Sum* (WRS) method compares the medians of ranked features to determine their resemblance [50], [51]. In a classification problem, the WRS scoring function for feature selection can be formulated as [17]

$$J(\mathbf{x}) = (n-1)\frac{\sum_{c\in\Omega} n_c(\mu(\mathbf{x}_{r,c}) - \mu(\mathbf{x}_r))^2}{\sum_{c\in\Omega}\sum_{i=1}^{n_c}(\mathbf{x}_{r,i,c} - \mu(\mathbf{x}_r))^2}$$

where n_c is the number of samples belonging to class $c \in \Omega$, and $\mathbf{x}_{r,i,c}$ is the rank of sample $\mathbf{x}_i$ of class c. The average rank of samples in class c is indicated by $\mu(\mathbf{x}_{r,c})$, while $\mu(\mathbf{x}_r)$ is the average rank of all samples.

Fisher Score

Feature selection by *Fisher Score* determines a subset of features that maximises the distance between classes while minimising the distance between samples of the same class [52].

The Fisher Score criterion function is given as [52]

$$J(\mathbf{x}) = \frac{\sum_{c\in\Omega} n_c(\mu(\mathbf{x}_c) - \mu(\mathbf{x}))^2}{\sum_{c\in\Omega} n_c\sigma(\mathbf{x}_c)^2}$$

Note the similarity between this method and the Wilcoxon Rank Sum approach described in the previous section.

2.4.2 Multivariate Filter Methods

Contrary to univariate filter methods, multivariate algorithms can detect predictive information from interacting features [53]. The *Relief*-based algorithms is a family of such multivariate filter methods.

ReliefF

The *ReliefF* algorithm represents the relevancy to some dependent feature y in terms of weights assigned to each feature. A weight of -1 indicates the least relevant feature, while a weight equal to 1 implies the opposite.

Two sets, M and H, are defined from the K nearest neighbours of a selected sample, $\mathbf{x}_i$ [53]. ReliefF calculates the distance between two features, $\mathbf{x}^{(j)}$ and $\mathbf{x}^{(l)}$, using the Manhattan metric

$$d\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = \sum_{i=1}^{n} \sqrt{\left(\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)})\right)^2}$$

All elements in H belong to the same class as $\mathbf{x}_i$, while M is the set complementary to H. That is,

$$M = \{\mathbf{x}_k \mid \mathbf{y}_k \neq \mathbf{y}_i\}_{k=1}^{K}$$
$$H = \{\mathbf{x}_k \mid \mathbf{y}_k = \mathbf{y}_i\}_{k=1}^{K}$$

The choice of K can be optimised, but is restricted to the smallest class

$$K \leq \min_{c \in \Omega} \{|c|\}$$

A feature weight, w, is calculated according to

$$w_{(t+1)} := w_t + \frac{1}{n \cdot K} \sum_{i=1}^{n} \sum_{k=1}^{K} d\left(\mathbf{x}, \mathbf{x}_i, H_k\right) - d\left(\mathbf{x}, \mathbf{x}_k, M_k\right) \tag{2.6}$$

where

$$d(\mathbf{x}, \mathbf{x}_i, I) = \frac{|\mathbf{x}_i - I|}{\max \mathbf{x} - \min \mathbf{x}}$$

represents the difference between all samples in $\mathbf{x}$ and a selected sample. The variable I is a placeholder for samples from either M or H.

According to Equation 2.6, if $\mathbf{x}_i$ belongs to the same class as the samples in M, $\mathbf{x}_i$ is assumed to be informative of $\mathbf{y}$ and w is increased [53]. On the contrary, w is decreased to signify that $\mathbf{x}_i$ is not considered predictive of $\mathbf{y}$. Since H and M contains the same K number of elements, ReliefF inherently corrects for class imbalance.

MultiSURF

The *MultiSURF* algorithm is based on the principles of ReliefF but with the number of neighbors, K, determined by the algorithm [54]

$$K = T_i - \frac{\sigma(\mathbf{x})}{2}$$

where T_i is the average Manhattan distance between a sample and all other samples. That is, MultiSURF considers all the observations within a distance of T_i from a selected sample rather than K selected samples. The H and M sets defined as described for ReliefF in the previous section, but the MultiSURF feature weight update is given by

$$w := w + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{d\left(\mathbf{x}, \mathbf{x}_i, H_k\right)}{|H|} - \frac{d\left(\mathbf{x}, \mathbf{x}_i, M_k\right)}{|M|}$$

where division by K is replaced by division with the number of elements, $|\cdot|$, in the H or M sets.

2.5 Classification

Classification algorithms seek to construct a discriminative function that organises observations into distinct groups [55]. Contrary to clustering, classification is a *supervised learning* method that utilises examples of sample memberships to categorise observations.

2.5.1 Quadratic Discriminant Analysis

The *Quadratic Discriminant Analysis* (QDA) classification algorithm aims to maximise separability between classes under the assumption that samples are normally distributed [56]. The Bayes theorem [57] is used to model the probability of an observation belonging to a particular class

$$P(\mathbf{y}_i = c \mid \mathbf{x}) \propto P(\mathbf{x} \mid \mathbf{y}_i = c)P(\mathbf{y})$$

Class priors are estimated from training data as the proportion of samples in each class [56]. Moreover, the probability distribution of samples belonging to a particular class, $P(x \mid y = c)$, is also assumed to be normal. Observations are assigned to the class that maximises the quadratic discriminative function

$$\frac{1}{2}\log|\mathbf{C}_c| - \frac{1}{2}(\mathbf{x} - \mu(\mathbf{x}_c))^T\mathbf{C}_c^{-1}(\mathbf{x} - \mu(\mathbf{x}_c)) + \log\frac{n_c}{n}$$

where $\mathbf{C}_c$ is the co-variance matrix of class c. Thus, QDA calculates one co-variance matrix for each class, which enables construction of both linear and quadratic decision surfaces.

Shrinkage

Shrinkage can be used to regularise the QDA model by using a penalised estimate for the co-variance matrix of the form [58]

$$\mathbf{C}(\alpha) = \alpha\mathbf{C} + (1 - \alpha)\sigma^2$$

The α parameter ranges between zero and one and can be used to adjust the degree of regularisation. In high-dimensional problems, shrinkage relaxes the correlation between features [59].

2.5.2 Support Vector Machine

The *Support Vector Machine* (SVM) constructs an N-dimensional hyper-plane to achieve maximal separation of samples according to some objective, such as classification [60]. In a classification problem, the optimal hyper-plane offers maximum separability between classes.

A linear hyper-plane is obtained from samples satisfying

$$\left\{\mathbf{x} : \mathbf{w}^T\mathbf{x} - b = 0\right\}$$

for which the parameters $\mathbf{w}$ and b are determined by the learning algorithm. The *Support Vector Classification* (SVC) optimisation problem can be expressed as

$$\min_{\mathbf{w},b,\xi} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \beta \sum_{i=1}^{n} \xi_i \right\}$$

subject to

$$\begin{cases} \mathbf{y}_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

The non-negative variables ξ are referred to as *slack* variables allowing samples to violate the decision boundary. The degree of decision boundary violation is controlled by the parameter β. The dual formulation of the SVC problem is given by

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \beta_i - \frac{1}{2} \sum_{i,j=1}^{n} \beta_i \mathbf{y}_i \beta_j \mathbf{y}_j K(\mathbf{x}_i, \mathbf{x}_j) \right\}$$

subject to

$$\begin{cases} 0 \leq \beta_i \leq \beta_U \\ \sum_i \mathbf{y}_i \beta_i = 0 \end{cases}$$

The dual formulation is derived by using *Lagrange* coefficients [61], β. Solving the SVC dual optimisation problem involves computation of feature dot products, which opens for application of the *kernel trick* [62]. The kernel trick involves implicitly mapping features to a higher dimensional space

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

by to a kernel function $\phi(\cdot)$. Table 2.1 lists a selection of kernel functions [63].

Table 2.1: Kernel functions for the *Support Vector Classification* algorithm [63]. Parameters: The intercept, r, of the sigmoid and polynomial kernel of degree d, and the positive scaling parameter γ.

Name	**Kernel**
Linear	$\mathbf{x} \cdot \mathbf{x}^T$
Polynomial	$\left(\gamma(\mathbf{x} \cdot \mathbf{x}^T) + r\right)^d$
Radial Basis	$\exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}^T\right\|^2\right)$
Sigmoid	$\tanh\left(\gamma(\mathbf{x} \cdot \mathbf{x}^T) + r\right)$

The β parameter from the dual SVC optimisation problem and the choice of kernel function can be optimised [60].

2.5.3 Logistic Regression

Logistic Regression is a linear classification model that assigns probabilities to different classes according to the logistic function [64].

The Logistic Regression optimisation objective depends on the choice of regularisation term. With L_1 regularisation, the optimisation becomes [64]

$$\min_{\mathbf{w},b} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \beta \sum_{i=1}^{n} \log\exp\left(-\mathbf{y}_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) + 1\right) \right\}$$

while for L_2 regularisation, the objective is

$$\min_{\mathbf{w},b} \left\{ \|\mathbf{w}\| + \alpha \sum_{i=1}^{n} \log\exp\left(-\mathbf{y}_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) + 1\right) \right\}$$

The α parameter controls the regularisation strength, which depends on the problem. Sparse solutions can be obtained with L_1 regularisation, which is an example of embedded feature selection.

2.5.4 Ridge Classification

Ridge classification corresponds to a regularised Ordinary Least Squares (OLS) problem [65].

The OLS problem can be formulated as [66]

$$\min_{\mathbf{w}} \left\{ \|\mathbf{x} \cdot \mathbf{w} - \mathbf{y}\|^2 \right\}$$

for which the weights, w, of the Ridge regression model are obtained by [65]

$$\min_{\mathbf{w}} \left\{ \|\mathbf{x} \cdot \mathbf{w} - \mathbf{y}\|^2 + \beta \, \|\mathbf{w}\|^2 \right\}$$

The parameter β controls the amount of regularisation imposed on the problem. Increasing β corresponds to enforce generality in the model, which enables the model to handle co-linearity.

The Ridge model can be used for binary classification where the predictions are given by [67]

$$\widehat{\mathbf{y}} = \begin{cases} 0 & \text{if } \mathbf{x} \cdot \mathbf{w} \leq 0 \\ 1 & \text{if } \mathbf{x} \cdot \mathbf{w} > 0 \end{cases}$$

which is equivalent to a linear decision surface.

2.5.5 K-Nearest Neighbors

The *K-Nearest Neighbors* (KNN) algorithm is an example of a lazy learner. Instead of learning a discriminative function, the algorithm memorises the training data [68]. Classification is performed by comparing samples to the K most similar observations from the memorised training data and assigning the dominant class label to this sample. In order to determine the K closest training observations to a new sample, the algorithm applies a distance function. Examples of such distance functions are given in Table 2.2.

Table 2.2: Functions to quantify the distance between samples in $\mathbf{x}$ and $\mathbf{x}_k$. The p parameter is arbitrary.

Name	**Distance Metric**
Euclidean	$\sqrt{\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{x}_k)^2}$
Manhattan	$\sum_{i=1}^{n}\sqrt{(\mathbf{x}_i - \mathbf{x}_k)^2}$
Chebyshev	$\max\left\{\sqrt{(\mathbf{x}_i - \mathbf{x}_k)^2}\right\}_{i=1}^{n}$
Minkowski	$\left(\sum_{i=1}^{n}\sqrt{(\mathbf{x}_i - \mathbf{x}_k)^2}^{p}\right)^{1/p}$

The KNN algorithm calculates the distances between an observation $\mathbf{x}_i$ and all the memorised observations $\mathbf{x}$. The K observations closest to $\mathbf{x}_i$ [68], and a prediction is made by a majority vote

$$\widehat{\mathbf{y}} = \text{mode}\,\{\mathbf{x}_k\}_{k=1}^{K}$$

That is, the dominant class among the K neighbours of a sample x_i is assigned to the sample. The parameter K and choice of distance function can be optimised for each problem.

2.5.6 Decision Tree

A *Decision Tree* infers a set of decision rules by recursive partitioning of features. Each decision rule is learned from the data based on a metric quantifying the quality of a partitioning. In classification trees, the leaves represent the class labels, while regression trees hold continuous values in the tree leaves [69].

Let $\left\{Q^{(m)}(\cdot)\right\}_{m=1}^{M}$ represent a set of feature queries performed at each node m in a tree of M nodes. Moreover, at each node, conditionals

$$q(\mathbf{x}, \tau_m) = \mathbf{x}^{(j)} \leq \tau_m$$

are imposed on a feature by some threshold τ_m to evaluate candidate splits [69]. These queries partitions the data into subsets, $Q_L^{(m)}(q)$ and $Q_R^{(m)}(q)$, given by

$$Q_L^{(m)}(q) = \{(\mathbf{x}_i, \mathbf{y}_i) \mid \mathbf{x}_i \leq \tau_m\}_{i=1}^{d}$$
$$Q_R^{(m)}(q) = Q \setminus Q_L(q)$$

To select a split from amongst all the candidate splits, each split is evaluated in terms of an *information gain* criterion given by Equation 2.7

$$I(Q^{(m)}, q) = I(Q^{(m)}, q) - \frac{n_L}{n} H(Q_L^{(m+1)}(q)) - \frac{n_R}{n} H(Q_R^{(m+1)}(q)) \tag{2.7}$$

where n_L and n_R is the number of samples in the nodes L and R resulting from a split operation, and $H(\cdot)$ is a measure of impurity. The Decision Tree objective is to maximise the information gain at each tree node

$$\widehat{q} = \arg\max_{q} I(Q, q)$$

and the choice of impurity function depends on whether the problem concerns classification or regression. Table 2.3 shows impurity measures for classification and regression problems [69].

Table 2.3: Decision tree impurity measures. Parameters: the proportion of class c at node m over a region R_m in the tree with n_m observations, $p_{m,c}$.

Objective	**Impurity**	H
Classification	Gini	$\sum_k p_{m,c} \cdot (1 - p_{m,c})$
	Entropy	$-\sum_k p_{m,c} \log p_{m,c}$
	Misclassification	$1 - \max p_{m,c}$
Regression	Mean Squared Error	$\frac{1}{n_m} \sum_{i \in n_m} (\mathbf{y}_i - \mu(\mathbf{y}_m))^2$
	Mean Absolute Error	$\frac{1}{n_m} \sum_{i \in n_m} \sqrt{(\mathbf{y}_i - \mu(\mathbf{y}_m))^2}$

The recursive partitioning of the data into subsets proceeds until each node contains a specific number of samples, or the recursion has reached a given depth [69]. Determining these constraints contributes to regularising the tree against over-fitting. Moreover, Decision Trees performs embedded feature selection by evaluating and splitting a subset of features.

2.5.7 Bootstrap Aggregation

Bootstrap Aggregation (bagging) combines multiple versions of a base model in parallel to improve the robustness over a single model. Each base model is built from a bootstrap sample of training data. Majority voting is used to produce a prediction and contributes to reducing the variance in the model [70].

Random Forest

The *Random Forest* algorithm combines Decision Trees as base models [71]. These base models are typically referred to as *weak learners*, which refers to a model with learning capacity similar to random guessing. Each split in a Decision Tree is typically performed with a random subset of features at each node, although this is not strictly necessary. Performing splits with random feature subset can potentially increase bias, but also reduce the variance. Predictions are obtained from a majority vote

$$\widehat{\mathbf{y}} = \text{mode}\left\{b_m(\mathbf{x})\right\}_{m=1}^{M}$$

over all M trees in the model.

Alternatives for selecting the size of the feature subset from p original features could be $\log p$ or $\sqrt{p}$ [67].

Extremely Randomised Trees

The *Extremely Randomised Trees* (ET) algorithm is based on the same principles as the Random Forest algorithm, but the optimal Decision Tree threshold, τ_m, is randomly selected at each split [72]. That is, the ET algorithm imposes random conditionals on a random subset of features at each node a Decision Tree base model. Random selection of conditional thresholds may contribute to reducing variance at the expense of increased bias.

2.5.8 Boosting

Contrary to bagging, boosting combines base models sequentially [70]. That is, each base model, $\{G_i(\cdot,\cdot)\}_{i=10}^{M}$, are combined to form an ensemble

$$f(x) = \sum_{m=1}^{M} \beta G_m(\mathbf{x}, \theta)$$

A set of weights β is used to regularise the contribution of each model, and to emphasise the models with the strongest predictive performance.

The model ensemble is built over M boosting rounds where model number m is selected according to [73]

$$G_m(\mathbf{x}) = G_{m-1}(\mathbf{x}) - \gamma_m \sum_{i=1}^{n} \nabla_G \mathcal{L}(\mathbf{y}_i, G_{m-1}(\mathbf{x}_i)) \tag{2.8}$$

which describes the *Gradient Boosting* procedure. The model selected in each boosting round is determined by minimisation of a criterion, $\mathcal{L}(\cdot, \cdot)$. The γ_m parameter, known as *learning rate*, is derived from

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{y}_i, G_{m-1}(\mathbf{x}_i) - \gamma \frac{\partial \mathcal{L}(\mathbf{y}_i, G_{m-1}(\mathbf{x}_i))}{\partial G_{m-1}(\mathbf{x}_i)}\right)$$

and represents the step length in the negative direction of the gradient in Equation 2.8 towards the minimum of $\mathcal{L}$ [74], [73]. Boosting may contribute to reduce both bias and variance compared the performance of a single base model [75].

A *Gradient Boosting Decision Tree* (GBDT) is an ensemble model consisting of *Classification and Regression Trees* (CART) base models [73]. A CART model differs from a DT in that the CART model holds prediction scores, instead of decision values, in each leaf. Moreover, models are trained sequentially in the GBDT scheme from the residuals of previous boosting rounds.

Extreme Gradient Boosting

The *Extreme Gradient Boosting* refers to a particular implementation of the GBDT algorithm based on CART models [76]. Compared to the gradient boosting optimisation objective, given by Equation 2.8, the Extreme Gradient Boosting objective also includes a regularisation term which gives

$$G_m(\mathbf{x}) = G_{m-1}(\mathbf{x}) - \gamma_m \sum_{i=1}^{n} \nabla_G \mathcal{L}(\mathbf{y}_i, G_{m-1}(\mathbf{x}_i)) + \sum_{j=1}^{m} R(G_j)$$

in boosting round m. The regularisation term, $R(G_j)$ penalises the complexity of the tree base model according to

$$R(G_j) = \gamma T + \frac{1}{2}\lambda \sum_{t=1}^{T} \mathbf{w}_t^2$$

where the γ and λ coefficients are arbitrary parameters, and $\mathbf{w}$ are the weights in the leaves of the tree model.

The Extreme Gradient Boosting also performs feature sub-sampling and weight scaling to prevent over-fitting, in addition to regularisation of the model [76]. Feature sub-sampling selects a subset of features when splitting samples at each tree node, analogous to Random Forest and Extremely Randomised Trees. Weight scaling adjusts the weights in the CART model leaves in each step of the boosting. Similarly to the learning rate parameter, scaling reduces the influence of individual base models and leaves space for future trees to improve the model.

Light Gradient Boosting Machine

The *Light Gradient Boosting Machine* (LGBM) algorithm is, similar to the XGB algorithm, also based on the GBDT procedure [77], [78]. The LGBM extends GBDT with a *Gradient-based One-Side Sampling* (GOSS) procedure and an *Exclusive Feature Bundling* (EFB) procedure to improve computational efficiency and embedded feature selection.

In LGBM, a new candidate model for the ensemble is constructed from a subset of samples rather than the complete training set [77], [78]. This sample subset is selected by the GOSS method according to the n_v largest gradients, $\nabla_G \mathcal{L}(y_i, G_{m-1}(x_i))$, given in Equation 2.8. Moreover, a random subset of n_s samples, independent of gradient magnitude, are combined with the initial subset. The GOSS procedure assumes that sample gradients are proportional to the contribution of each sample to the information gain, given in Equation 2.7. A constant

$$\frac{1 - n_v}{n_s}$$

is used to adjust for changes in the sample distribution after sub-sampling observations to calculate the information gain [77], [78].

The EFB method performs feature down-sampling, or *bundling* [79]. More specifically, the EBF algorithm combines features with similar values into one single feature. A threshold is used to determine feature similarity. The computational efficiency and potential feature redundancy can be reduced by merging features that have been grouped together [77], [78].

2.6 Model-Based Estimation of Feature Relevance

Estimation of feature relevance, or feature importance, may improve the understanding of the model behaviour. Some methods of feature importance quantification, such as *gain* and *split count*, have been shown to be *inconsistent* [80]. Inconsistency means that a highly ranked feature may be perceived as more important than features receiving a lower rank.

2.6.1 Shapley Additive Explanations

The *Shapley Additive Explanations* (SHAP) method was proposed by Lundberg, Erion, and Lee (2018) as a consolidated measure of feature importance. This method builds on Shapley values from game theory, which has been used to quantify the contribution of each participant in collaborative games [81].

Let the expected prediction, $\mathbb{E}_{\widehat{\mathbf{y}}}$, of a model, $\lambda(\phi, \cdot)$, trained on a subset of features, $\widetilde{\mathbf{X}} \subset \mathbf{X}$, be defined as

$$\mathbb{E}_{\widehat{\mathbf{y}}}\left(\widetilde{\mathbf{X}}\right) \equiv \mathbb{E}\left\{\lambda(\phi, \mathbf{X}) \mid \widetilde{\mathbf{X}}\right\}$$

where $\mathbf{X}$ represents the original set of p features. An algorithm for estimation of $\mathbb{E}_{\widehat{\mathbf{y}}}$ is outlined in the paper by Lundberg, Erion, and Lee (2018) [81]. The SHAP feature importance measure, ν_j, for a feature $\mathbf{x}^{(j)}$ is calculated by

$$\nu_j = \sum_{\widetilde{\mathbf{X}} \subseteq \mathbf{X} \setminus \left\{\mathbf{x}^{(j)}\right\}} \frac{\left|\widetilde{\mathbf{X}}\right|!(p - \left|\widetilde{\mathbf{X}}\right| - 1)!}{p!} \left(\mathbb{E}_{\widehat{\mathbf{y}}}\left(\widetilde{\mathbf{X}} \cup \left\{\mathbf{x}^{(j)}\right\}\right) - \mathbb{E}_{\widehat{\mathbf{y}}}\left(\widetilde{\mathbf{X}}\right)\right) \tag{2.9}$$

In Equation 2.9, the difference in expected model predictions from including and excluding a feature $\mathbf{x}^{(j)}$ is calculated. Note that the order of which features are included in $\widetilde{\mathbf{X}}$ may affect the importance estimates. Therefore, all possible permutations of the feature subset are evaluated. The final feature importance estimate is then averaged across the importance estimate of each permutation. Hence, a SHAP value represents the average feature contribution to a model prediction.

2.7 Comparing Prediction Models

Hyperparameter configurations affect the ability of a model to learn patterns [82]. According to the *No Free Lunch* theorems [23] should the performance of different models be compared in order to select the optimal one for a problem.

2.7.1 Model Selection

Model selection refers to the task of selecting a model by evaluating different combinations of hyperparameters. Given an optimisation criterion, $\mathcal{L}(\cdot, \cdot)$, the optimal model $\lambda(\widetilde{\phi}, \cdot)$ is selected according to

$$\lambda(\widetilde{\phi}, \mathbf{x}) = \arg\min_{\phi \in \mathbf{\Phi}} \mathcal{L}(\mathbf{y}, \lambda(\phi, \mathbf{x}))$$

The function $\mathcal{L}(\cdot, \cdot)$ is used to quantify the model performance for different parameter configurations $\phi \in \mathbf{\Phi}$ [83].

Bayesian approaches to model selection [84] have demonstrate superior performance compared to techniques such as random search [85], and evolutionary [86] and gradient-based methods [87].

Sequential Model-Based Optimisation

The Bayesian protocol for model selection is formalised through the *Sequential Model-Based Optimisation* (SMBO) algorithm. The SMBO algorithm constructs a surrogate model, $\mathcal{M}$, to predict the behaviour of a target algorithm, λ. By modelling λ, a computationally more efficient model, namely $\mathcal{M}$, can be used to select configurations for λ without requiring explicit evaluation of λ. Algorithm 1 outlines the SMBO protocol [88], [86].

Algorithm 1 Sequential Model-Based Optimisation

```
    Input: Learning algorithm, $\lambda$, hyperparameter domain $\mathbf{\Phi}$, training set, $\mathbf{x}$, ground truth, $\mathbf{y}$.
    Output: A hyperparameter configuration, $\widetilde{\phi}$.
procedure SMBO($\lambda$, $\mathbf{\Phi}$, $\mathbf{x}$, $\mathbf{y}$)
    for $t \in T$ do
        $\phi_t^{(*)} \leftarrow \underset{\phi \in \mathbf{\Phi}}{\text{argmax}} \{A(\phi, \mathcal{M}_{t-1})\}$
        $\mathcal{L}_t \leftarrow \mathcal{L}\left(\mathbf{y}, \lambda\left(\phi_t^{(*)}, \mathbf{x}\right)\right)$
        $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup (\phi_t^{(*)}, \mathcal{L}_t)$
        $\mathcal{M}_t \leftarrow \mathcal{M}_{t-1}(\mathcal{D})$
    $\widetilde{\phi} \leftarrow \arg\min_{t \in T}\ \mathcal{L}_t$
    return $\widetilde{\phi}$
```

In Algorithm 1, an acquisition function, $A(\cdot)$, is used to select hyperparameter configurations for the surrogate model. Over a budget of T iterations, the surrogate determines a candidate configuration, $\phi_t^{(*)}$, that is given to the target algorithm, λ. The predictions of the model, $\lambda(\phi_t^{(*)}, \cdot)$, is compared to the ground truths, $\mathbf{y}$, using an optimisation criterion, $\mathcal{L}$. This quantified model performance, $\mathcal{L}_t$, is added to the historical set along with the evaluated configuration, $\mathcal{D} := \mathcal{D} \cup (\phi_t^{(*)}, \mathcal{L}_t)$. The historical set is used to improve the predictions of the surrogate model [88].

Surrogate models that can be used with SMBO includes Gaussian process [89], Random Forest [88] or the Tree-Structured Parzen estimator [90].

Sequential Model-based Algorithm Configuration

The *Sequential Model-based Algorithm Configuration* (SMAC) protocol selects parameter configurations using a Random Forest (RF) of regression trees as surrogate. The RF model builds a set of B regression trees using T bootstrap samples from the historical set, $\mathcal{D}$. A randomly selected subset, $\phi_q \subset \phi$, of parameters is used to evaluate node splits. The number of parameters in each subset used to consider a split is calculated as

$$q = \lceil |\phi| \cdot r \rceil \tag{2.10}$$

for a constant r [88]. Using a RF as surrogate model opens for conditional constraints to be imposed on the hyperparameter space. Consider, for instance, the kernel function of an SVC model. Optimising the degree of a polynomial kernel is only relevant if the kernel has been selected to be a polynomial in the first place. Such conditional constraints are not supported with Gaussian process surrogates [84].

The Acquisition Function

In SMAC, the *Expected Improvement* (EI) criterion is used in hyperparameter acquisition [91], [88]. The EI defines a balance between exploring new areas in the parameter space and exploiting areas that are already known to be favorable configurations.

Let $\mathcal{L}_t$ denote the highest performance obtained by the surrogate model at step t given by

$$\mathcal{L}_t = \mathcal{M}_t(\phi_t, \mathcal{D}_t)$$

Moreover, let $u(\phi)$ be a function defined as

$$u(\phi) \equiv \max\{0, \mathcal{L}_{t-1} - \mathcal{L}_t\}$$

Recall that the surrogate objective is to minimise $\mathcal{L}$. The situation where $u(\phi)$ is greater than zero corresponds to an improvement upon the previous configuration

ϕ_{t-1}, since configuration ϕ_t leads to a reduction in $\mathcal{L}_t$ compared to $\mathcal{L}_{t-1}$. On the contrary, no improvement has been made if $u(\phi)$ equals zero.

The EI acquisition function is given as the expectation of $u(\phi)$ according to

$$\mathrm{EI}(\phi) = \mathbb{E}\left[u(\phi_t) \mid \phi_{t-1}, \mathcal{D}_{t-1}\right] = \int_{\infty}^{\mathcal{L}_{t-1}} (\mathcal{L}_{t-1} - \mathcal{L}_t) P(\mathcal{L}_t \mid \phi_{t-1}; \mathcal{D}_{t-1}) d\phi \tag{2.11}$$

Hence, parameter candidates are selected by which are expected to maximise the improvement upon $\mathcal{L}_{t-1}$.

The Predictive Distributions

The EI assumes that the predictive distribution of the surrogate model, $P(\mathcal{L} \mid \phi; \mathcal{D})$, for a configuration, ϕ, is Gaussian [91]

$$P(\mathcal{L} \mid \phi; \mathcal{D}) \sim \mathcal{N}(\phi \mid \mu; \sigma^2)$$

An advantage of using a Gaussian predictive distribution is that it gives a closed-form expression for Equation 2.11.

A Gaussian predictive distribution is also assumed for the RF surrogate [88]. From the predictions, $b(\mathbf{x})$, of the individual regression trees, the empirical mean, $\widehat{\mu}$, and variance, $\widehat{\sigma}^2$, given by

$$\widehat{\mu} = \frac{1}{|B|} \sum_{b \in B} b(\mathbf{x})$$
$$\widehat{\sigma}^2 = \frac{1}{|B| - 1} \sum_{b \in B} (b(\mathbf{x}) - \widehat{\mu})^2$$

is calculated to condition the predictive distribution of the surrogate according to

$$P(\mathcal{L} \mid \phi; \mathcal{D}) \sim \mathcal{N}(\mathcal{L} \mid \widehat{\mu}; \widehat{\sigma}^2)$$

That is, the prior distribution for the EI and SMAC protocol is defined as a Gaussian distribution.

2.7.2 Stratified K-Fold Cross Validation

Random partitioning procedures create artificial training and validation data from the original training set that can be used to assess the average model performance. Stratified K-fold cross-validation (CV) is one approach to perform such random partitioning [92].

The K-fold CV protocol partitions a data set, $D = \{(\mathbf{X}_j, \mathbf{y}_j)\}_{j=1}^{n}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$, into K equally sized subsets,

$$D = \left\{ D_k \in \mathbb{R}^{\frac{n}{K} \times (p+1)} \right\}_{k=1}^{K}$$

These training and validation sets are created by selecting one subset to be the validation set, $D_V = D_k$, while the $K - 1$ other subsets are combined into a training set $D_T = D \setminus D_V$. Each subset is selected once as the validation set. The CV estimate, $\mathcal{L}_{\text{CV}}$, is given as the average model performance over the K validation sets

$$\mathcal{L}_{\text{CV}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}(\mathbf{y}_k, \widehat{\mathbf{y}}_k)$$

where $\mathbf{y}$ is the ground truth, and $\widehat{\mathbf{y}}$ represents the model prediction. In a classification problem, stratification ensures that the proportion of classes in the original data set is reflected in each fold.

2.7.3 General Model Performance Estimation

The general model performance is a fundamental concept in statistical learning theory and refers to the difference between the model training and validation error [93]. Carrying out both model selection and evaluation using the same CV fold gives optimistically biased estimates. However, a nested CV protocol can be used to assess the general performance, or error, of a model [94], [95].

From the training, D_T, and validation, D_V, partitions of K-fold CV, the nested CV scheme creates a second set of validation and training partitions using D_T [94], [95].

These data subsets are used in (1) model selection, and (2) performance estimation. Algorithm 2 outlines the nested CV protocol.

Algorithm 2 Model Performance Estimation

Input: Learning algorithm λ, optimisation criterion, $\mathcal{L}$, model selection protocol, SMBO, hyperparameter domain $\mathbf{\Phi}$, data set, D.

Output: The general model error estimate, $\overline{\pi}$.

```
procedure NestedKFoldCV($\lambda$, SMBO, $\mathcal{L}$, $\mathbf{\Phi}$, $D$, $\mathcal{M}$,)
    $\pi \leftarrow \emptyset$
    for $k \in [1, K]$ do                                  ▷ Outer cross validation loop.
        $D_V \leftarrow D_k$
        $D_T \leftarrow D \setminus D_V$
        $\widetilde{\phi} \leftarrow$ SMBO$(\lambda, \Phi, D_T)$            ▷ Inner cross validation loop.
        $\lambda \leftarrow \lambda\left(\widetilde{\phi}, D_T\right)$                              ▷ Train model.
        $\pi_k \leftarrow \mathcal{L}\left(\mathbf{y}_V, \lambda\left(\widetilde{\phi}, D_V\right)\right)$
    return $\overline{\pi} = \frac{1}{K}\sum_k \pi_k$
```

Algorithm 2 begins with creating training and test sets in the outer CV loop, and assumes that CV model selection is included in the SMBO procedure. That is, for each configuration proposed by SMBO, the performance of the model given this configuration is a CV estimate. However, any model selection procedure may be used with Algorithm 2.

2.7.4 A Non-Parametric Confidence Interval

To obtain a confidence interval (CI) around a model performance estimate, some methods require that these performance estimates are normally distributed. However, the bootstrap method described by Wang (2001) is a non-parametric approach to construction of CI estimates [96].

Assuming a set of model scores, $\mathcal{L} = (\mathcal{L}_1, \cdots, \mathcal{L}_T)$, has been obtained from T experiments. Let $B = (B_1, \cdots, B_K)$ represent K bootstrap samples, each of T scores, sampled from $\mathcal{L}$. Let μ_B and σ_B denote the mean and standard deviation of B given by

$$\mu_B = \frac{1}{K}\sum_{k=1}^{K}\mu(B_k)$$

$$\sigma_B = \sqrt{\frac{\sum_{t=1}^{T}\left(\mathcal{L}_t - \mu(\mathcal{L})\right)}{T-1}}$$

Thus, a CI estimate with $(1-\alpha)$ confidence level is given by

$$\mu_B \pm Z_{\alpha/2}\sigma_B$$

where $Z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution [96].

Chapter 3

Materials and Methods

The two goals of this thesis were (1) to study methods for radiomics data analysis, and (2) exploration of features to identify biomarkers of *disease-free survival* in *head and neck cancers*. It was hypothesised that a model superior in classifying patient treatment outcomes would have recognised the prognostic value of each feature. By investigating which features were utilised by this model, potential biomarkers could be identified.

This chapter use the notation and definitions from Chapter 2, in addition to the following. An image is represented as a stack of $A \in \mathbb{R}^{L \times C}$ *slices*. Each slice represents is matrix of size $L \times C$ with intensity values. Thus, each image represents a three-dimensional volume, $\mathbf{I} \in \mathbb{R}^{A \times L \times C}$, which is also referred to as a *stack*. An element of this image volume is referred to as a *voxel*.

Radiomics feature extraction is indicated by application functions, $F_i : \mathbb{R}^3 \mapsto \mathbb{R}$, to an image, $\mathbf{I}$, which results in scalar values. These values represent different properties of the image, such as the average image intensity.

Abbreviations of radiomics texture feature categories are given in table 3.1.

Table 3.1: Abbreviations of radiomics texture feature categories.

Abbreviations			
GLCM	Gray Level Co-occurrence Matrix	GLSZM	Gray Level Size Zone Matrix
NGTDM	Neighbouring Gray Tone Difference Matrix	GLRLM	Gray Level Run Length Matrix
GLDM Gray Level Dependence Matrix			

3.1 Software

All procedures for data analysis were implemented using the *Python*^{TM} [97] programming language, version 3.6.2., and made publicly available via the *GitHub*$^{©}$ web-based hosting service. The scripts were combined into a package named *biorad*, and can be accessed from `https://github.com/gsel9/biorad` [25].

The main protocols included in biorad concerns (1) radiomics feature extraction, described in Section 3.4, and (2) model comparison experiments, described in Section 3.8. The location of the files containing the code for these procedures, relative to the main folder of the package, is illustrated in Figure 3.1.

Note that Figure 3.1 does not include all the contents of biorad, but highlights the relevant files for radiomics feature extraction, and classification experiments.

The material for carrying out radiomics feature extraction is located in *feature_extraction.py*. This material builds on the *PyRadiomics* [16] package that was used to calculate the radiomics features. An application of feature extraction functionality can be found in the *Jupyter Notebook* [98] named, *feature_extraction.ipynb*. Examples on how to configure the feature extraction procedure is given in *parameter_files*. More information on the feature extraction settings is given in the PyRadiomics documentation [99].

The material required to perform a model comparison experiment is located in the *experiments* folder. A setup for experiments is available in *main.py*. Moreover, an explanation of the contents in this file is given in Appendix D. The *algorithms* folder

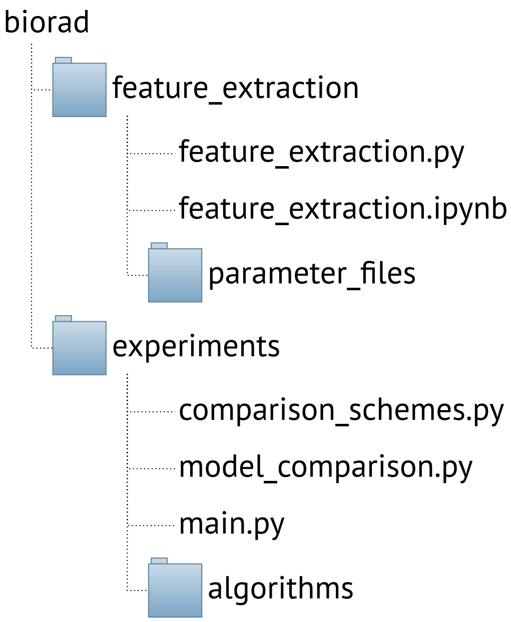


Figure 3.1: The *biorad* [25] package folder tree.

contains the implementations of the classification and feature selection algorithms used in this thesis.

3.2 Hardware

Radiomics feature extraction, described in Section 3.4, was performed with a MacBook Pro (13-inch, 2015) with a 1.6 GHz Intel® Core i5 processor and 8 GB memory.

Classification experiments, described in Section 3.8, were carried out with a Lenovo ThinkStation P720 with 20 x 2.20 GHz Intel® Xeon® processor and 129 GB memory.

3.3 The Data Set

The data set used in this thesis included clinical factors, PET parameters and pre-treatment *18F-fluorodeoxyglucose Positron Emission Tomography/Computed Tomography* (PET/CT) images of 198 head and neck cancer patients. These cancer patients received radiotherapy at the Oslo University Hospital between January 2007 and December 2013. Further details on the patient cohort and the image acquisition procedure are available in Moan et al. (2019) [100].

3.3.1 Images

The images had been obtained with a Siemens Biograph 16 PET/CT scanner, as described by Moan et al. (2019) [100]. All CT images were contrast-enhanced to make the tumour more visible. The PET images had been filtered after reconstruction with a Gaussian kernel using a *Full Width at Half Maximum* [101] of 3.5 mm. Image sizes ranged from 341 × 341 × 341 mm^3 to 682 × 682 × 396 mm^3. Isotropic voxels of 1 mm^3 had been obtained from co-registration of the PET and CT images onto a common image frame. The original PET and CT spatial resolutions were 3 × 3 × 2 mm^3 and 1 × 1 × 2 mm^3, respectively. The CT images contained *Hounsfield units* [102] shifted by -1024, while PET images contained *standardised uptake values* (SUVs) [103].

Supplementing each PET/CT was a binary mask image that included only the region of the tumour volume (ROI). Figure 3.2 illustrates how this binary image was used in segmentation of the ROI.

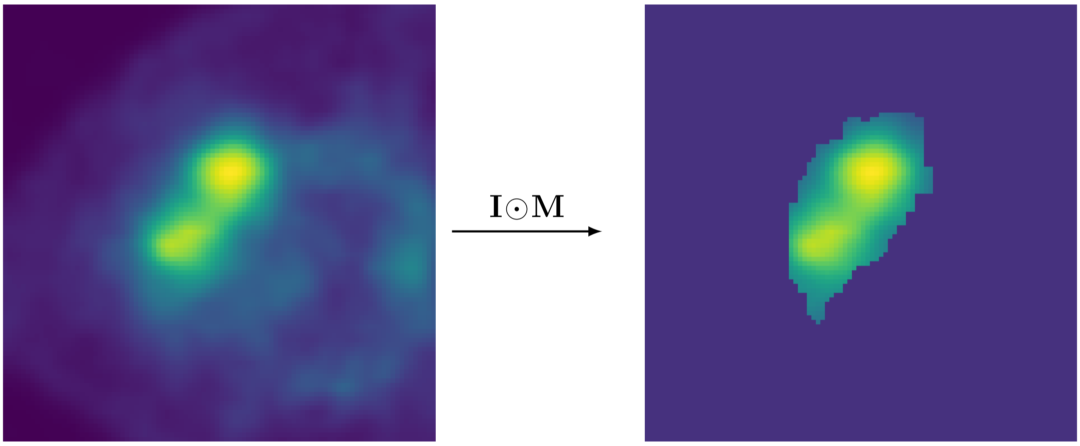


Figure 3.2: Segmentation of the tumour region in a PET slice, $\mathbf{I}$, by element-wise multiplication with a binary image, $\mathbf{M}$.

As illustrated in Figure 3.2, the ROI of each CT and PET was segmented by multiplying each image with the corresponding mask image.

3.3.2 Clinical Factors

A summary of the patient tumour and pre-treatment characteristics, referred to as *clinical factors*, are given in Table 3.2.

Table 3.2: The median, minimum and maximum values of selected tumour and pre-treatment characteristics of the patient cohort.

Factor	Description	
Total number of patients		198
Age (years)		60, (40, 80)(1)
Gender	Male	50 (25 %)
	Female	148 (74 %)
Tumour stage	T1/T2	96 (48 %)
	T3/T4	102 (52 %)
Tumour site	Oral cavity	17 (9 %)
	Oropharynx	144 (73 %)
	Hypopharynx	16 (8 %)
	Larynx	21 (10 %)
Tumour volume (cm^3)		14.7, (0.800, 285)(1)
HPV status	Positive	83 (42 %)
	Negative	18 (9 %)
	Unknown	97 (49 %)

(1): median, (minimum, maximum)

All the clinical factors included in the data set are summarised in Table A.1 in Appendix A.

3.3.3 PET Parameters

The PET parameters had been obtained from the ROI of each PET image as described by Moan et al. (2019) [100]. Among these parameters were only *SUV peak*, [104], *metabolic tumor volume* (MTV) [105] and *total lesion glycolysis* (TLG) [100] used in this thesis. The SUV peak was defined as the highest mean SUV within a 1 cm^3 spherical subregion of the ROI. Calculation of MTV had been performed based on the voxels in the ROI corresponding to an intensity greater or equal to 41 % of the SUV peak. The TLG was calculated according to [100]

$$\text{TLG} = \text{MTV} \cdot \mu\left(\text{SUV}_{\text{MTV}}\right)$$

where $\mu\left(\text{SUV}_{\text{MTV}}\right)$ denotes the mean SUV of the voxels used to calculate MTV.

The tumour volume and maximum SUV were also available PET parameters, but these were not used due to the resemblance with radiomics features. That is, features in the radiomics shape and first-order categories, described in Section 3.4, included both tumour volume and maximum SUV.

3.3.4 Patient Treatment Response

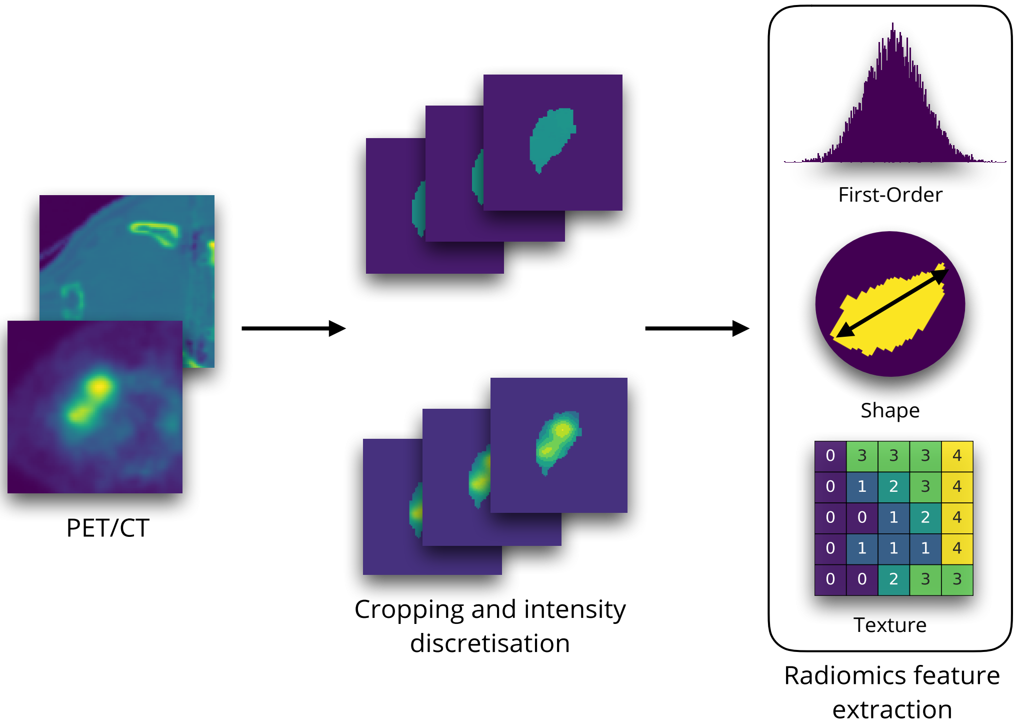
The studied clinical endpoint, namely disease-free survival [21], was described by an indicator function

$$\mathbf{y}_i = \begin{cases} 0 & \text{if disease-free survival} \\ 1 & \text{otherwise} \end{cases} \tag{3.1}$$

where $\mathbf{y}_i$ denotes the treatment response of patient i. Among the 198 patients constituted disease-free survival about 68 $\%$ of the outcomes.

3.4 Radiomics Feature Extraction

Figure 3.3 illustrates the protocol used in this thesis to extract radiomics features from PET and CT images of the 198 patients with head and neck cancers.


First-Order
Shape
Texture
PET/CT
Cropping and intensity discretisation
Radiomics feature extraction

0	3	3	3	4
0	1	2	3	4
0	0	1	2	4
0	1	1	1	4
0	0	2	3	3

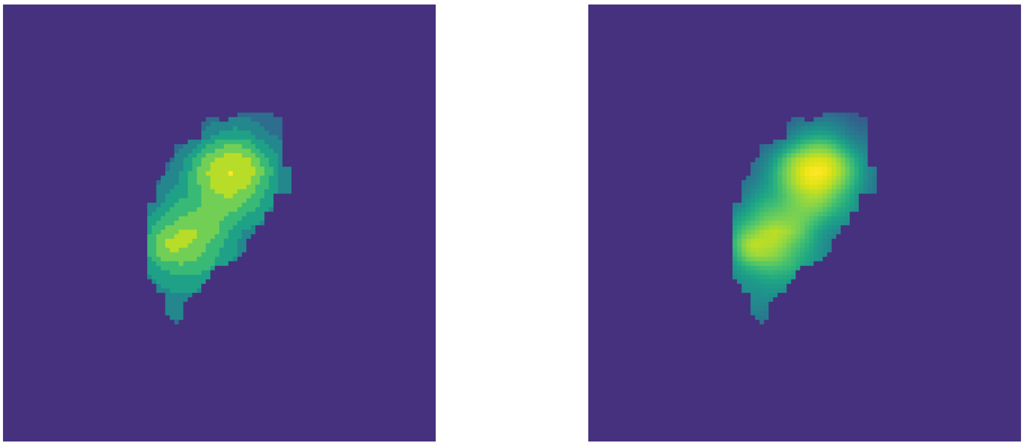
Figure 3.3: The main steps in extraction of radiomics feature from PET and CT images.

The steps in Figure 3.3 represent

1. segmentation of the image ROI (Section 3.3.1)
2. discretisation of image intensities into a fixed number of bins (Section 3.4.1)
3. and calculation of radiomics features, including ROI shape characteristics, first-order statistics and image textures (Section 3.4).

3.4.1 Image Discretisation

Discretisation of the image intensities can be used to optimise image texture features [106], [107]. Different discretisation schemes produce different images, which in turn give rise to different sets of features. Moreover, binning of image intensities has shown to be contributing to increase feature stability and reduce noise [108]. Figure 3.4 illustrates the effect of image intensity discretisation using 32 and 128 intensity bins.



(a) 32 bins.

(b) 128 bins.

Figure 3.4: Discretisation of PET intensities using 32 and 128 intensity bins.

Figure 3.4 illustrates how courser image textures are obtained by using a smaller number of intensities. Thereby, different texture characteristics can be obtained.

Images were discretised only before calculation of first-order and texture features. Binning of intensity values was performed for each image according to Equation 3.2 [108]

$$\boldsymbol{I}_{b,k} = \left\lfloor \frac{\boldsymbol{I}_k - \min \boldsymbol{I}}{w} \right\rfloor + 1 \tag{3.2}$$

where $\min \boldsymbol{I}$ is the minimum intensity of an image stack, $\boldsymbol{I}_k$ is an image voxel and w represents the intensity bin width. Application of Equation 3.2 produces an image,

$\boldsymbol{I}_{b,k}$, with b intensities. A recommended approach to discretisation is to use a fixed width, w [108]. Contrary to a fixed bin count, a fixed bin width has been shown to improve the reproducibility and comparability of PET features [109]. A bin width, w_b, corresponding to b intensity bins was calculated with

$$w_b = \frac{\max \boldsymbol{I} - \min \boldsymbol{I}}{b}$$

where $\max \boldsymbol{I}$ is the maximum intensity in the image. Given n stacks, $\boldsymbol{I_1}, \cdots, \boldsymbol{I_n}$, the bin widths were averaged to obtain comparable intensity distributions between the images according to

$$\overline{w} = \frac{1}{b \cdot n} \sum_{i=1}^{n} (\max \boldsymbol{I}_i - \min \boldsymbol{I}_i) \tag{3.3}$$

where $\overline{w}$ is the average bin width over n stacks used to discretise the PET and CT images. Three bins widths, corresponding to 32, 64 and 128 intensity bins, were used in this thesis.

3.4.2 Calculation of Radiomics Features

Radiomics features corresponding to the definitions in Appendix E were extracted from the PET and CT images discretised into 32, 64 and 128 bins. Table 3.3 shows the total number of features in each category that were calculated from each set of images.

Table 3.3: The number of radiomics features extracted in this thesis according to feature category. Abbreviations are defined in Table 3.1.

		Texture				
Shape	**First-Order**	GLCM	GLRLM	GLDM	NGTDM	GLSZM
14	18	24	16	14	5	16

In total: 107

As shown in Table 3.3, the texture category includes sub-categories of 75 features in total.

Shape Features

Radiomics shape features describe the three-dimensional rendering and voxel geometry of the ROI [108]. Moreover, these features are independent of intensity values [16].

In this thesis, the shape features were calculated from the binary mask images described in Section 3.3.1. Thus, these features were independent of the PET and CT images.

First-Order Statistics

First-order features represent the distribution of image intensities [108]. These histogram-based features describe properties such as symmetry and dispersion of intensities and depend on the number of bins used to create the histogram [16].

Six sets of first-order features were extracted in this thesis. Each set of features was derived from either PET or CT, discretised into 32, 64 or 128 bins.

Texture Features

Image texture features describe the spatial relationship between image intensities. These features can be used to quantify intratumour heterogeneity. Different descriptive matrices are typically used to study different aspects of image texture [108], [16], [110].

In this thesis, six sets of each texture feature sub-category were extracted. These feature sets were derived from PET and CT discretised into 32, 64 or 128 bins.

A Z-scoring [45] of the CT intensities was carried out before calculation of texture features. The Z-scoring was performed according to

$$\mathbf{I}_z = \frac{\mathbf{I} - \mu\left(\mathbf{I}\right)}{\sigma\left(\mathbf{I}\right)} \tag{3.4}$$

where $\mathbf{I}_z$ is the transformed image stack, and $\mu\left(\mathbf{I}\right)$ and $\sigma\left(\mathbf{I}\right)$ is the average and standard deviation of the stack, respectively. By Z-scoring the CT images, the intensity resolutions become comparable across different tumours, which is a recommended approach to image texture analysis [19], [111]. On the contrary, the PET images were not Z-scored in order to preserve the intensity variations assumed to maintain a direct relationship with the tumour biology.

Gray Level Co-occurrence Matrix

An entry in the *Gray Level Co-occurrence Matrix* (GLCM) represents the number of times a combination of two intensities occurred within a distance δ, and along an angle θ [16]. That is, the GLCM quantifies the frequency of co-occurring combinations of intensities in a particular direction for a given distance.

During feature extraction, δ was equal as the default configuration in PyRadiomics [112]. This choice of δ corresponded to 26 voxels in each three-dimensional neighbourhood.

Gray Level Size Zone Matrix

The GLSZM counts the number of intensity zones in an image. A zone refers to the number of connected voxels with the same intensity value [113]. Two voxels are defined as connected if the spatial distance between these two voxels equals one [16]. Note that intensity discretisation may be a pre-requisite to calculate features such as GLSZM in order to obtain any zones.

Gray Level Run Length Matrix

A GLRLM quantifies the number of consecutively connected voxels with the same intensity [114]. Such a sequence of voxels with the same intensity is referred to as a *run*. Similar to GLSZM features, intensity discretisation may be required to identify any voxel runs. Each entry (i, j) of the GLRLM represents the number of times the intensity i occurred with length j in a direction θ of the image [16]. The GLRLM features that were calculated for different angles was combined by averaging over all angles.

Neighbouring Gray Tone Difference Matrix

The NGTDM describes the difference between the intensity of a voxel and the average intensity in a neighbourhood of voxels. The NGTDM matrix holds the sum of absolute differences for each image intensity value.

Gray Level Dependence Matrix

A GLDM quantifies dependencies between image intensities [108]. The dependency between intensities are defined as the number of connected voxels within a distance δ that depend on a centre voxel with intensity i [16]. A voxel with intensity j, neighbouring a voxel with intensity i, is considered dependent on i if

$$\sqrt{(i - j)^2} \leq \alpha$$

for some threshold value α.

In this thesis, α was equal to zero. This setting was motivate by the size of the smallest tumour, which is given in Table 3.2. The δ was arbitrarily selected as one, which was the default configuration in PyRadiomics [112]. Following from this choice of δ, each voxel neighbourhood comprised 26 voxels in three dimensions.

3.4.3 Feature Post-Processing

After the extraction of radiomics features, those features that appeared to be invariant of image intensity discretisation were grouped according to their definitions and replaced by the group average.

Let $F(\mathbf{I}_b)$ be a feature extracted from an image, $\mathbf{I}$, discretised into b intensities. Thereby, a group of features was defined as

$$\mathbf{G} \equiv \{F(\mathbf{I}_{32}), F(\mathbf{I}_{64}), F(\mathbf{I}_{128})\}$$

Each group included three versions of a feature extracted from images discretised into 32, 64 and 128 intensity levels. An accumulated variance was calculated according to

$$\tau \equiv \frac{1}{2} \sum_{F \in \mathbf{G}} \left(F - \mu\left(\mathbf{G}\right)\right)^2 \tag{3.5}$$

where $\mu\left(\mathbf{G}\right) \in \mathbb{R}^n$ represents the mean across the features in a group

$$\mu\left(\mathbf{G}\right) = \frac{1}{|\mathbf{G}|} \sum_{F \in \mathbf{G}} F$$

The value τ, calculated with Equation 3.5, was compared to an arbitrary selected threshold of 10^{-15} according to

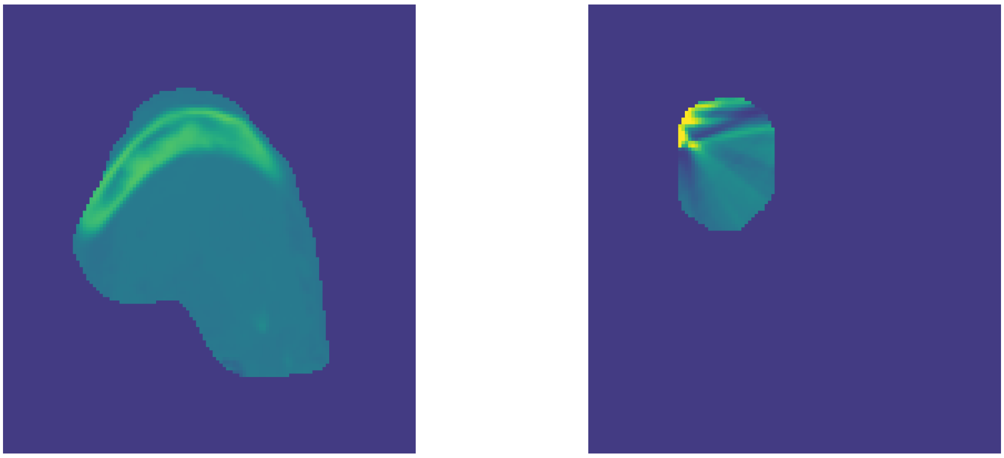
$$\mathbf{G} := \begin{cases} \mu\left(\mathbf{G}\right) & \text{if } \tau \leq 10^{-15} \\ \mathbf{G} & \text{if } \tau > 10^{-15} \end{cases} \tag{3.6}$$

Equation 3.6 was used to settle if the features in G should be replaced by their average in order to reduce the redundancy. If τ was less than the pre-defined threshold of 10^{-15}, the features in G were replaced by the average across the group members.

Note that the variance between the features in a group could also have be determined with the *Intraclass Correlation Coefficient* [32], described in Section 5.4.1.

3.4.4 Removing Image Artefacts

An additional set of images was created from the set by removing CT slices, and PET and CT images that contained streaks from dental fillings and bone structures inside the tumour region. Both phenomena are referred to as *artefacts*. Figure 3.5 shows an example of such artefacts found in CT slices.



(a) A bone structure.

(b) Streaks.

Figure 3.5: Examples of artefacts found in the tumour region of two CT slices.

In a head and neck cancer study of CT artefacts, Ger et al. (2018) proposed to remove only the slices influenced by artefacts from the CT images, instead of excluding the complete stack from the data set [115]. Ger et al. (2018) found that up to 50 % of the original ROI could be removed before image features changed significantly according to a *pairwise t-test* [51].

Detecting and Removing Artefacts

To detect artefacts in CT slices, the intensity values in each slice was compared to a predefined range of intensities. This range was assumed to cover only the intensities that were not found in artefacts. Thus, the goal of this range was not to capture only true ROI intensities, but to exclude intensities that could originate from artefacts.

The range, ζ, was defined as 100 intensities from the median, $\widetilde{I}$, of the ROI. That is,

$$\zeta \equiv \left[\widetilde{I} - 100, \widetilde{I} + 100\right]$$

The width of 200 intensities was arbitrarily selected after studying the distributions of a random selection of slices. Each image slice, $I(a, L, C)$, was evaluated according to an indicator function

$$\mathbb{I}(I(a, L, C)) = \begin{cases} 0 & \text{if } I_k \in \zeta \\ 1 & \text{if } I_k \notin \zeta \end{cases} \tag{3.7}$$

where I_k is the intensity of voxel k in slice $I(a, L, C)$. The function $\mathbb{I}$ was used to indicate anomalies in the tumour region of a slice, in which case the slice was visually inspected. If more than 50 % of the ROI in a CT stack was considered influenced by artefacts, the patient was removed from the data set.

Artefact Corrected Radiomics Features

Radiomics features were extracted from the PET and CT images subjected to artefact correction, according to the procedures described in Section 3.4.2. Shape features were calculated from the original binary mask, and were therefore not affected by removal of slices.

Assessment of Feature Stability Towards Slice Removal

The distribution of the artefact-corrected features relative to a normal probability distribution [28] was evaluated in terms of the *Shapiro-Wilk* [29] and *D'Agostino's* K^2 [30] tests, described in Sections 2.1.2 and 2.1.3. A *Bonferroni* correction [116] was preformed to adjust for multiple testing effects, as recommended by Parmar et al. (2018).

The *Wilcoxon Signed-Rank* test [27], described in Section 2.1.1, was used to compare the feature distributions before and after removing artefacts.

All tests were performed with a 95 % level of confidence.

3.5 The Feature Matrices

A feature matrix, named the *standard* feature matrix, was constructed by concatenating clinical factors, PET parameters and radiomics features extracted from the image data containing artefacts. The standard feature matrix included 513 features and 198 patients.

By combining radiomics features extracted from artefact filtered images with PET parameters and clinical factors, the *artefact corrected* feature matrix was constructed. This matrix included only 187 patients after removal of image stacks.

Moreover, the *clinical* feature matrix contained only clinical factors for each of the 198 patients.

All categorical features were dummy encoded [45], followed by a Z-score transformation as given by Equation 3.4. Carrying out a Z-scoring of dummy encoded features was motivated by Tibshirani (1997) [118].

3.6 Addressing Intra-Feature Correlations

In study by Hassan et al. (2018) [119], correlations between CT texture features and the number of image intensity values was demonstrated. This association is referred to in this thesis as *intra-feature correlations*. Hassan et al. (2018) proposed modifications of feature definitions to relax the dependency between texture features and image discretisation.

Recall from Section 3.4.3 that a feature group was defined as

$$\mathbf{G} \equiv \{F(\mathbf{I}_{32}), F(\mathbf{I}_{64}), F(\mathbf{I}_{128})\}$$

where $F(\mathbf{I})$ is a feature extracted from an image, $\mathbf{I}$, discretised into 32, 64 and 128 intensity levels. Let $\mathcal{H} : \mathbb{R}^n \mapsto \mathbb{R}^n$ represent the operation of adjusting the definition of a feature according to the modifications proposed by Hassan et al. (2018) [119]. These modifications are given in Table 3.4.

Table 3.4: Adjustments to radiomics texture feature proposed by Hassan et al. (2018). The original feature definition is denoted F, while N_g is the number of image intensity bins. Abbreviations are given in Table 3.1.

Feature Name	**Adjusted Definition**
GLCM DifferenceEntropy	$F/\log N_g^2$
GLCM JointEntropy	$F/\log N_g^2$
GLCM SumEntropy	$F/\log N_g^2$
GLCM Contrast	F/N_g^2
GLCM DifferenceVariance	F/N_g^2
GLCM SumAverage	F/N_g
GLCM DifferenceAverage	F/N_g
GLRLM GrayLevelNonUniformity	$F \cdot N_g$
GLRLM HighGrayLevelRunEmphasis	F/N_g^2
GLRLM ShortRunHighGrayLevelEmphasis	F/N_g^2
NGTDM Contrast	F/N_g
NGTDM Complexity	F/N_g^3
NGTDM Strength	F/N_g^2

Modifying features using the definitions given in Table 3.4 produces a new set of features, $\mathbf{G}'$, where

$$F' = \mathcal{H}(F)$$

is the adjusted version of F included in $\mathbf{G}'$.

Using the standard and artefact corrected feature matrices, defined in Section 3.5, the features given in Table 3.4 were adjusted accordingly. The *Intraclass Correlation Coefficient* (ICC) from Section 5.4.1 was calculated for all groups of features [31]. Then, each feature group was evaluated after

$$\mathbf{G} := \begin{cases} \mu\left(\mathbf{G}\right) & \text{if ICC}\left(\mathbf{G}\right) \geq 0.8 \\ \mu\left(\mathbf{G}'\right) & \text{if ICC}\left(\mathbf{G}'\right) \geq 0.8 \\ \mathbf{G} & \text{otherwise} \end{cases} \tag{3.8}$$

to remove intra-correlated features. In equation 3.8, $\mu\left(\mathbf{G}\right), \mu\left(\mathbf{G}'\right) \in \mathbb{R}^n$ represents the mean across the features in a group. Note that the original features were only retained if the ICC score was strictly less than 0.8. This threshold was adopted from Hassan et al. (2018) [119].

3.7 Searching for Latent Patterns

A *Spectral Co-Clustering*, described in Section 2.3.2, was performed with the standard feature matrix from Section 3.5. The goal of this clustering experiment was to assess the ability of the features to group the patients according to clinical endpoint. All features were Z-scored according to Equation 3.4. One to eight number of biclusters were evaluated to determine the optimal number of clusters in terms of the *Transposed Virtual Error* [41], defined in Section 2.3.3.

3.8 Model Comparison Experiments

Experiments were carried out to estimate the general ability of supervised models to classify disease-free survival in the head and neck cancers cohort. The goal with these experiments was to identify one model with superior performance of predicting patient treatment response. This model would then be used to search for biomarkers.

3.8.1 Measuring Model Performance

The predictive performance of a model was measured based on the area under the *Receiver Operating Characteristics* curve (AUC) [120]. This curve can be used to quantify the performance of a binary classification model, and has been previously used in radiomics [117], [18], [121], [19], [122].

In this thesis, the AUC score was weighted by the proportion of outcomes in each category. This weighting scheme was used to account for imbalanced distributions of patient treatment outcomes. The weighted AUC score, denoted wAUC, was calculated according to

$$\text{wAUC} = \frac{w_0 \text{AUC}(\widehat{\boldsymbol{y}}_{\mathbf{0}}, \boldsymbol{y}_{\mathbf{0}}) + w_1 \text{AUC}(\widehat{\boldsymbol{y}}_{\mathbf{1}}, \boldsymbol{y}_{\mathbf{1}})}{w_0 + w_1} \tag{3.9}$$

where $\widehat{\mathbf{y}}_0$ and $\widehat{\mathbf{y}}_1$ are the predicted treatment outcomes. Moreover, the weights, w_0 and w_1 were calculated according to

$$w_0 = \frac{|c_0|}{|c_0| + |c_1|}$$
$$w_1 = \frac{|c_1|}{|c_0| + |c_1|}$$

using the number of outcomes in each category, $|c|$. The wAUC score ranges from zero to one, where wAUC equal to one indicates that all predictions are correct. Confidence intervals of the wAUC scores were calculated with the bootstrap method described in Section 2.7.4.

To assess whether a model had gained statistical power, the wAUC score for each model was compared to the *no information rate* defined as

$$\nu = \frac{\max\left\{|\boldsymbol{c}_0|, |\boldsymbol{c}_1|\right\}}{n} \tag{3.10}$$

in a cohort of n patients.

3.8.2 Hyperparameter Optimisation

The SMAC framework [88], described in Section 2.7.1, was used to select hyperparameter configurations using the default surrogate model. This random forest surrogate consisted of ten regression trees. Each node in a regression tree was split with a parameter subset of size given by Equation 2.10 for r equal to 5/6. Ten data samples were required to perform a split at each node.

The average model performance was obtained for hyperparameter configuration using the stratified K-fold cross-validation (CV) scheme outlined in Algorithm 3. The SMAC budget for evaluating the objective function, T in Algorithm 1, was set to 80 evaluations. This means that a budget of 80 parameter configurations was used regardless of the number of hyperparameters each model associated.

Algorithm 3 Hyperparameter Optimisation Objective

Input: Learning algorithm, λ, hyperparameter configuration, $\boldsymbol{\phi}$, ground truth, $\boldsymbol{y}$, feature matrix, $\boldsymbol{X}$.

Output: The average loss, $\overline{\pi}$, of cross-validated model performance.

```
procedure ObjectiveCV($\lambda$, $\boldsymbol{\phi}$, $\boldsymbol{y}$, $\boldsymbol{X}$)
    $\boldsymbol{\pi} \leftarrow \emptyset$
    for $k \in [1, K]$ do
        $\boldsymbol{X}_V \leftarrow \boldsymbol{X}_k$, $\boldsymbol{y}_V \leftarrow \boldsymbol{y}_k$
        $\lambda \leftarrow \lambda(\boldsymbol{\phi}, \boldsymbol{X}_V)$                ▷ Train model
        $\boldsymbol{X}_T \leftarrow \boldsymbol{X} \setminus \boldsymbol{X}_V$, $\boldsymbol{y}_T \leftarrow \boldsymbol{y} \setminus \boldsymbol{y}_V$
        $\boldsymbol{\pi}_k \leftarrow$ wAUC$(\boldsymbol{y}_V, \lambda(\boldsymbol{\phi}, \boldsymbol{X}_V))$
    return $\overline{\pi} = \frac{1}{K}\sum_k \boldsymbol{\pi}_k$
```

Model performance estimates, $\{\pi_k\}_{k=1}^{K}$, were obtained for each iteration in Algorithm 3. Each performance estimate was produced by training and evaluating a learning algorithm λ, given a hyperparameter configuration, ϕ, training, $\boldsymbol{X}_T$, and validation, $\boldsymbol{X}_V$, sets. All hyperparameter domains, $\boldsymbol{\Phi}$, were modeled as uniform probability distribution, since this was the only option in SMAC, version 0.10.0 [123].

The general model performance was measured in terms of the wAUC, given by Equation 3.9, averaged over all K validation folds

$$\overline{\pi} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\pi}_k = \frac{1}{K}\sum_{k=1}^{K}\text{wAUC}_k$$

The model maximising $\overline{\pi}$ was selected as optimal.

3.8.3 Classification Experiments

Predictive models were constructed from combinations of seven feature selection and 10 classification algorithms. The feature selection and classification algorithms used in this thesis are described in Sections 2.4 and 2.5. Feature selection, λ_{FS}, and classification, λ_{CLF}, was jointly performed to reduce bias and over-fitting [124]. That is, each model, $\lambda(\phi, \cdot)$, was given as

$$\lambda(\phi, \cdot) = \lambda_{CLF}(\phi_{CLF}, \lambda_{FS}(\phi_{FS}, \cdot))$$

where

$$\phi = \phi_{CLF} \cup \phi_{FS}$$

is the set of hyperparameters for both algorithms.

Due to the computational complexity associated with wrapper methods, this thesis was limited to filter and embedded feature selection methods [43]. Note that embedded feature selection was performed by tree-based and regularised classification models.

To ensure that features contained only positive values when performing χ^2 feature selection, each feature, $\boldsymbol{x}$, were shifted by

$$\boldsymbol{x} := \boldsymbol{x} + \sqrt{(\min \boldsymbol{x})^2} + 1$$

prior to feature selection.

The nested stratified CV scheme in Algorithm 2 from Section 2.7.3 was used in combination with Algorithms 1 and 3, from Sections 2.7.1 and 3.8.2, to estimate the general performance of each candidate model. Algorithm 2 was used to evaluate the hyperparameter configurations obtained with Algorithm 1. The general performance of the model selected according to Algorithm 1 was obtained with Algorithm 2. The main protocol for performing S repeats of a model comparison experiment is given in Algorithm 4.

Algorithm 4 Model Comparison Experiments

Input: Learning algorithm, λ, model selection protocol, SMBO, hyperparameter domain, $\boldsymbol{\Phi}$, ground truth, $\boldsymbol{y}$, feature matrix, $\boldsymbol{X}$.

Output: Training and validation performances, and the optimal hyperparameter configuration of each experimental repeat.

```
procedure ModelComparison($\lambda$, SMBO, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\boldsymbol{X}$)
    $\boldsymbol{\pi} \leftarrow \emptyset$
    for $s \in S$ do
        $\boldsymbol{\pi}_s \leftarrow$ NestedKFoldCV($\lambda$, SMBO, wAUC, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\boldsymbol{X}$)
    return $\boldsymbol{\pi}$
```

Each experiment included 40 repeats of nested CV using different random seeds due to the the stochastic nature of CV and SMAC. A component of the variance in model error estimations has been found to stem from the partitioning of training and validation folds [125]. It has therefore been recommended to repeat random splitting protocols with different split configurations to include information on random variations [126].

Experiments 1-2: Testing the Radiomics Hypothesis

Two classification experiments were performed to assess the prognostic value of combining PET parameters and radiomics features with clinical factors to predict disease-free survival. Previous studies have demonstrated increased ability to predict clinical outcomes by combining clinical factors with radiomics features, as opposed to analysing only clinical factors [18], [17], [22]. Experiment 1 included only clinical factors, while Experiment 2 was performed with the standard feature matrix, described in Section 3.5. Five folds were used in the nested CV scheme.

Experiment 3: Removing Image Artefacts

Having removed image artefacts as described in Section 3.4.4, the goal of Experiment 3 was to study the effect of artefact correction on feature selection and model performances. A classification experiment was performed with the artefact corrected feature matrix from Section 3.5. The nested CV scheme was configured with five folds.

Experiment 4: Removing Intra-Correlated Features

Subjecting the standard feature matrix, defined in Section 3.5, to a filtering and removal of intra-correlated, described in Section 3.6, gave a subset of features. A classification experiment was performed including these features modified by Equation 3.8. Information leakage was avoided since ICC thresholding is an unsupervised operation [93]. Moreover, note that ICC thresholding in was applied to all features dependent on the number of image intensity bins, and not just the features shown in Table 3.4. However, only the features in Table 3.4 were modified. A z-scoring was performed of the resulting feature matrix, given by Equation 3.4, and five folds were used in the nested CV scheme.

Experiment 5: Removing Intra- and Inter-Correlated Features

Using the feature matrix obtained from removing intra-feature correlations, derived in the previous section, the *Spearman's Rank Correlation* (SCC) coefficient was calculated for the remaining features [34]. For each pair of features that were correlated by at least 0.95 SCC, one of the features in the correlated pair was arbitrarily removed. The aim of the experiment was to evaluate the effect of removing both intra- and inter-feature correlations on model performances. The threshold of 0.95 SCC was arbitrarily selected. Model comparisons experiments were configured with 5-fold nested stratified CV.

Preliminary Feature Relevance

The model corresponding to the highest wAUC score in the classification experiment was used to rank features according to relevancy for predicting disease-free survival. Information from all patients was used to retrain the model and infer feature importance. The model was configured with the average of the hyperparameter configurations selected in the experiment. Features were initially ranked using the feature selection algorithm, and a subset was selected, including the average number of features selected in the classification experiment. Furthermore, the subset of features was ranked by using *Shapley Additive Explanations* (SHAP) values [81], described in Section 2.6.1, and the selected classification model.

Experiments 6-9: HPV Subgroup Analyses

The patient cohort was divided into two subgroups referred to as (1) HPV related and (2) HPV unrelated. This partitioning of patients was motivated by studies reporting an association between HPV status and clinical outcomes [18], [127], [128], [129], [130]. Furthermore, Moan et al. (2019) suggested that the relation between disease-free survival, PET parameters and tumor volume was stronger for the HPV unrelated patients compared to the HPV related patients in this cohort [100].

The HPV related group included patients with positive HPV status and tumour located in the *oropharynx*, while the HPV unrelated group consisted of patients with either negative HPV status and a tumour in the oropharynx, or a primary tumour outside of oropharynx regardless of HPV status. A total of 149 patients were eligible for analysis since HPV status could not be obtained for 49 patients. Table 3.5 summarises the patient characteristics of the HPV related and unrelated cohorts.

Table 3.5: Patient characteristics of the HPV related and unrelated cohorts.

Clinical Factor	HPV Related	HPV Unrelated
Total number of patients	82	67
PFS (%)	73	53
Age (years)	60, (40, 80)$^{(1)}$	62, (43, 77)$^{(1)}$
Tumour volume (cm^3)	13.4, (0.826, 145)$^{(1)}$	

(1): median, (min, max)

A total of four experiments were performed including the two patient sub-cohorts and features from the standard feature matrix. For each patient subgroup, Experiments 6 and 8 included the standard feature matrix. Moreover, the features used in Experiments 8 and 9 were subjected to removal of intra- and inter-feature correlations, described in Sections 3.8.3 and 3.8.3. The number of folds in the CV procedure was increased from five to 10 due to the reduced number of patients in each analysis compared to previous experiments.

Biomarker Identification

The model corresponding to the highest wAUC score was used to rank features according to relevancy for predicting disease-free survival. Feature importance to treatment response was quantified with SHAP values, described in Section 2.6.1, and the selected classification model. The model was configured with the average of the hyperparameter configurations selected in the experiment. Features were initially ranked using the feature selection algorithm, and a subset of 26 features was selected. Furthermore, this feature subset was ranked by using SHAP values [81], described in Section 2.6.1, and the selected classification model. Only the patients from the HPV unrelated cohort was used.

A learning curve was constructed to investigate the 10-fold stratified CV training and validation performance of the selected model for different training set sizes [45]. The learning curve was used to evaluate the number of CV folds, as well as a selected hyperparameter configuration.

Experiment 10: Reassessment of the Radiomics Hypothesis

The classification experiment was performed to assess the prognostic value of the clinical factors for disease-free survival in the HPV unrelated cohort. The experiment included the clinical factors from the clinical feature matrix described in Section 3.5, and 5-fold nested stratified CV.

Chapter 4

Results

For clarity, the abbreviations of classification models discussed in this chapter are given in Table 4.1.

Table 4.1: Abbreviations of classification algorithms.

Abbreviations			
KNN	K-Nearest Neighbours	LR	Logistic Regression
LGBM	Light Gradient Boosting Machine	RF	Random Forest
ET	Extremely Randomised Trees	QDA	Quadratic Discriminant Analysis
DT	Decision Tree	Ridge	Ridge Classifier
SVC	C-Support Vector Classifier	XGB	Extreme Gradient Boosting

Furthermore, abbreviations of radiomics texture feature categories are given in Table 4.2.

Table 4.2: Abbreviations of radiomics texture feature categories.

Abbreviations			
GLCM	Gray Level Co-occurrence Matrix	GLSZM	Gray Level Size Zone Matrix
NGTDM	Neighbouring Gray Tone Difference Matrix	GLRLM	Gray Level Run Length Matrix
GLDM	Gray Level Dependence Matrix		

4.1 Data Set Exploration

The data set included contrast enhanced pre-treatment *18F-fluorodeoxyglucose Positron Emission Tomography/Computed Tomography* (PET/CT) images of 198 head and neck cancer patients. Moreover, the data set also included PET parameters and clinical factors for each patient, defined in Section 3.3.

4.1.1 Exploring the Image Data

The PET and CT images were explored by examining the distribution of intensities. Figure 4.1 shows the maximum, mean, median and minimum intensities in each of the PET and CT image stacks.

Figure 4.1 illustrates more variation in the maximum intensity of the PET and CT stacks compares to the minimum, mean and median intensities. Moreover, the maximum CT intensities in Figure 4.1 b) appears to be divided into two groups. Compared to CT, the distribution of maximum PET intensities in Figure 4.1 a) is more randomly distributed.

Assuming two clusters, the K-means++ algorithm, described in Section 2.3.1, was applied to the Euclidean distances between the maximum CT intensities in Figure 4.1 b). The algorithm found that the intensities could be divided into two groups by a horisontal line from intensity value 3279. Distortions from the clustering, described in Section 2.3.1, for the evaluated number of clusters are shown in Figure 4.2.

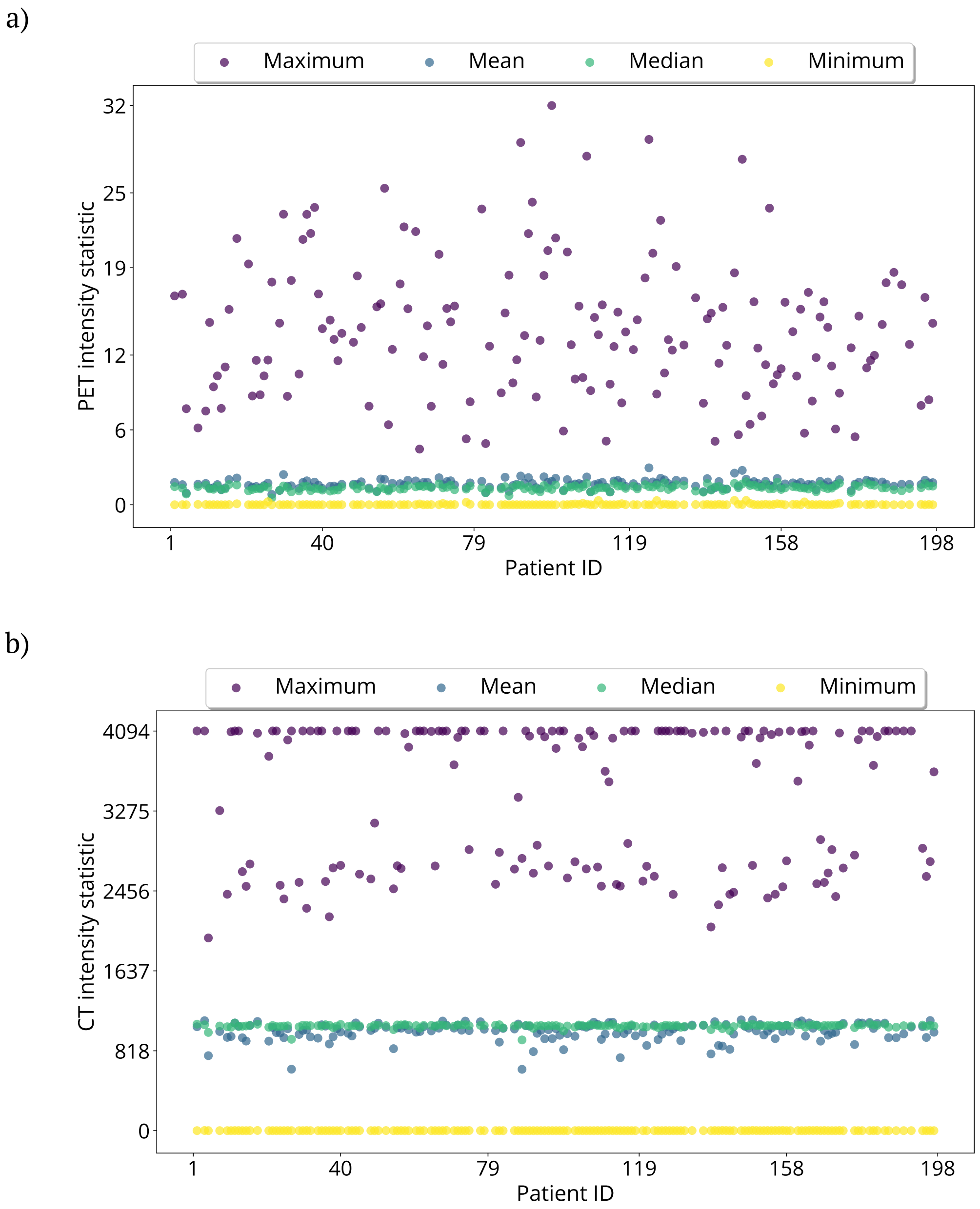


Figure 4.1: The maximum, mean, median and minimum statistics calculated from the intensities of a) PET stacks and b) CT stacks for each patient.

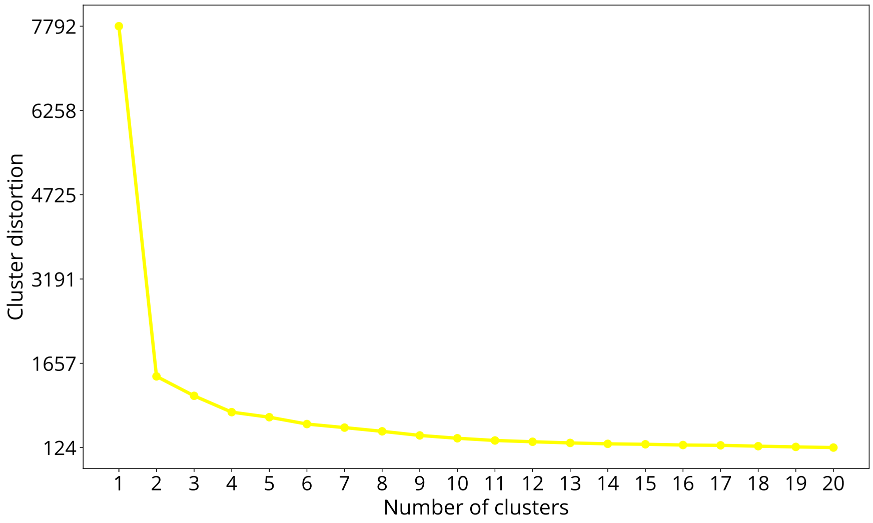


Figure 4.2: Cluster distortions obtained by K-means++ clustering, for one to 20 target clusters of the CT maximum intensities. Smaller distortion indicates a higher quality of clusters.

The largest reduction of cluster distortion in Figure 4.2 occurred for two clusters, which indicates the presence of two clusters.

Figure 4.3 shows the maximum, mean, median and minimum intensities calculated from only the tumour volume (ROI) of the PET and CT image stacks.

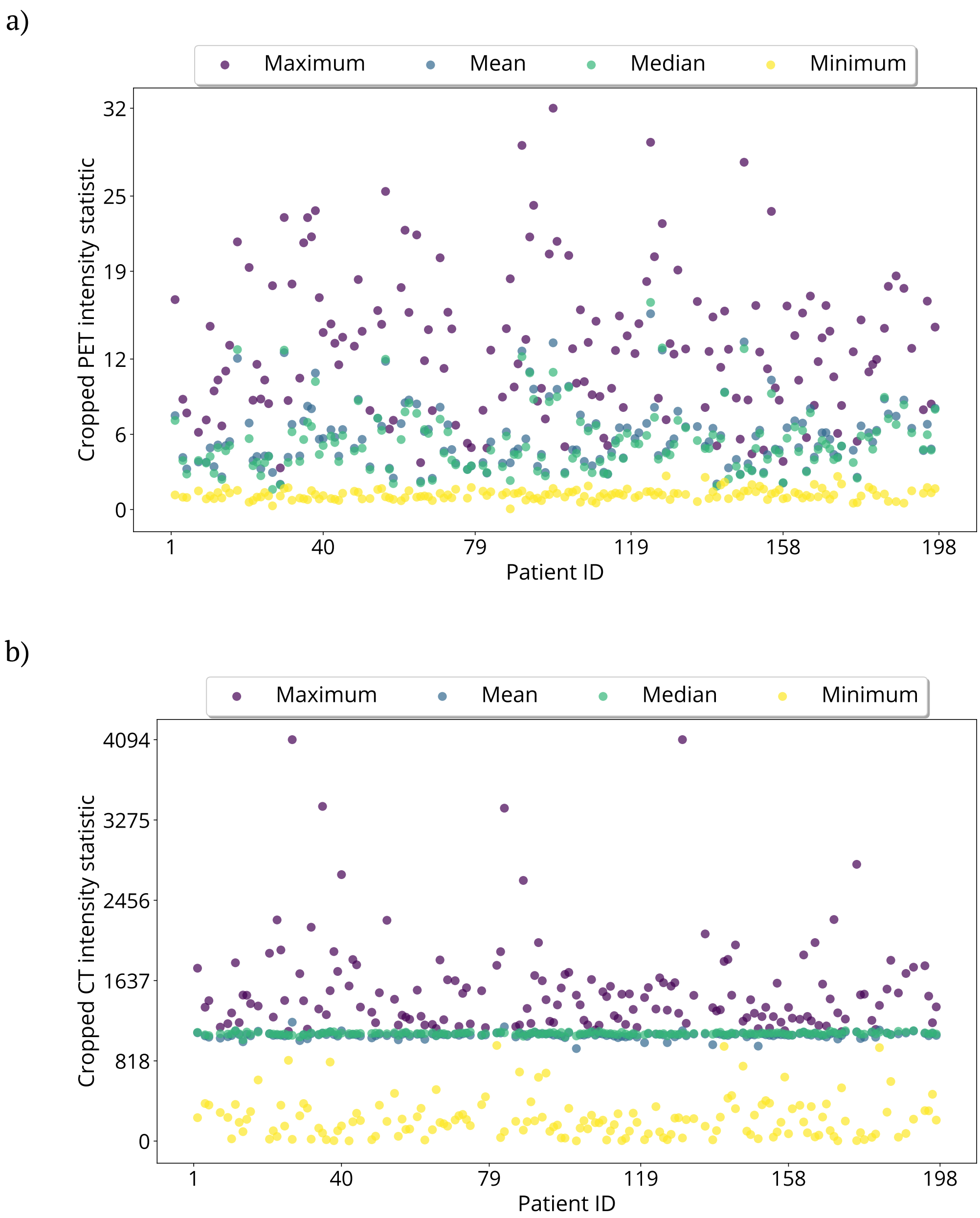


Figure 4.3: The maximum, mean, median and minimum statistics calculated from the tumour volume of a) PET stacks and b) CT stacks for each patient.

Apart from the minimum intensity, Figure 4.3 a) shows more variation in the PET intensity statistics across patients compared to the CT intensities in Figure 4.3 b).

Note the stability in the CT median intensity, used in Section 3.4.4, for image artefact correction, compared to the mean. Figure 4.3 b) shows four CT stacks with intensities exceeding 3279, which was the threshold determined with K-means++ to separate the maximum intensities in Figure 4.1 b). However, visual inspection of a random selection of CT images from each of these clusters did not reveal any particular differences.

4.1.2 The Standard Feature Matrix

The distributions of the radiomics features in the standard feature matrix from Section 3.5, were explored with scatter plots given in Appendix A, Figure A.1. Despite some extreme feature values, the shape, first-order and texture features appeared to be relatively randomly distributed across patients. Extreme observations among the shape features, shown in Figure A.1 a), were found to stem from the tumour volume.

4.1.3 Spectral Co-Clustering of Features

Spectral Co-clustering was performed in an unsupervised approach to discriminate between clinical endpoints using the standard feature matrix. Recall from Section 2.3.2 that this algorithm simultaneously clusters rows and columns by grouping together observations with similar values.

From one to eight clusters were evaluated in terms of the *Transposed Virtual Error*, defined in Section 2.3.3, to quantify the cluster quality. Three clusters were found to be optimal by giving the lowest Transposed Virtual Error, as shown in Figure 4.4.

Figure 4.4 shows that the Transposed Virtual Error increased after attempting more than three clusters. However, using one, two or three clusters gave similar results, which indicates that the algorithm did not recognise differences in clinical outcomes based on the standard features.

A re-arrangement of the feature matrix, in Figure 4.5, illustrates the three detected clusters inside bounding boxes along the main diagonal.

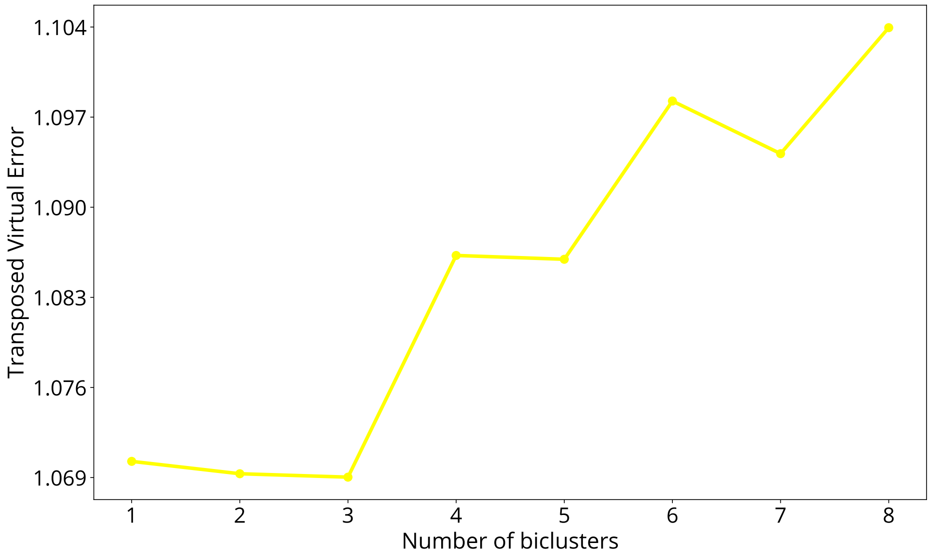


Figure 4.4: The *Transposed Virtual Error* scores (vertical axis) from Spectral Co-clustering of the standard feature matrix using one to eight target clusters (horisontal axis). A score indicates higher cluster quality.

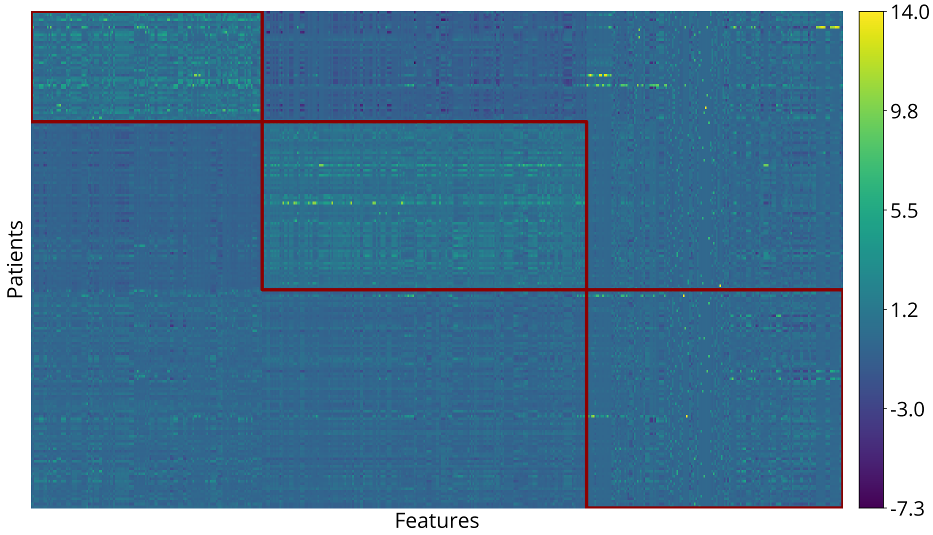


Figure 4.5: Spectral Co-clustering of the standard feature matrix with 513 features (columns) and 198 patients (rows). Red bounding boxes enclose each detected cluster. The colour bar indicates the magnitude of each feature value.

Figure 4.5 illustrates the shape of each of the detected cluster, representing the size of the feature and patient subsets. The top left-most cluster consisted of 44 patients and 146 features, the middle cluster held 67 patients and 205 features, while the lower right-most cluster included 162 features and the remaining 87 patients.

The distribution of features by category in each of the clusters shown in Figure 4.5 is given in Figure 4.6. Note that the upper left-most cluster in Figure 4.5 is counted as cluster number one, while the lower right-most cluster is referred to as cluster number three.

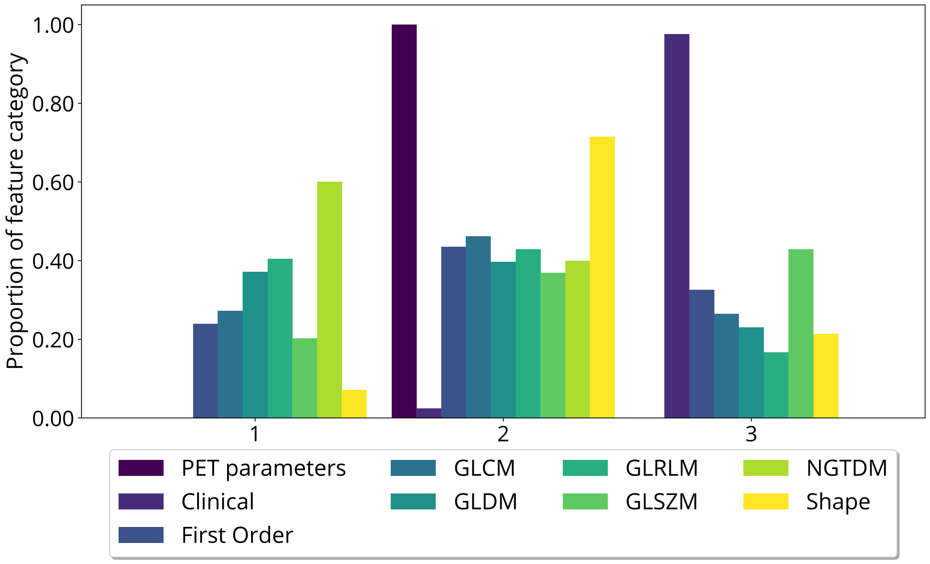


Figure 4.6: The distribution of features in each detected cluster grouped by feature category. Clusters are numbered from left to right along the horisontal axis of Figure 4.5.

Figure 4.6 shows that cluster 2, the middle cluster in Figure 4.5, was assigned all of the PET parameters and that cluster 3 contained the majority of clinical factors. Cluster 2 also included the most shape features compared to cluster 1 and 3. Thus, Figure 4.6 indicates that PET parameters, shape features and clinical factors were distinguishable by the Spectral Co-clustering algorithm.

The distribution of clinical outcomes in the detected clusters are shown in Figure 4.7 where cluster indicators corresponds to Figure 4.6. Combining the information in

Figure 4.6 and Figure 4.7 illustrates the distribution of feature categories and clinical endpoints in each cluster shown in Figure 4.5.

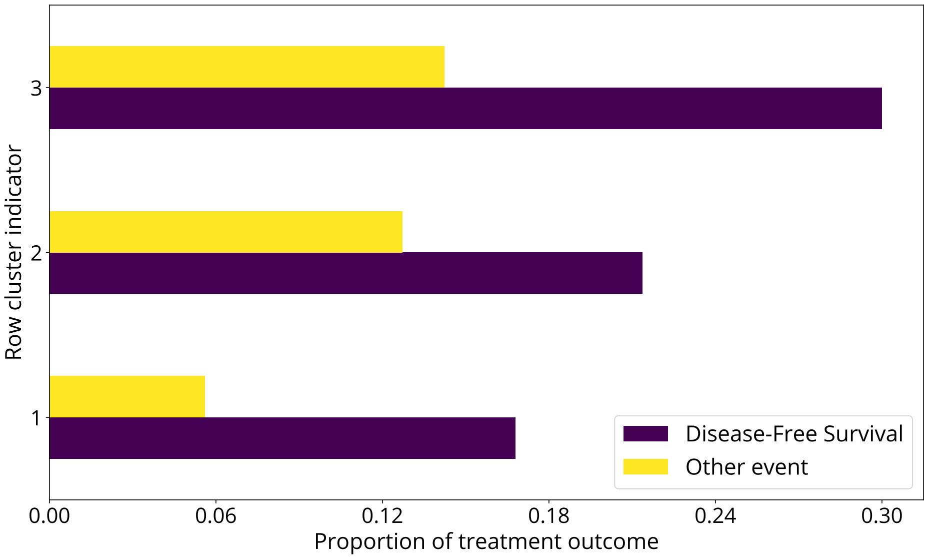


Figure 4.7: The distribution of clinical endpoints in detected clusters. Clusters are numbered from top to bottom along the vertical axis of Figure 4.5.

In Figure 4.7, each detected cluster contains both categories of treatment outcomes. Cluster 1 and 3 included 62 % and 67 % cases of *disease-free survival*, while this outcome constituted 75 % of cluster 2. Thus, the Spectral Co-clustering algorithm did not discriminate between clinical outcomes by using information from the standard feature matrix.

4.2 Learning to Predict Disease-Free Survival

A total of 10 model comparison experiments, described in Section 3.8.3, were performed to identify the model with the superior ability to classify disease-free survival as response to radiotherapy. The average run time for a classification experiment was approximately 42 hours. A weighted area under the *Receiver Operating Characteristic curve* (wAUC), outlined in Section 3.8.1, was used to measure the performance of each model.

4.2.1 Validation of the Radiomics Hypothesis

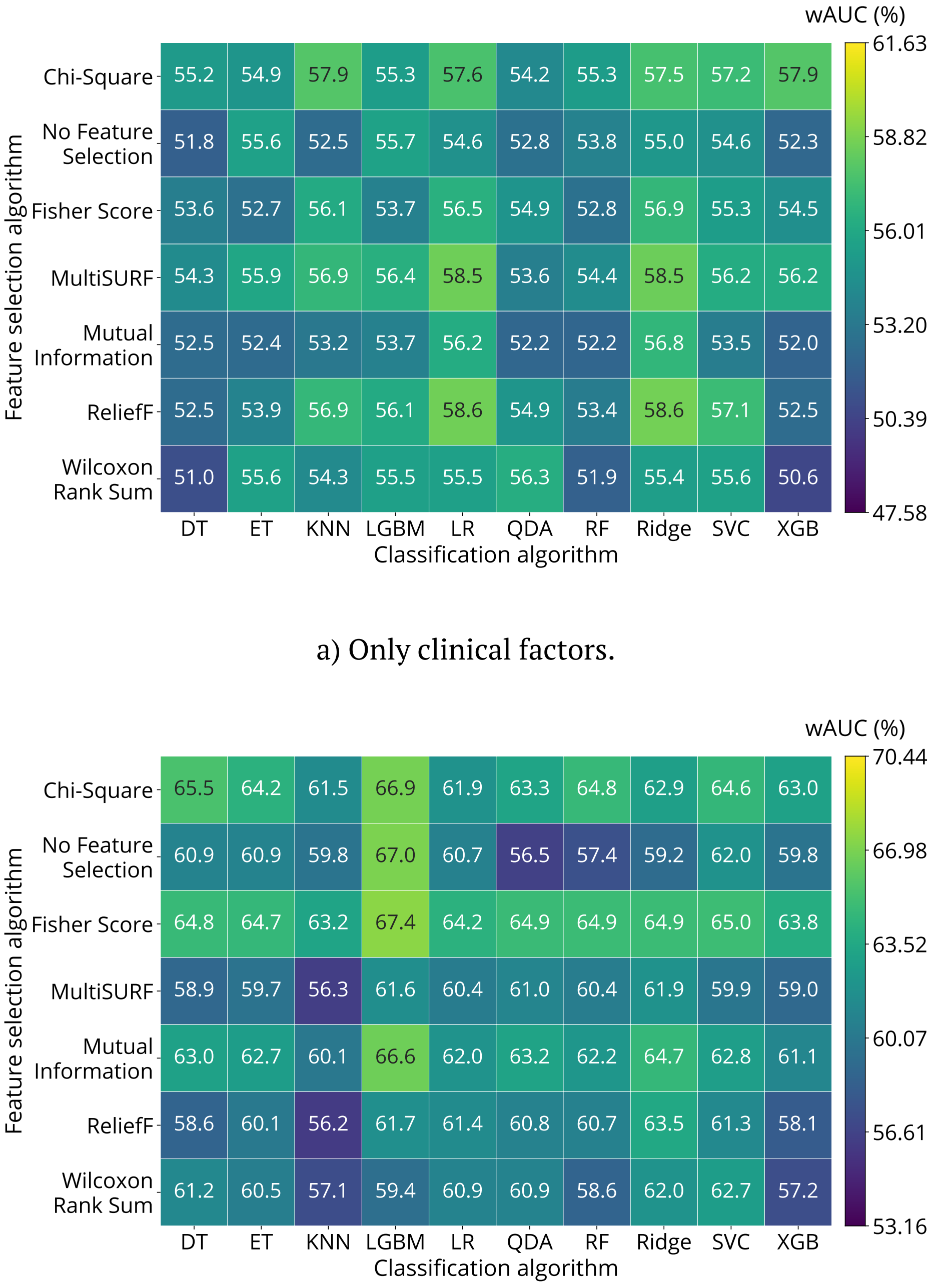
Experiments 1 and 2, described in Section 3.8.3, were performed to evaluate the effect of combining clinical factors with PET parameters and radiomics features on model performances.

The results of classifying treatment outcomes using the clinical and standard feature matrices, defined in Section 3.5, are shown in Figure 4.8. The *no information rate*, introduced in Section 3.8.1, was 67 % disease-free survival in Experiments 1 and 2. Abbreviations to classification algorithms are given in Table 4.1.

Figure 4.8 a) shows that models combining *ReliefF* or *MultiSURF* feature selection with either *Logistic Regression* (LR) or *Ridge Classification* gave the highest scores of approximately 59 % wAUC when only clinical factors were used. When PET parameters and radiomics features were also included, the combination of *Fisher Score Light Gradient Boosting Machine* (LGBM) model gave the highest wAUC score of about 67 % (Figure 4.8 b). Moreover, LGBM showed superior performance without prior feature selection, as well as in combination with *Chi Square* or *Mutual Information*. The standard deviation of the wAUC scores for each model in Figure 4.8 ranged between 4 % and 8 % wAUC, which indicates relatively stable models.

4.2.2 An Attempt to Handle Image Artefacts

A total of 11 CT image stacks were found to contain bone structure and streak artefacts, described in Section 3.4.4, in at least 50 % of the ROI. These PET and CT stacks were removed from the data set, reducing the cohort from 198 to 187 patients. Only slices were removed from the remaining CT stacks that were less influenced by such artefacts, while all slices were retained in the PET images. Figure 4.9 shows the percentage of the ROI removed from each CT stack.



a) Only clinical factors.

b) Clinical factors, PET parameters and radiomics features.

Figure 4.8: Average wAUC (%) from including a) only clinical factors, and b) the standard feature matrix to classify disease-free survival with combinations of feature selection (vertical axis) and classification (horisontal axis) algorithms. The colour bar shows that a higher score corresponds to more correct classifications.

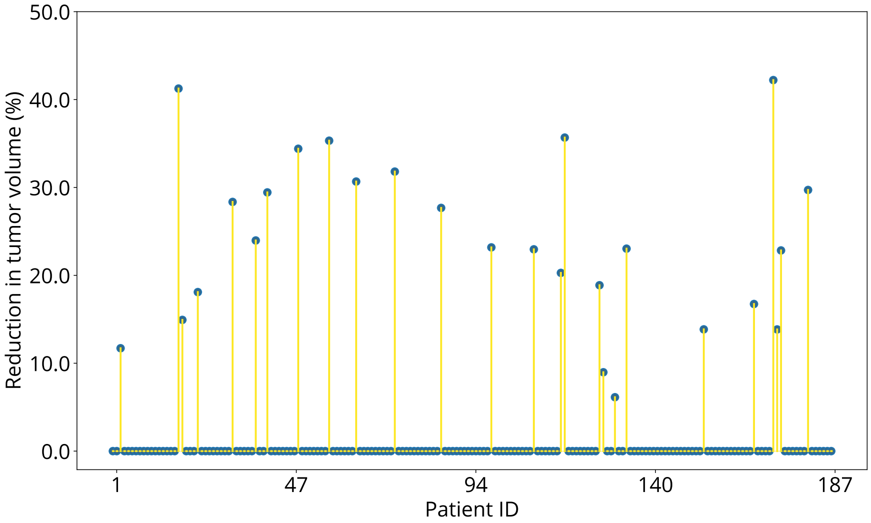


Figure 4.9: The relative proportion of the ROI (vertical axis) in CT stacks (horisontal axis) removed due bone and streak artefacts.

The largest reduction in the ROI shown in Figure 4.9 amounted to 42 %, while 24 % was removed on average from the 26 CT images with identified artefacts.

Changes in Radiomics Features After Artefact Removal

The *Wilcoxon Signed-Rank* test (WSR) was used with a 95 % confidence level to compare the distributions of PET and CT features before and after removal of image artefacts, as described in Section 3.4.4. Table 4.3 summarises the WSR test outcomes. Recall that the original image masks were used to calculate shape features, and that artefact correction only affected first-order and texture features.

The relative proportion of PET and CT features that were considered insignificantly changed by removal of image artefacts according to the *Wilcoxon Signed-Rank* test with a 95 % level of confidence.

Table 4.3: The relative proportion of PET and CT features that were not significantly affected by removal of image artefacts according to the *Wilcoxon Signed-Rank* test using a 95 % confidence level.

	Features	
Imaging Modality	**First-Order (%)**	**Texture (%)**
PET	94	70
CT	84	76

Table 4.3 shows that texture features were more influenced by the removal of image artefacts compared to first-order features. Still, at least 70 % of the texture features were likely to originate from the same distribution before and after slice removal according to the WSR test.

Classifying Disease-Free Survival using Artefact Corrected Features

Results from classifying treatment outcomes using the *artefact corrected* feature matrix, described in Section 3.8.3, are shown in Figure 4.10. The artefact corrected feature matrix, defined in Section 3.5, included radiomics features extracted from artefact corrected images, PET parameters and clinical factors. The no information rate in this cohort of 187 patients was 68 % disease-free survival. Abbreviations to classification algorithms are given in Table 4.1.

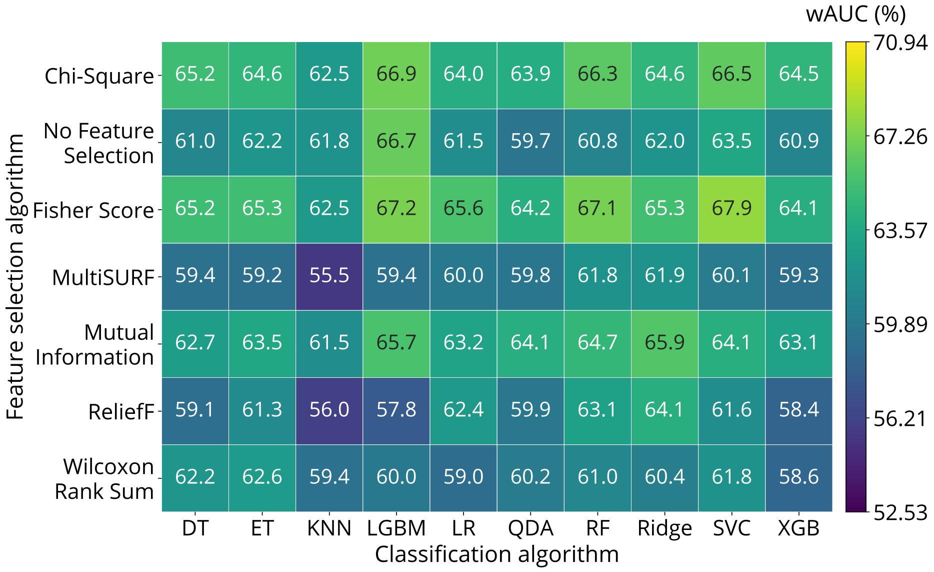
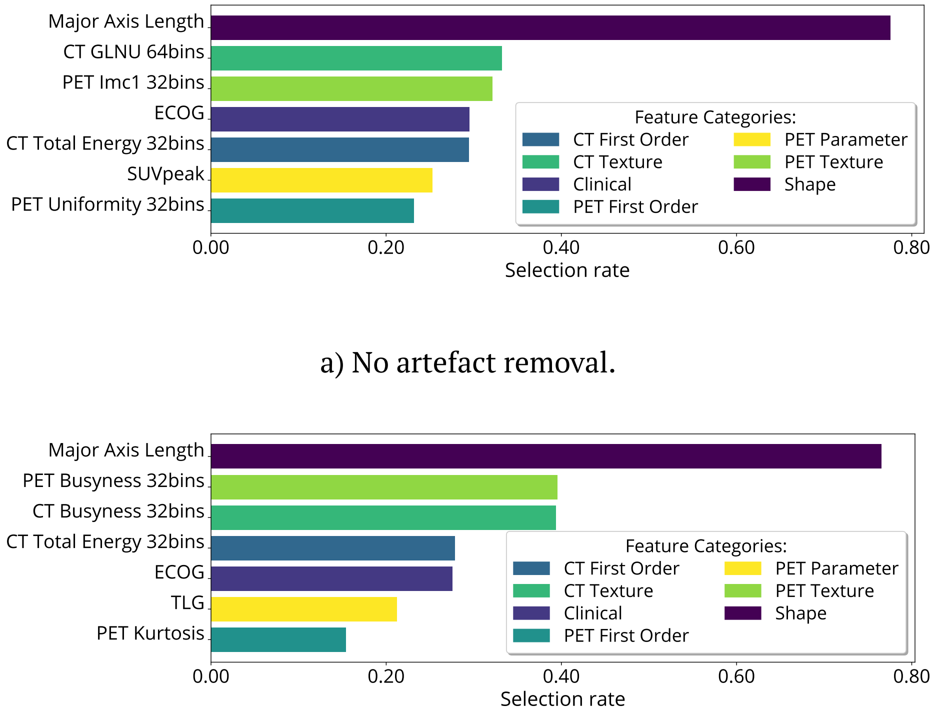


Figure 4.10: Average wAUC (%) from including the artefact corrected feature matrix to classify disease-free survival with combinations of feature selection (vertical axis) and classification (horisontal axis) algorithms. The colour bar shows that a higher score corresponds to more correct classifications.

Almost 68 % wAUC was achieved by combining Fisher Score feature selection with *Support Vector Classification* (SVC), which is the highest score shown in Figure 4.10. Fisher Score combined with Logistic Regression (LR) or Ridge Classification gave about 65 % wAUC. The highest performance obtained with LGBM, also in combination with Fisher Score, was close to 67 % wAUC. The standard deviations of the wAUC scores in Figure 4.10 ranged from 4 % to 8 % wAUC as indications of model stability.

Feature Selection Before and After Artefact Removal

Figure 4.11 shows which features were the most often selected in each category before and after removal of artefacts. The *selection rate* represents the number of times a feature was selected relative to the total number of feature selection operations. Thus, a feature selected on each occasion during a classification experiment receives a selection rate equal to one. Recall that feature definitions are available in Appendix E.



a) No artefact removal.

b) Artefact removal.

Figure 4.11: The most selected features (vertical axis) in classification experiments with a) the standard feature matrix, and b) the artefact corrected feature matrix. A higher selection rate (horisontal axis) signifies increased feature selection. Abbreviations: *Total Lesion Glycolysis* (TLG), *Gray Level Non-Uniformity* (GLNU), *Eastern Cooperative Oncology Group* (ECOG), *Informational Measure of Correlation 1* (Imc1).

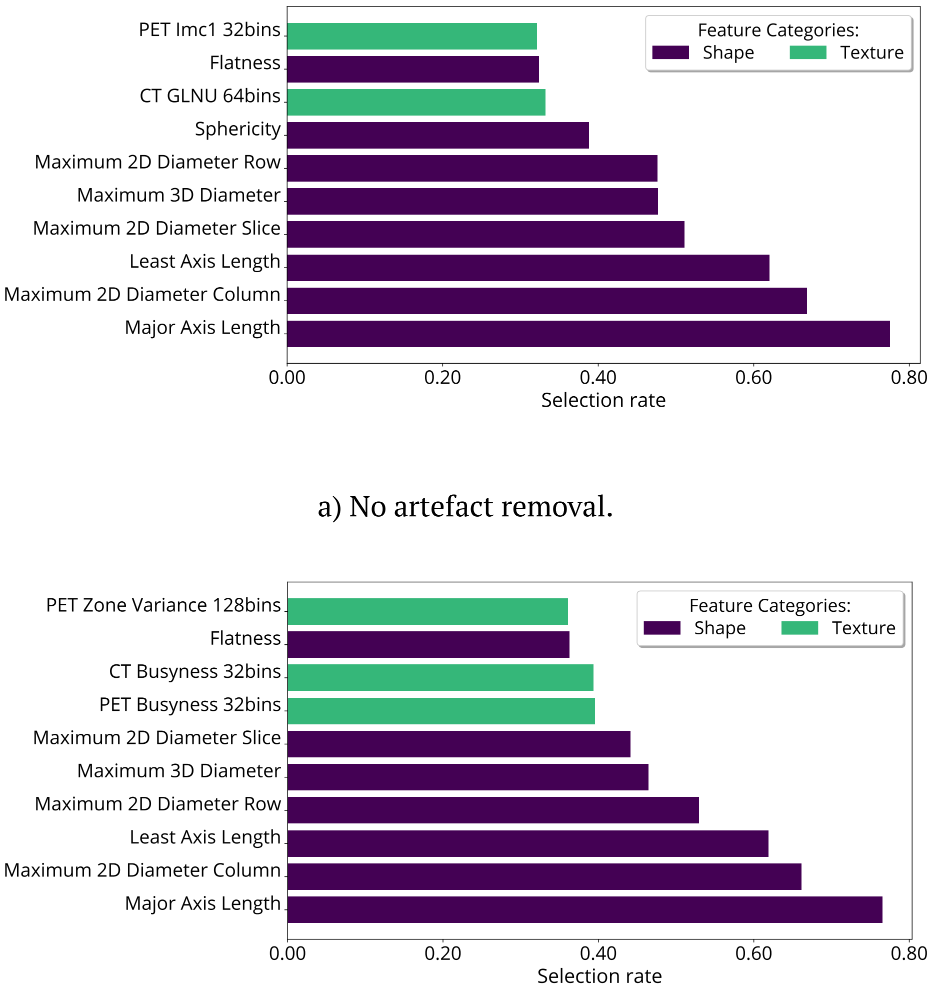
Although not being directly affected by slice removal, Figure 4.11 shows that *Major Axis Length* was the most selected feature before and after artefact correction. Major Axis Length, defined in Appendix E, Section E.2, describes the longest axis in the shape of the ROI. Removing artefacts replaced PET *Informational Measure of Correlation 1* (Imc1) CT *Gray Level Non-Uniformity* (GLNU) with PET and CT *Busyness* as the most selected texture features. Busyness, Lmc1 and GLNU features quantify characteristics of intratumor heterogeneity [16].

The 10 most selected features before and after the removal of artefacts, regardless of category, is shown in Figure 4.12. Abbreviations to texture feature categories are given in Table 4.2.

Figure 4.12 conveys that shape was the dominant category in classification experiments. Furthermore, selection of PET *Zone Variance*, describing the variance over regions of voxels with the same intensity, increased after artefact correction. The *Flatness* shape feature describes the ratio between the longest and shortest axis of the ROI shape and appeared to be unaffected by changes to the image data. The average selection rate for the 503 features included in the standard and artefact corrected feature matrices, but not in Figure 4.12, was approximately 0.35.

The *Spearman's Rank Correlation Coefficient* (SCC) [34], described in Section 2.2.2, between the features in Figure 4.12 and the ROI size is given in Figure 4.13. That is, Figure 4.12 illustrates the association between the 10 most selected features and the size of the tumour volume.

Apart from *Flatness* and *Sphericity*, all shape features were correlated by at least 0.8 SCC with ROI, according to Figure 4.13. Sphericity measures the roundness of the ROI relative to a circle. Notice the different degrees of correlation between the ROI, and PET and CT Busyness.



a) No artefact removal.

b) Artefact removal.

Figure 4.12: The 10 most selected features (vertical axis) in classification experiments with a) the standard feature matrix, and b) the artefact corrected feature matrix. A higher selection rate (horisontal axis) signifies increased feature selection. Abbreviations: *Gray Level Non-Uniformity* (GLNU), *Informational Measure of Correlation 1* (Imc1).

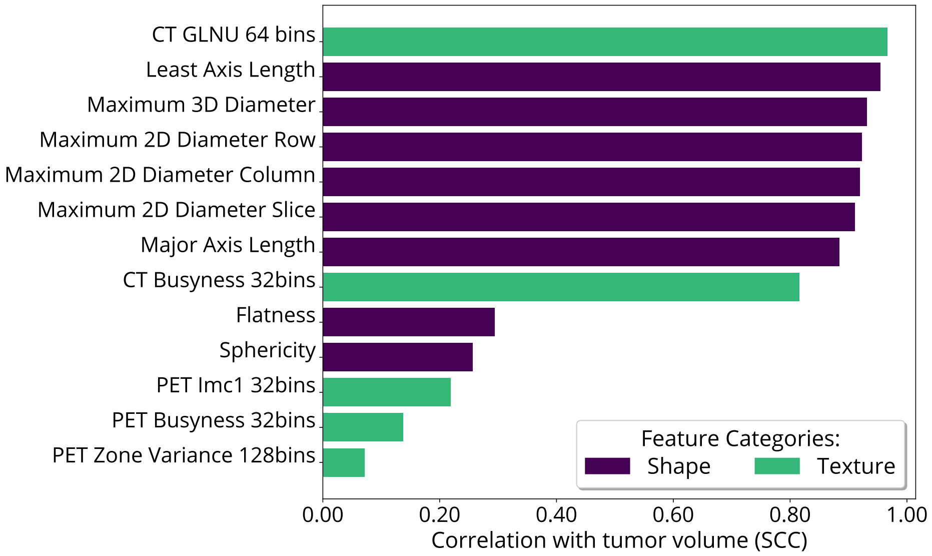


Figure 4.13: The *Spearman's Rank Correlation Coefficient* (SCC) between the 10 most selected features in classification experiments with the standard and artefact corrected feature matrices and the ROI size. Abbreviations: *Gray Level Non-Uniformity* (GLNU), *Informational Measure of Correlation 1* (Imc1).

4.2.3 Investigating Feature Redundancy

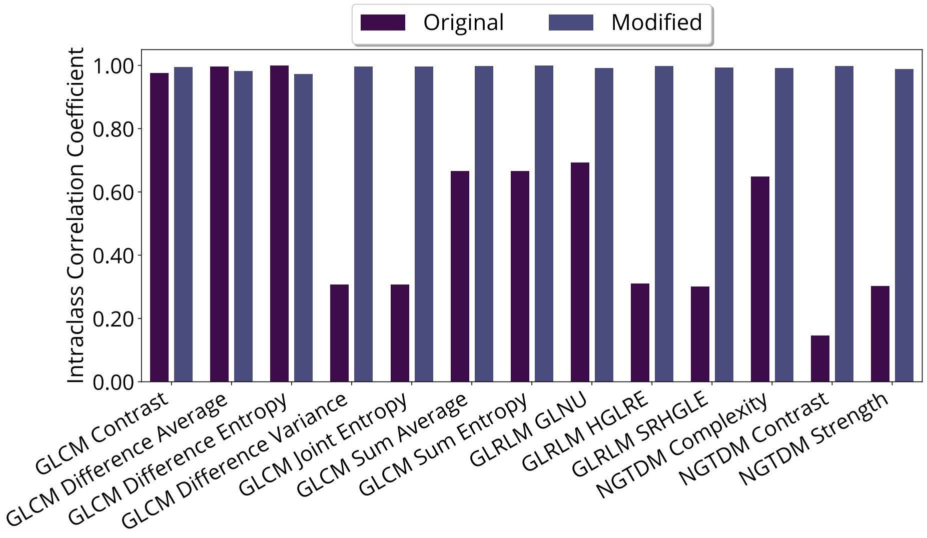
Classification experiments, described in Section 3.8.3, were performed to investigate *redundancy* among clinical factors, PET parameters and radiomics features. Redundancy is referred to in this thesis as repeated feature information, measured in terms of *intra*- and *inter-feature* correlations.

Adjustments to Reduce Intra-Feature Correlations

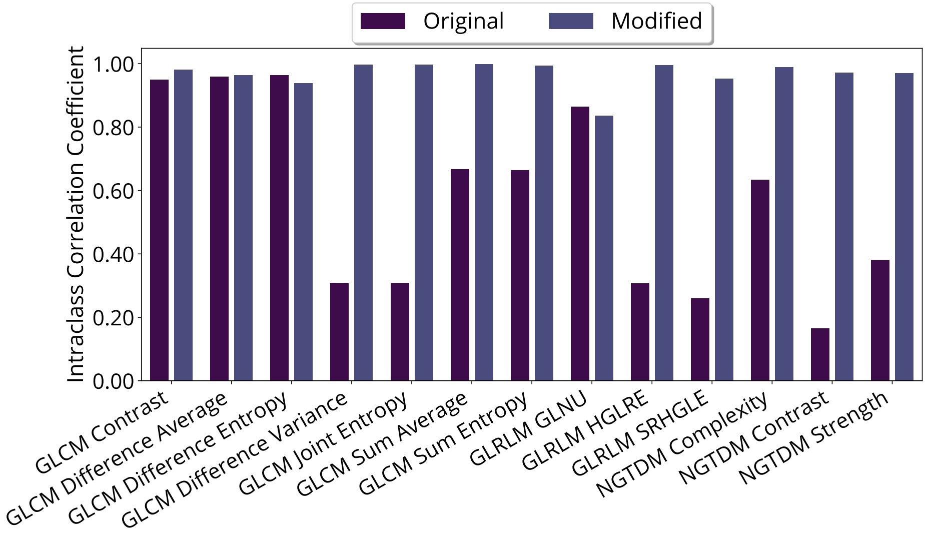
Recall from Section 3.8.3 that radiomics features were modified after Hassan et al. (2018) [119] to reduce their dependency on the number of image intensity bins. This dependency, referred to as intra-feature correlation, was measured using the *Intra-class Correlation Coefficient* (ICC) [31]. The ICC was calculated for groups of features extracted from differently discretised images using the same feature definition. This procedure was applied to features from the standard and artifact corrected feature matrices. A higher ICC score after feature modification implies that the correlation between the image discretisation level and the feature was reduced. Abbreviations to texture feature categories are given in Table 4.2.

Figure 4.14 shows that adjustments to features, given in Table 3.1, increased the ICC for all PET and CT texture features. Thus, by modifying these features, information on image discretisation was incorporated to relax the association between features and the level of discretisation.

The ICC scores of features extracted from artefact-filtered images are shown in Figure 4.15. In Figure 4.15, the ICC score exceeds 0.8 for all features prior to modifications. Thereby, these features appeared to have become invariant to image discretisation after artefact correction. Moreover, adjustments to these features decreased the ICC, meaning that the correlation between the features and the number of image intensity bins increased.

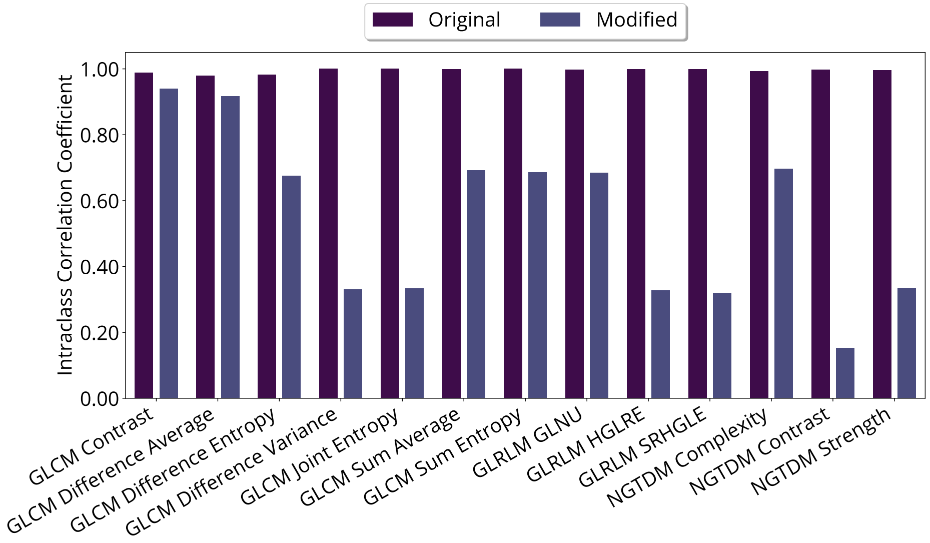


a) PET texture features.

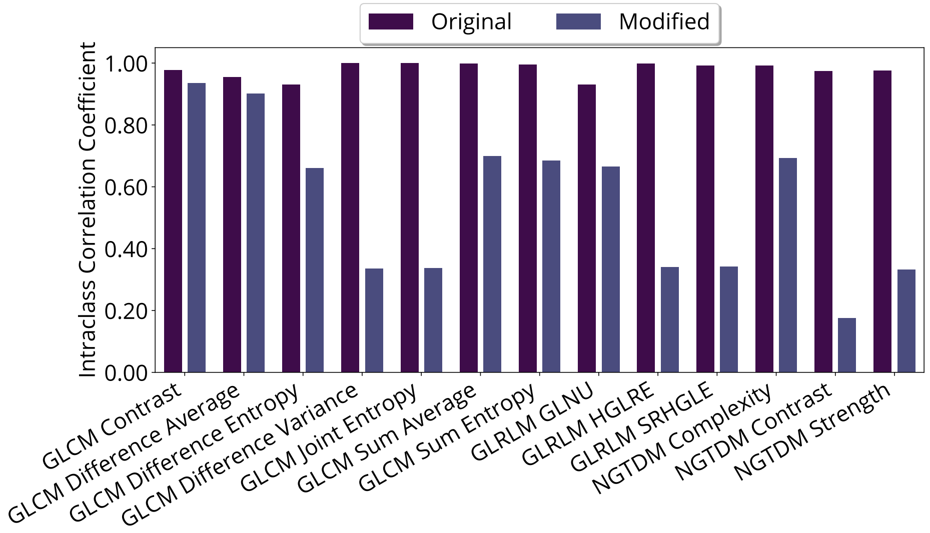


b) CT texture features.

Figure 4.14: The *Intraclass Correlation Coefficient* (vertical axis) of a) PET and b) CT texture features (horisontal axis) extracted from the original images discretised at 32, 64 and 128 bins. The *Original* and *Modified* labels refers to the original and adjusted feature definitions, to account for image discretisation levels.



a) PET texture features.



b) CT texture features.

Figure 4.15: The *Intraclass Correlation Coefficient* (vertical axis) of a) PET and b) CT texture features (horisontal axis) extracted from the artifact corrected images discretised at 32, 64 and 128 bins. The *Original* and *Modified* labels refers to the original and adjusted feature definitions, to account for image discretisation levels. [119] and described in Section 3.8.3.

Replacing the features in a group corresponding of least 0.8 ICC by their average, as outlined in Section 3.6, reduced the number of features in the standard feature matrix from 513 to 341. Results from classifying disease-free survival using this subset of 341 features, described in Section 3.8.3, are given in Appendix B.1, Figure B.1. Omitting feature selection with the Light Gradient Boosting Machine classifier gave the highest wAUC score close to 67 %, while Fisher Score feature selection and the Support Vector Classifier (SVC) gave approximately 65 % wAUC. The lowest score of about 55 % wAUC was obtained with *Wilcoxon Rank Sum* feature selection and Logistic Regression.

Removal of Intra- and Inter-Correlated Features

Further reduction of feature redundancy was carried out based on the subset of 341 features retained after removal of intra-feature correlations, described in the previous section. The SCC was used to quantify inter-feature correlations. In a pair of features correlated by at least 0.95 SCC was one of the features arbitrarily selected and removed. Performing this operation removed 188 features to produce a feature matrix of 152 features.

Figure 4.16 shows a) the SCC calculated between the 513 original features, and b) the SCC for the 152 features retained after removal of intra- and inter-feature correlations.

Originally, 115 features in the standard feature matrix were correlated by at least 0.95 SCC, as shown in Figure 4.16 a). Clinical factors, described in Section 3.3.2, were correlated by less than 0.25 SCC, while PET parameters, defined in Section 3.3.3, were removed during SCC thresholding, which is given in Figure 4.16 b).

The relationship with the features correlated to the PET parameters are illustrated in Figure 4.17.

Figure 4.17 illustrates the relationship between the features determined to be the most correlated to PET parameters. In addition, Figure 4.17 b) and c) shows observations deviating from the majority.

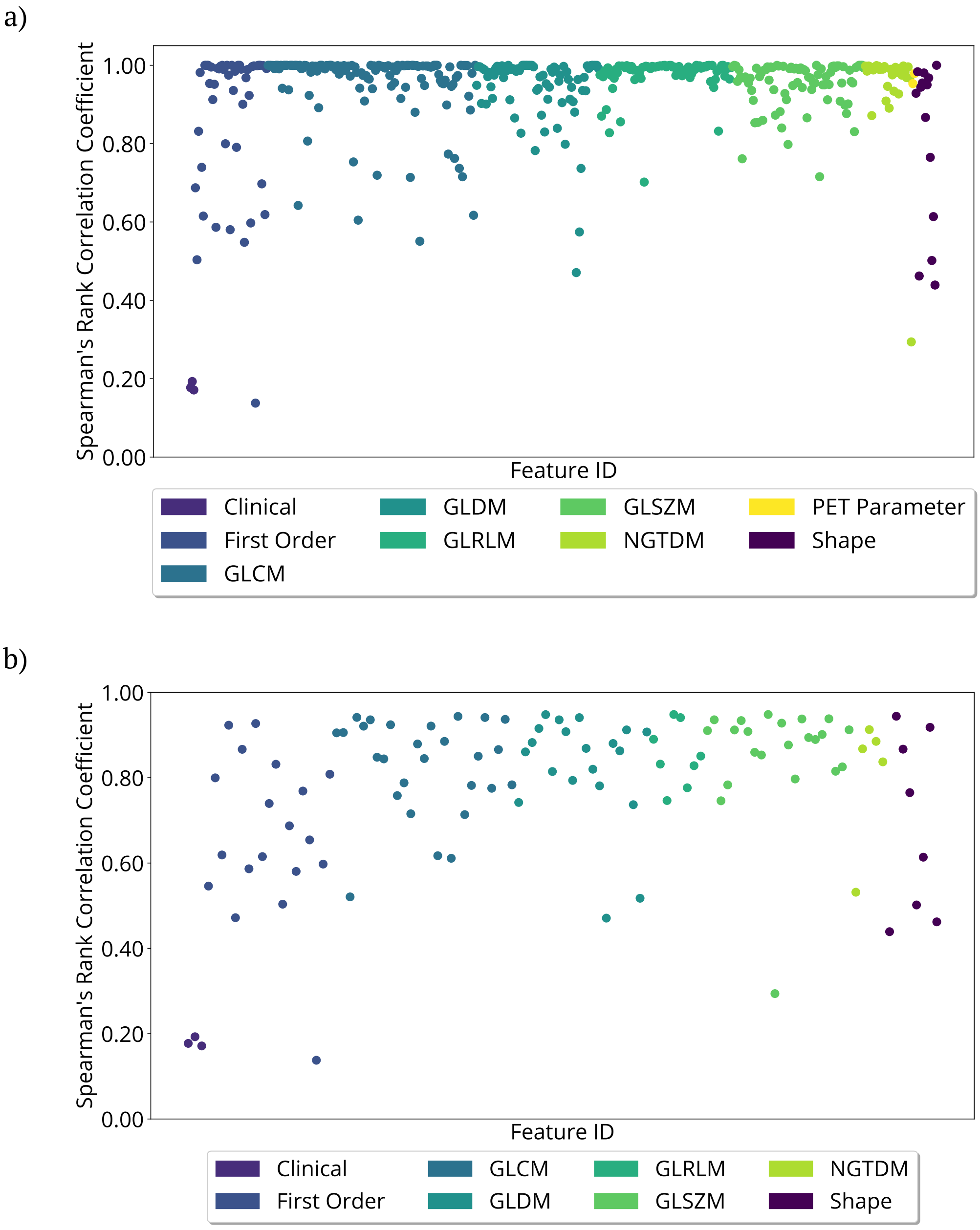


Figure 4.16: The *Spearman's Rank Correlation Coefficient* of a) the 513 in the standard feature matrix, and b) 152 features retained after removal of intra- and inter-feature correlations. Abbreviations to feature categories are given in Table 4.2.

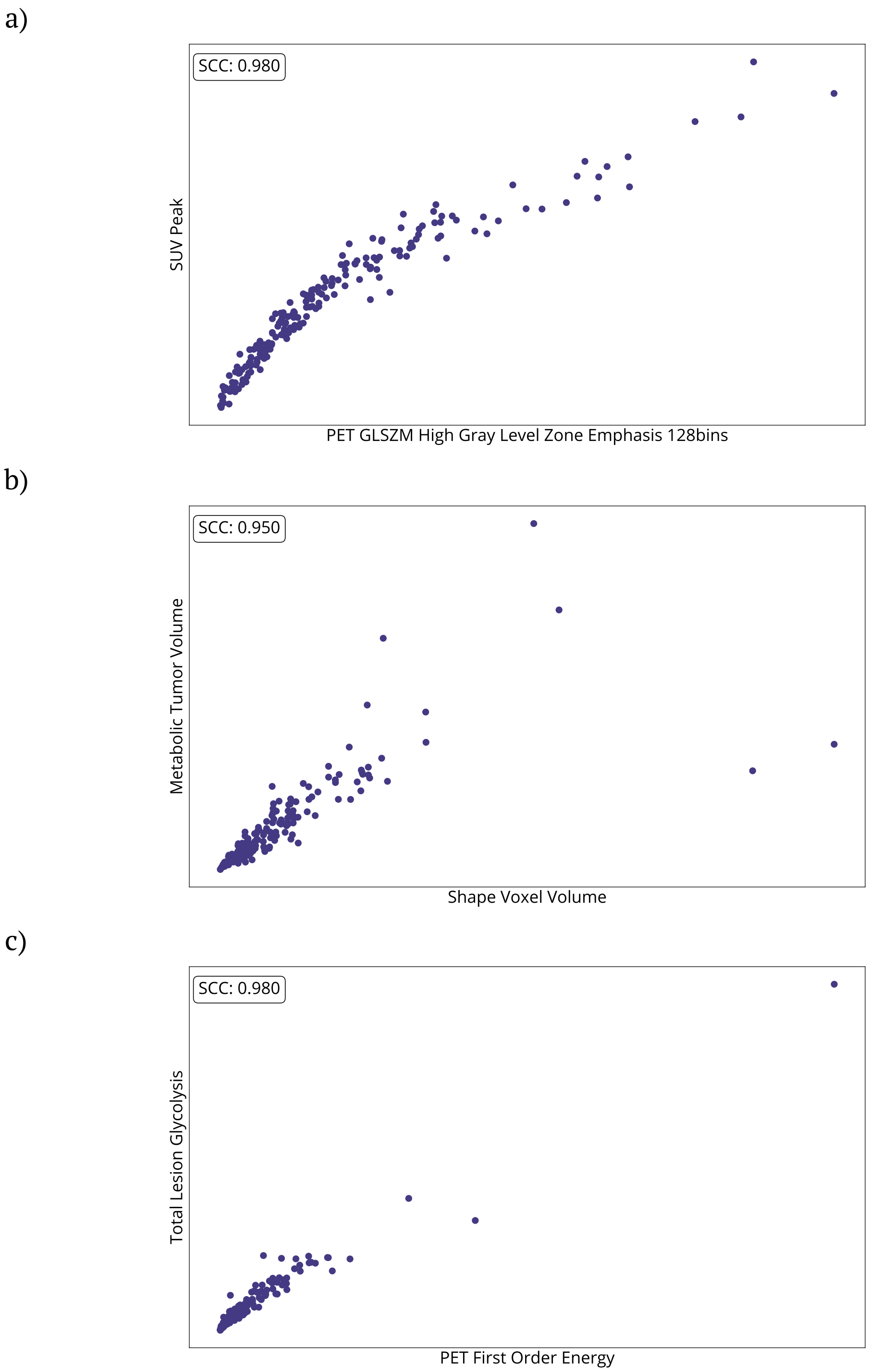


Figure 4.17: The relationship between PET parameters (vertical axis) and radiomics features (horisontal axis). Abbreviations: Spearman's Rank Correlation Coefficient (SCC).

Classifying Disease-Free Survival Using Redundancy Filtered Features

The mean wAUC scores from classifying disease-free survival using the standard feature matrix subjected to filtering and removal of intra- and inter-correlated features, as described in Section 3.8.3, is given in Figure 4.18. The no information rate in Experiment 5 was 67 % disease-free survival. Abbreviations to classification algorithms are given in Table 4.1.

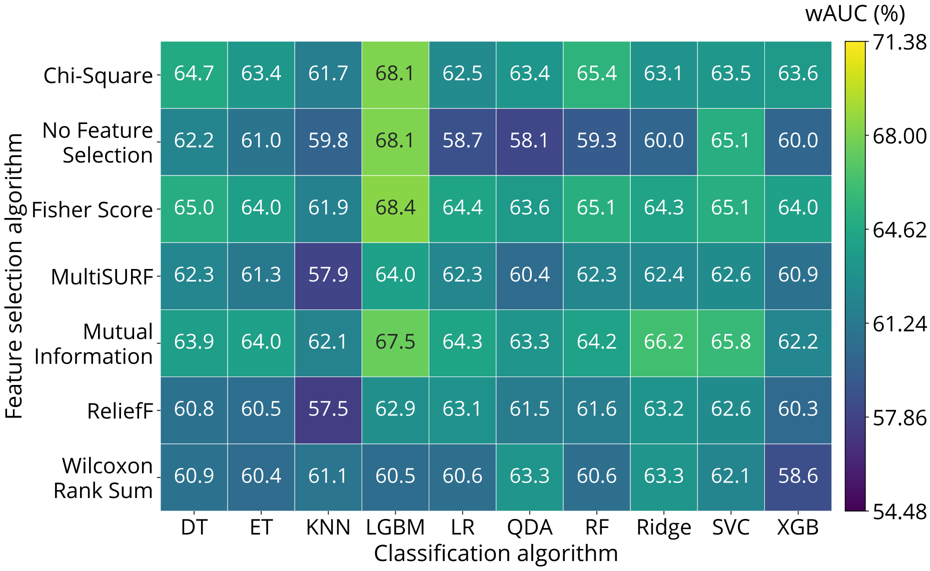


Figure 4.18: Average wAUC (%) from including features retained after removal of intra- and inter-feature correlations to classify disease-free survival with combinations of feature selection (vertical axis) and classification (horisontal axis) algorithms. The colour bar shows that a higher score corresponds to more correct classifications.

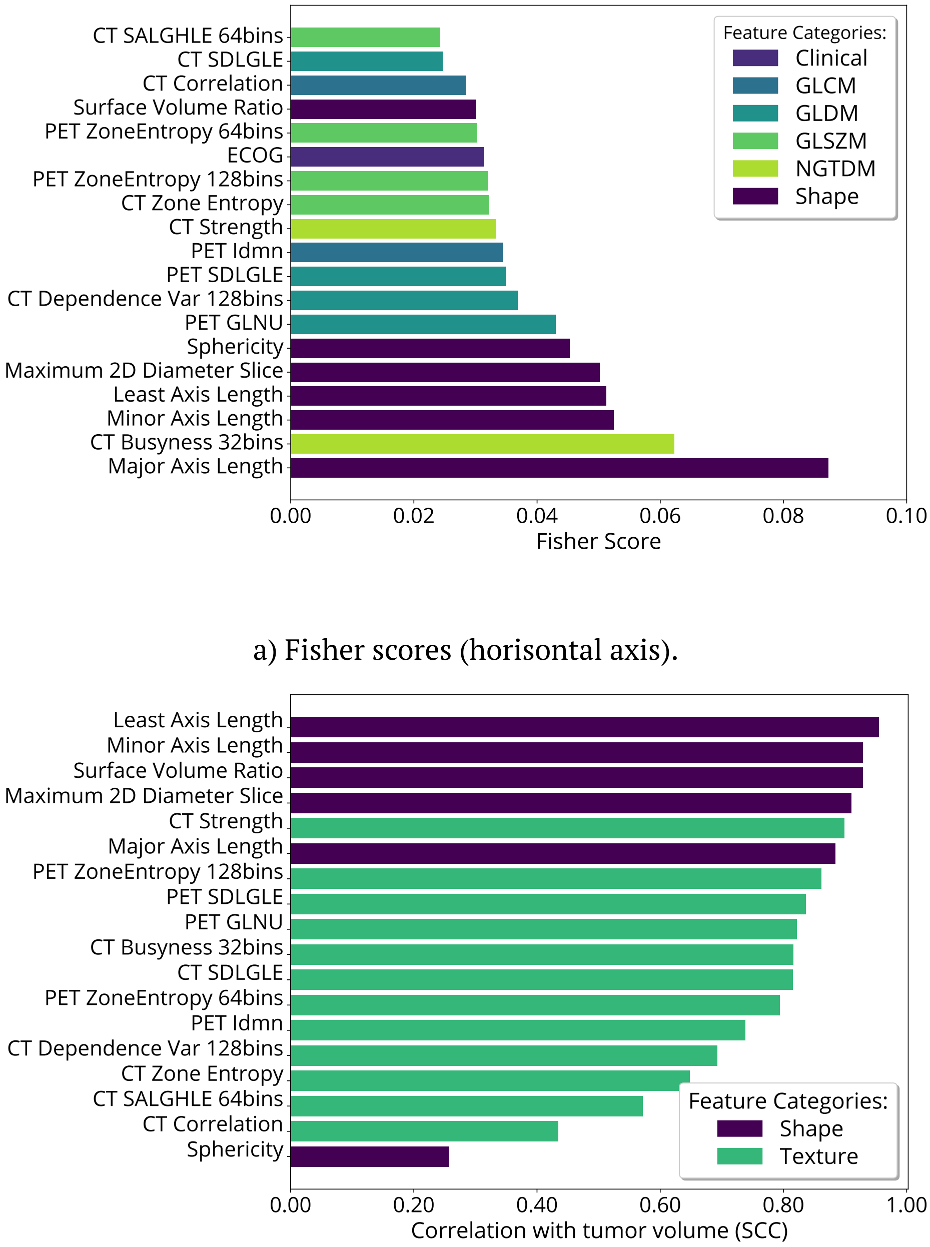
The highest score in Figure 4.18, exceeding 68 % wAUC, was achieved with Fisher score feature selection and the Light Gradient Boosting Machine (LGBM). Note that superior performance was also obtained with LGBM either in combination with Chi-Square or Mutual Information feature selection or by omitting feature selection. ReliefF and *K-Nearest Neighbours* (KNN) gave the lowest score of about 58 % wAUC. Combining Fisher score with Logistic Regression (LR), Ridge Classification or Support Vector Classifier (SVC) gave approximately 65 % wAUC. The standard deviations of the wAUC scores in Figure 4.18 ranged from 13 % to 19 % wAUC.

4.2.4 Preliminary Estimates of Feature Relevance

The Fisher Score and LGBM algorithms were used to rank the subset of 152 features retained after removing intra- and inter-correlated features from the standard feature matrix. These two algorithms were selected based on their superior performance, illustrated by Figure 4.18. The procedures are described in Section 3.8.3.

The Fisher Score algorithm was configured to select 19 features, as the average number of selected features in the classification experiment. The average hyperparameter configuration selected for LGBM is given in Appendix B.2, Table B.1. Table B.1 shows that the average LGBM model included 101 tree base estimators, with an average depth of 165 levels, which illustrates the complexity of the model.

The features corresponding to the 19 highest Fisher scores, determined by the Fisher Score algorithm, is given in Figure 4.19 a). Moreover, Figure 4.19 b) shows the SCCs between these features and the ROI. Abbreviations to texture feature categories are given in Table 4.2.



a) Fisher scores (horisontal axis).

b) SCC values (horisontal axis).

Figure 4.19: The a) 19 features (vertical axis) determined by Fisher scores (horisontal axis) as the most relevant for classifying disease-free survival, and b) the SCC (horisontal axis) between these features (vertical axis) and the ROI. A higher Fisher score indicates higher relevance towards disease-free survival.

Primarily shape features, such as the Major Axis Length and Sphericity, have been assigned the highest ranks in Figure 4.19 a). Moreover, CT Busyness and PET Gray Level Non-Uniformity are the highest ranked texture features. Figure 4.19 a) also shows that first-order features were not recognised by Fisher Score as particularly prognostic of clinical endpoints. However, only Sphericity was less correlated to ROI size than 0.8 SCC, according to Figure 4.19 b).

The 19 features shown in Figure 4.19 were ranked using the LGBM classifier and *Shapley Additive Explanations* (SHAP) values [81], described in Section 2.6.1. A SHAP value represents the average contribution of a feature to the predictions of a model. Only the three of the 19 features ranked by Fisher Score associated non-zero mean absolute SHAP values. The SHAP values for these features are shown in Figure 4.20. Note that a higher SHAP values signify greater feature relevance, while a SHAP value of zero indicates that a feature does not affect model predictions.

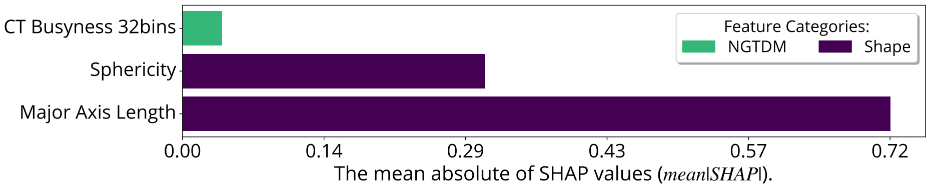


Figure 4.20: The relevancy of features (vertical axis) towards disease-free survival in terms of *Shapley Additive Explanations* (SHAP) values (horisontal axis). A higher mean absolute SHAP value indicates greater relevance.

According to Figure 4.20, the Major Axis Length and Sphericity shape features were the most relevant to the predictions of the LGBM model.

4.2.5 Classification of HPV Subgroups

The patient cohort was divided into two subgroups referred to as *HPV related* and *HPV unrelated*, as described in Section 3.8.3.

Classifying the HPV Related Cohort

Figure B.2 in Appendix B.3 shows the results from classifying treatment outcomes in the HPV related cohort. The highest wAUC score in both classification experiments, including the standard feature matrix and the features retained after correlation filtering and removal, was 70 % wAUC. A combination of Wilcoxon Rank Sum feature selection algorithm and *Decision Tree* (DT) classification achieved this score with features from the standard feature matrix. The HPV related cohort no information rate was 73 % disease-free survival.

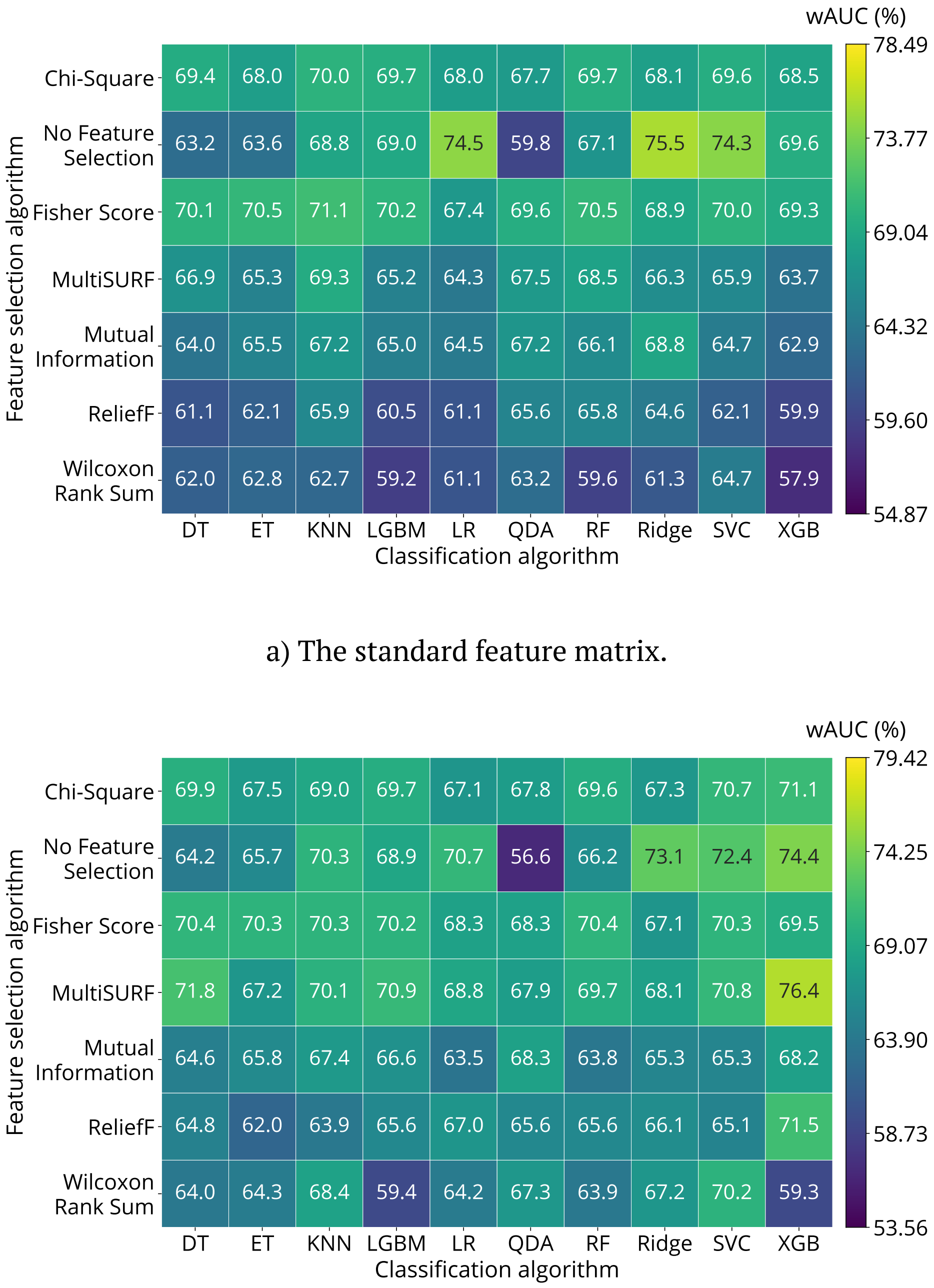
Classifying the HPV Unrelated Cohort

Results from classifying disease-free survival in the HPV unrelated cohort are shown in Figure 4.21. The no information rate in the HPV unrelated cohort was 54 % disease-free survival. Abbreviations to classification algorithms are given in Table 4.1.

Using the standard feature matrix, the highest performance in Figure 4.21 a) of 75.5 ± 15.1 % wAUC was obtained using Ridge Classification without prior feature selection. Furthermore, combining *MultiSURF* feature selection with *Extreme Gradient Boosting* (XGB) classification gave the highest score in Figure 4.21 b) of 76.4 ± 13.2 % wAUC. This score was achieved with the subset of features retained after removal of intra- and inter-correlated features from the standard feature matrix, described in Section 3.8.3. The standard deviation of the wAUC scores in Figure 4.21 ranged between 13 and 19 % wAUC in both experiments.

Training and Validation Performances

Figure 4.22 shows the training and validation performance of the MultiSURF and XGB model.



a) The standard feature matrix.

b) Removal of intra- and inter-correlated features.

Figure 4.21: Average wAUC (%) from including a) the standard feature matrix, and b) features retained after removal of intra- and inter-feature correlations to classify disease-free survival in the HPV unrelated cohort with combinations of feature selection (vertical axis) and classification (horisontal axis) algorithms. The colour bar shows that a higher score corresponds to more correct classifications.

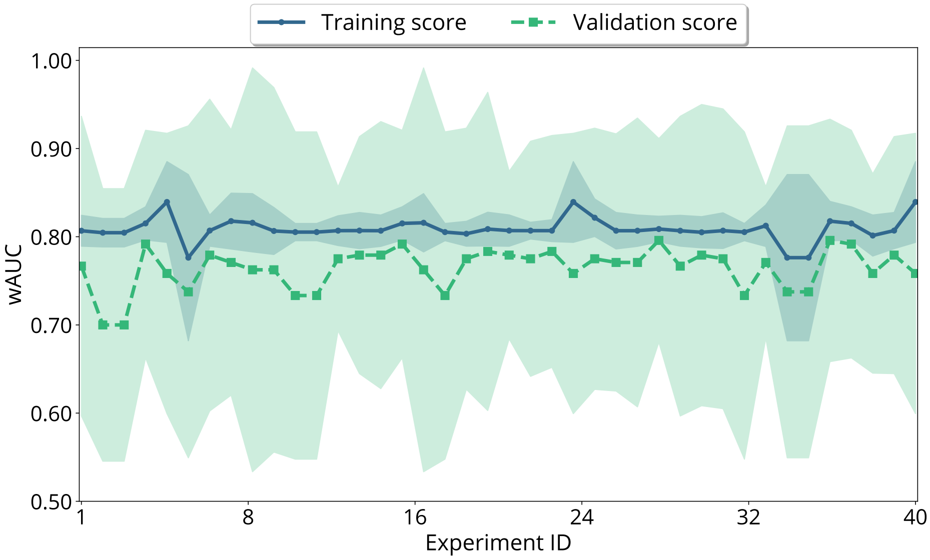


Figure 4.22: Average training and validation wAUC (vertical axis) of the combined MultiSURF and XGB model for each repeat (horisontal axis) of classifying disease-free survival in the HPV unrelated cohort. Shaded areas represent the standard deviation of the wAUC.

The largest difference between training and validation performance in Figure 4.22 was approximately 10 % wAUC. Moreover, the validation performance is consistently lower then the training performance.

The Ridge classification model without prior feature selection achieved the second-highest score in classification experiments. Figure 4.23 shows the training and validation performance of the Ridge model from classifying the HPV related cohort using the standard feature matrix.

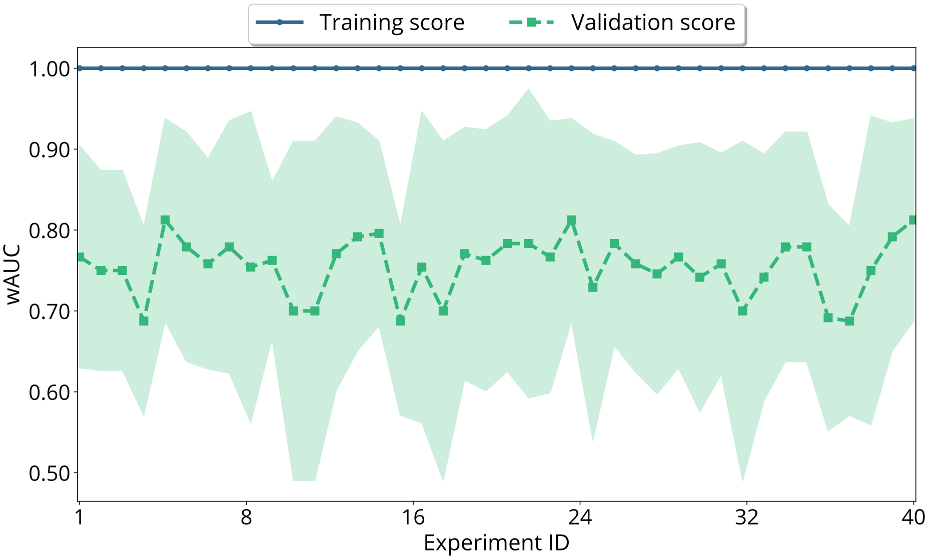


Figure 4.23: Average training and validation wAUC (vertical axis) of the Ridge Classification model for each repeat (horisontal axis) of classifying disease-free survival in the HPV unrelated cohort. Shaded areas represent the standard deviation of the wAUC.

Figure 4.23 shows that the Ridge model was maximally over-fitted in each repeat of the classification experiment.

4.2.6 Reassessment of the Radiomics Hypothesis

The wAUC scores from classifying the HPV unrelated cohort using only clinical factors, as described in Section 3.8.3, are shown in Figure 4.24.

MultiSURF Feature selection and a Support Vector Classifier (SVC) achieved 64 % wAUC as the highest score in Figure 4.24. This result was inferior to highest wAUC score in Figure 4.21 where also PET parameters and radiomics features were included. Standard deviations of the wAUC scores in Figure 4.24 ranged from 5 % to 19 % wAUC.

Figure 4.24 shows the selection rate of the 13 clinical factors included in analyses of the original (Section 4.2.1) and the HPV unrelated (Section 4.2.5) cohorts.

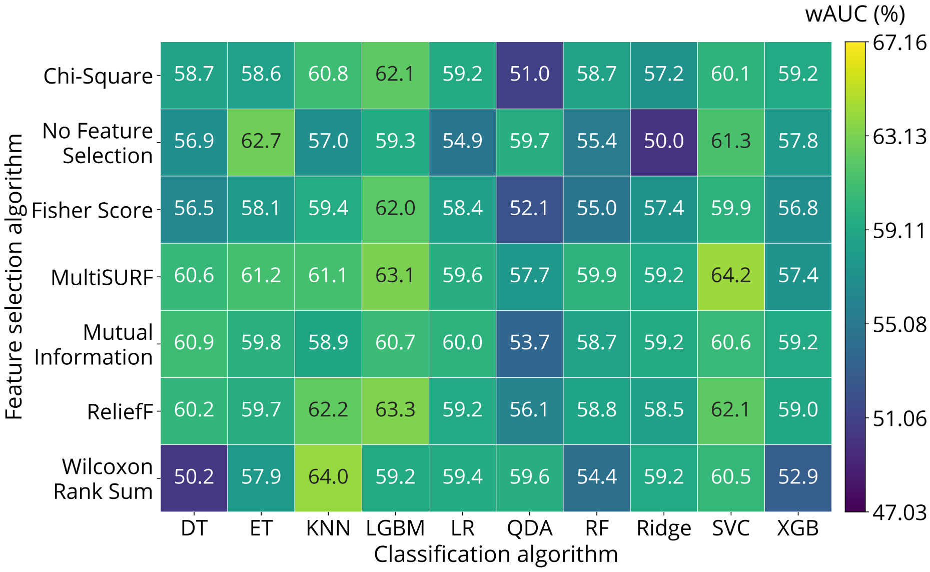
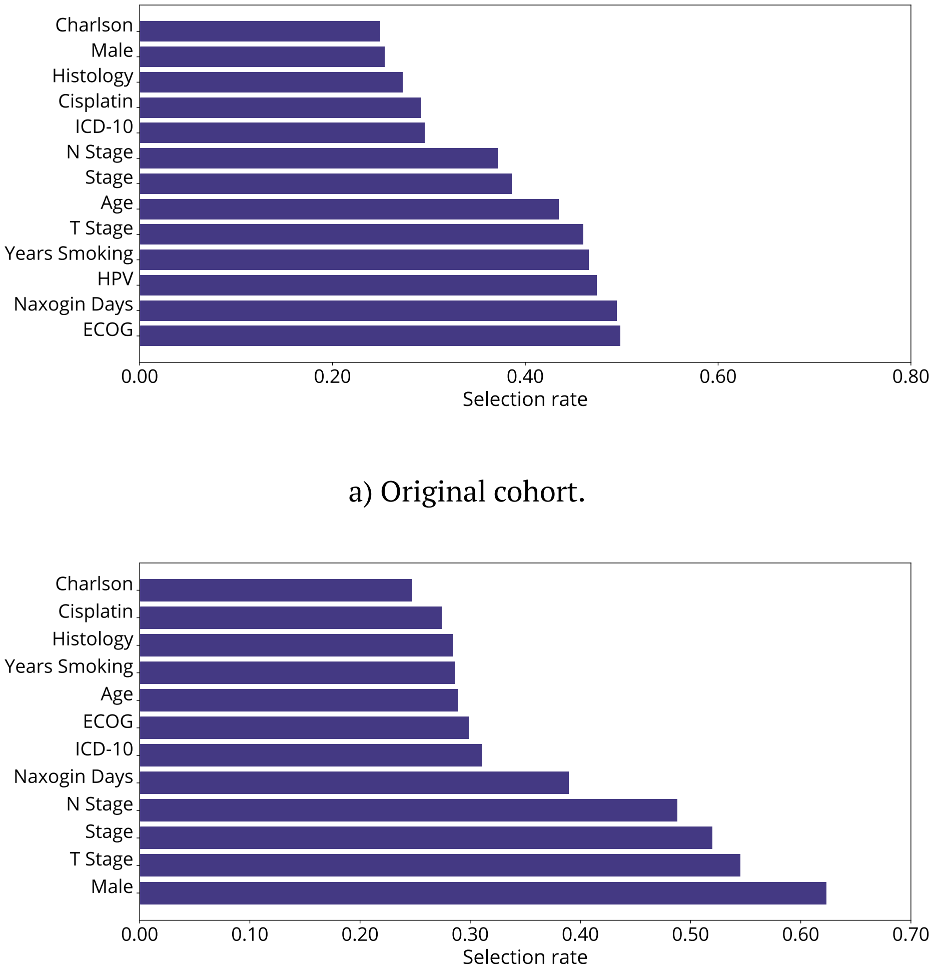


Figure 4.24: Average wAUC (%) from including only clinical factors to classify disease-free survival in the HPV unrelated cohort with combinations of feature selection (vertical axis) and classification (horisontal axis) algorithms. The colour bar shows that a higher score corresponds to more correct classifications.

Dividing the original cohort into HPV subgroups increased selection of tumour *Stage*, and *T*- and *N-Stage*, as shown in Figure 4.25. Moreover, the most selected feature shifted from ECOG to patient gender.

4.3 Selecting a Model for Inference on Feature Relevance

The model combining MultiSURF feature selection and XGB classification achieved the overall highest score from classifying treatment outcomes the HPV unrelated cohort, described in Section 4.2.5. A 95 % confidence interval of (70.5 %, 77.0 %) wAUC was calculated for the performance of this model using the bootstrap method from Section 2.7.4. The bootstrap method was used since the wAUC scores were normally distributed according to the Shapiro and the K^2 tests using a 95 % level of confidence.



a) Original cohort.

b) HPV unrelated cohort.

Figure 4.25: The selection rate (horisontal axis) of clinical factors (vertical axis) when classifying disease-free survival using a) the original cohort of 198 patients, and b) the 67 patients in the HPV unrelated cohort. A selection rate indicates that a feature is selected more often. Abbreviations: *International Classification of Diseases* (ICD), *Charlson Comorbidity Index* (Charlson).

Table 4.4: Descriptive statistics of the hyperparameter configurations selected for the combined MultiSURF and XGB model in the HPV unrelated classification experiment. Abbreviation: standard deviation (STD).

Hyperparameter	**Minimum**	**Mean**	**STD**	**Maximum**
Selected features	6	26	10	47
Learning rate	0.0110	19.0	16.0	47.0
Max tree depth	30	230	160	480
Min leaf samples	2	4	1	5
Number of trees	16	120	53	190
α-regularisation	4.20	5.90	1.20	7.90
λ-regularisation	0.00	47.0	33.0	99.0

A summary of the hyperparameter configurations selected for the combined MultiSURF and XGB model in the HPV unrelated classification experiment, is given in Table 4.4.

The learning curve for the MultiSURF and XGB model, using the average hyperparameter configuration in Table 4.4, is shown in Figure 4.26. The learning curve illustrates the model wAUC scores for a sequentially increasing number of training samples. Figure 4.26 only includes scores obtained from using at least 60 % of the training data since this amount of observations was required for the model to gain any statistical power.

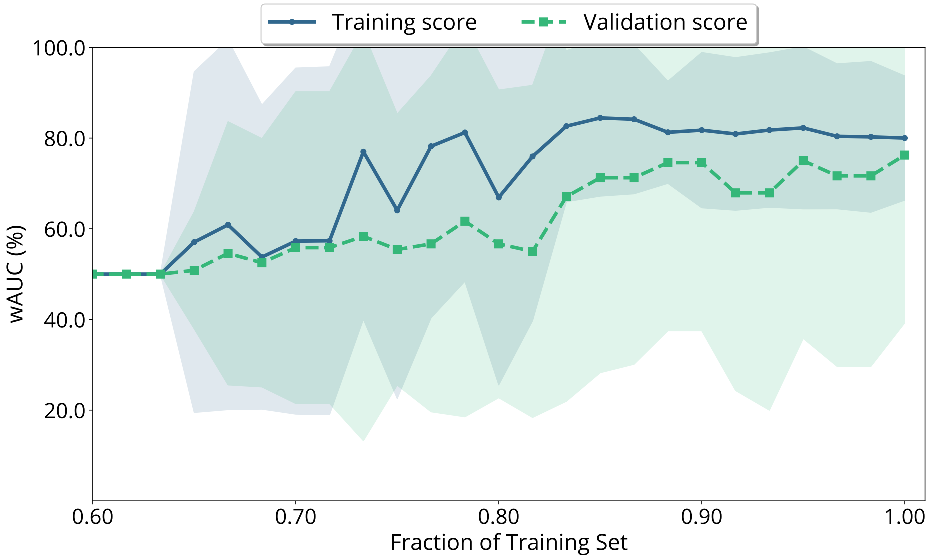


Figure 4.26: The wAUC scores (vertical axis) of the combined MultiSURF and XGB model for an increasing number of training samples (horisontal axis). Shaded areas represent the standard deviation of the wAUC scores.

Both the training and validation performance shown in Figure 4.26 increases successively with the size of the training until the performance stablises around 85 % of the training set. However, only the standard deviation of the training wAUC score decreases. The smallest training error was obtained using the full training set.

4.4 Potential Biomarkers

The subset of features obtained by removing intra- and inter-correlated features from the standard feature matrix was ranked, using the MultiSURF and XGB algorithms, according to relevance for classifying the HPV unrelated cohort. The procedure is outlined in Section 3.8.3.

Preliminary estimates of feature importance was obtained with MultiSURF for the 26 most prognostic features to disease-free survival. This number of features, given in Table 4.4, was the average configuration selected for MultiSURF during classification of the HPV unrelated cohort. Each feature was assigned a weight by MultiSURF, as

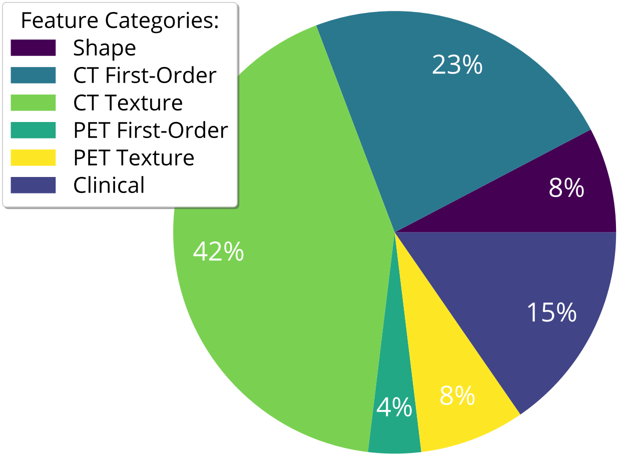


Figure 4.27: The categorical distribution of the 26 features determined by the MultiSURF algorithm as the prognostic to disease-free survival in the HPV unrelated cohort.

described in Section 2.4.2, to represent the relevance of the feature towards disease-free survival. The distribution of the selected features grouped according category is illustrated in Figure 4.27.

Figure 4.27 shows that the majority of the 26 features selected by MultiSURF originated from the CT texture category, followed by CT first-order and clinical factors. The PET first-order features constitutes the smallest category in Figure 4.27.

The weights of the 26 features selected by MultiSURF are shown in Figure 4.28. Larger weights signify higher relevance of disease-free survival. Abbreviations to texture feature categories are given in Table 4.2.

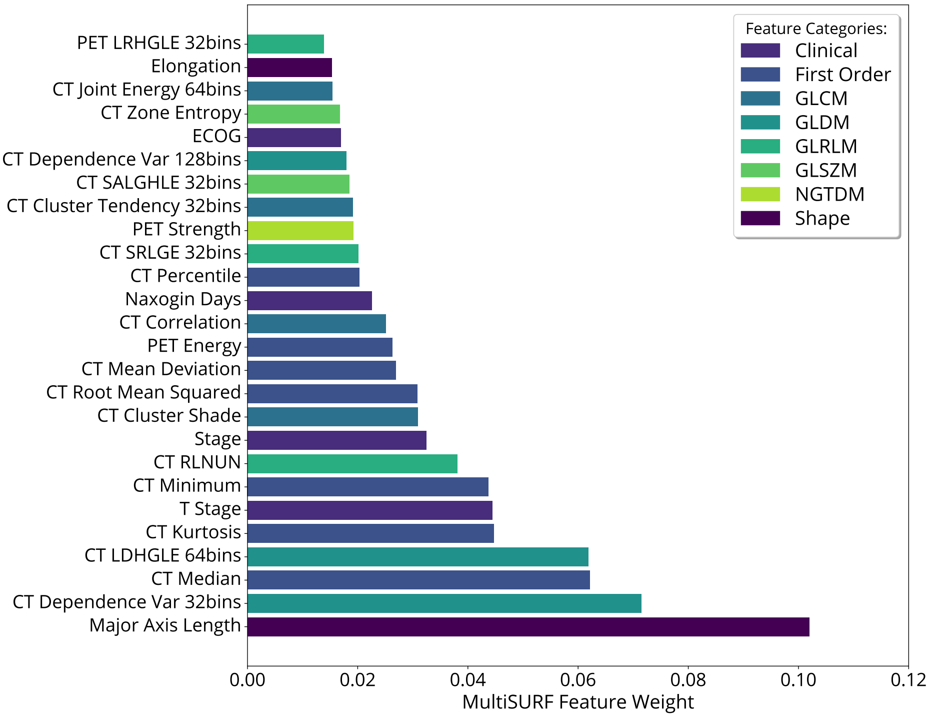


Figure 4.28: The 26 features (vertical axis) with the highest MultiSURF weights (horisontal axis) that quantifies feature relevance for classifying disease-free survival in the HPV unrelated cohort.

Figure 4.28 shows that Major Axis Length, median CT intensity, and the CT texture features *Dependence Variance* (Dependence Var) and *Large Dependence High Gray Level Emphasis* (LDHGLE) were the four highest ranked features according to MultiSURF weights. Note that the clinical factors *T Stage*, *Tumour Stage* (Stage) and ECOG were also selected.

The SCCs between the ROI size and the features show in Figure 4.28 are given in Figure 4.29.

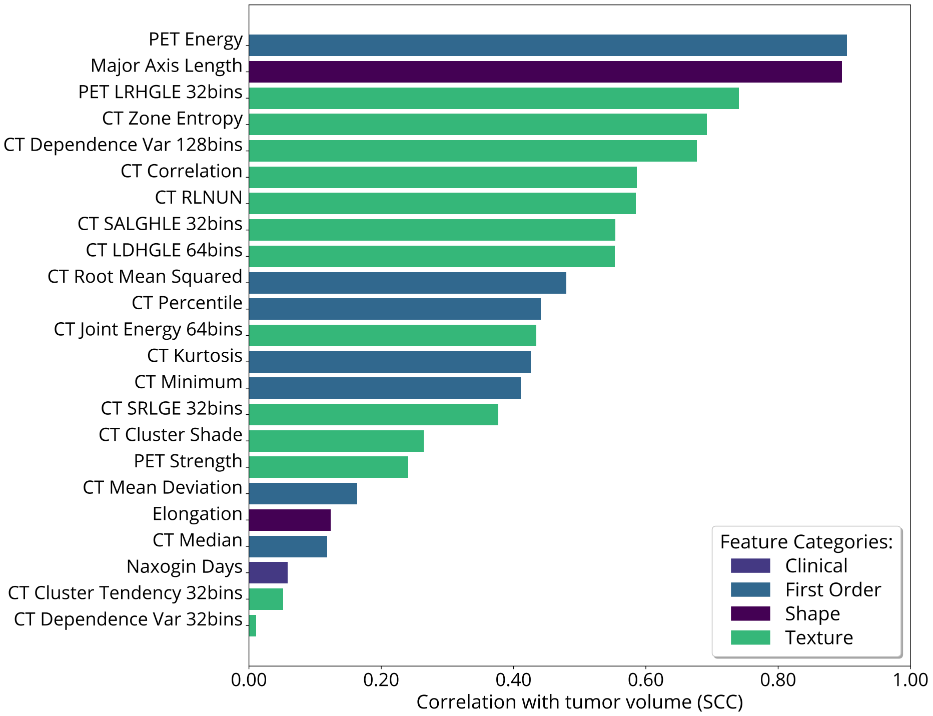


Figure 4.29: The SCC (horisontal axis) between the ROI and the 26 features (vertical axis) selected by MultiSURF as the most prognostic of disease-free survival.

Note that PET Energy corresponds to the highest SCC in Figure 4.29, as the feature strongest correlated to ROI, followed by Major Axis Length. Both 64 bins CT LDHGLE and 32 bins CT Dependence Variance associates less than 0.6 SCC, and are less correlated with ROI compared to to 128 bins CT Dependence Variance.

The feature relevance of disease-free survival using XGB was estimated by calculation of SHAP values, which is described in Section 3.8.3, based on the 26 features selected by MultiSURF. The SHAP values represent the average contribution of features to a model prediction. Four of the 26 ranked by MultiSURF were associated with a non-zero mean absolute SHAP value, as shown in Figure 4.30.

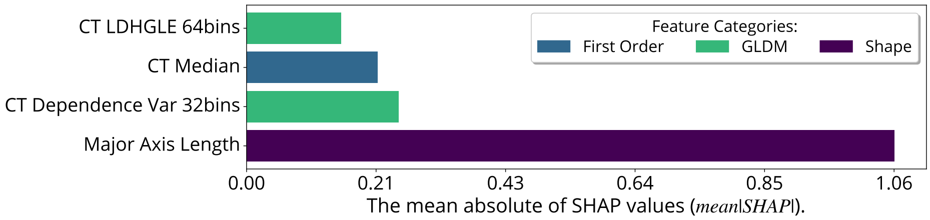


Figure 4.30: The relevancy of features (vertical axis) towards disease-free survival in terms of *Shapley Additive Explanations* (SHAP) values (horisontal axis). A higher mean absolute SHAP value indicates greater relevance.

Observe that the four features in Figure 4.30 corresponds to the four highest ranked features in Figure 4.28. Moreover, the SHAP value for Major Axis Length is more than three times as high as for the three other features.

The distribution of the features in Figure 4.30 are available in Appendix C, Figure C.1. Moreover, scatter plots illustrates the relationships between pairs features included in Figure C.2.

Chapter 5

Discussion

5.1 The Model Comparison Protocol

Motivated by the *No Free Lunch* theorems [23], an emphasised topic in this thesis was to reduce the bias in estimated classification error of the compared models. Studies on schemes to assess model performance have found that the nested stratified K-Fold cross-validation (CV) gave the least biased estimates [131], [92], [95], [94].

5.1.1 Nested Stratified K-Fold Cross-Validation

Despite the small bias, a drawback with the nested CV approach is the computational complexity of nested iterations [132]. Given C number of hyperparameter configurations to evaluate, the CV procedure trains $C \cdot K$ models. A nested CV scheme, however, trains $C \cdot K$ models as part of model selection as well as one model for each of the K validation folds in the outer loop. Thereby, the running time for the nested CV protocol depends quadratically on the choice of K which should be be selected in according to the study objective [133].

Five or 10 folds are typical choices of K [93]. A 5-fold nested CV scheme produces 25 models which requires less computational time compared to the 100 models trained during 10-fold nested CV. On the other hand, the effect of incrementing K has been shown to reduce the bias in model performance estimates, since more observations are assigned to the training fold [126], [125].

In this thesis, classification of the complete cohort of 198 patients (Experiments 1-5, Section 3.8.3) were performed with five folds to reduce the computational complexity of experiments. Moreover, the number of folds was increased to 10 in HPV subgroup analyses (Section 3.8.3) to account for fewer observations. The smallest number of patients included in a classification experiment were 67 patients. With this number of observations, 5- and 10-fold CV would produce training sets of approximately 54 and 60 patients, respectively. In a nested protocol, these sets are divided once more into training sets of 44 and 54 patients. Note that even though 10-fold CV produces a larger training set than 5-fold CV, the size of the validation set is proportionately diminished. Reduced size of the validation set may potentially increase the variability in model performance estimates. Moreover, although bias reduction may generally be referred to as the main objective in model selection, variance reduction has shown to be essential to reduce model over-fitting [95]. Furthermore, reducing the size of the validation set could hinder sample stratification [92].

5.1.2 Stratified Sampling

A property of stratification, as used in this thesis, is the reflection of the original distribution of clinical outcomes in each CV fold [92]. A further property is that stratification provides coverage for all subgroups in each fold. In their study on the impact of class distribution on classification trees, Provost and Weiss (2011) found that stratification of outcomes contributed to reducing model over-fitting [134]. This improved ability of the model to generalise was caused by preventing over- or under-representation of categories [134]. Nevertheless, stratification has, to the knowledge of the author, not been previously used in radiomics.

5.1.3 Hyperparameter Optimisation

The configuration of an algorithm affects the ability of the model to recognise patterns [82]. This makes hyperparameter optimisation a central part of model selection. However, as pointed out by Parmar et al. (2015) [17], manual optimisation may require a certain level of experience with the algorithm. Automated procedures, such as *Sequential Model-based Algorithm Configuration* (SMAC) [88], can be used to eliminate potential bias related to manual optimisation.

However, since SMAC uses a *Random Forest* (RF) [71] as surrogate model, the outcome can be sensitive to the parameters of this model. [88]. No attempts were made in this thesis to optimise the surrogate model, and the default configurations were used in all classification experiments. Previous studies have reported the superiority of SMAC to optimise hyperparameters of algorithms such as *Support Vector Machines* [135], *Logistic Regression* [84] and *deep neural networks* [136] without optimisation of the surrogate model. As for the eligibility of different surrogate models, Eggensperger et al. (2014) [136] concluded that surrogates based on RF outperformed the *Tree Parzen Estimator* [90] and the *Gaussian process* [135].

Furthermore, Hutter, Hoos, and Leyton-Brown (2011) describe the challenge of representing the surrogate surface [88]. That is, similarly to high-dimensional feature spaces, an exponentially increasing number of samples is required to represent the surrogate surface once the hyperparameter space increases in dimensionality. The number of hyperparameters optimised in classification experiments in this thesis varied from two to eight and over a fixed number of 80 combination evaluations (Section 3.8.2). This means that for some models with fewer parameters, a more thorough hyperparameter search could be conducted. For instance, given a model with two hyperparameters, 40 configurations could be suggested. However, if the algorithm associated eight hyperparameters, only 10 configurations were evaluated.

5.2 The Radiomics Hypothesis

Studies have demonstrated that medical images capture information on disease characteristics, which is hypothesised in radiomics [18], [17], [22]. Moreover, studies have also shown that combining radiomics features with clinical factors increase the ability of models to predict clinical outcomes [19], [137].

Classification experiments were performed to assess the prognostic value of combining clinical factors with PET parameters and radiomics features to predict disease-free survival (Section 3.8.3). Results, in terms of wAUC scores (Section 3.8.1), showed that model performances improved by combining these three sets of features (Figure 4.8). This means that the predictive performance of classification models improved by combining clinical factors with PET parameters and radiomics features. However, this thesis did not investigate the amount of predictive information included in each set of features. That is, the contribution from clinical factors, PET parameters and radiomics features to model performances separately was not quantified.

A combination *Fisher Score* feature selection and *Light Gradient Boosting Machine* (LGBM) classification achieved the superior performance in these experiments. The LGBM model appeared to be relatively robust to feature selection since *Fisher Score*, *Chi-Square* and *Mutual Information* combined with LGBM all gave wAUC scores of approximately 67 %. Omitting feature selection also gave 67 % wAUC, suggesting that LGBM performed internal selection of features, which is a property of tree-based models [69]. Furthermore, the *Exclusive Feature Bundling* (EFB) procedure (Section 2.5.8) combines features with similar values to create features that were not originally included in the data set [77]. As a consequence of EFB, feature redundancy is reduced.

Both LGBM and the *Extreme Gradient Boosting* (XGB) builds on the *Gradient Boosting Decision Tree* algorithm (Section 2.5.8). Still, the highest wAUC score achieved with XGB classification was about 64 % wAUC in combination with Fisher Score feature selection. As opposed to XGB, LGBM builds the tree ensemble from subsets of training data using the Gradient-based One-Side Sampling (GOSS) procedure (Section 2.5.8). The GOSS procedure can render LGBM more robust to sample noise and superior to XGB in some problems.

5.3 The Impact of Image Artefact Correction

Bone structures and streak artefacts were identified in approximately 14 % of the CT image stacks (Section 4.2.2). The procedure performed to remove these artefacts was based on the guidelines proposed by Ger et al. (2018) [115].

5.3.1 Feature Stability Towards Removal of Images and Slices

In Ger et al. (2018), the authors used the *pairwise t-test* [26] to assess the impact of CT slice removal on feature stability. However, statistical tests indicated that the difference between features calculated in this thesis from original and artefact-filtered images could not be assumed normally distributed (Section 3.4.4). Therefore, the *Wilcoxon Signed-Rank* test (WSR) [27] was used as an alternative to the t-test to compare features extracted from original and artefact-filtered images.

The WSR test showed that at least 76 % of the CT texture features extracted before and after removal of image slices were likely to originate from the same distribution (Table 4.3). This result supports Ger et al. (2018), where the impact of removing CT slices was also considered statistically insignificant for 76 % of the features [115].

According to the WSR test, more texture features then first-order features were changed by artefact correction. This outcome may be explained for CT features by the changes to the spatial arrangement of voxels in the ROI caused by the removal of slices. However, excluding 11 image stacks alone rendered 30 % of the PET texture features significantly different from their original distributions. This means that removing stacks had a greater impact on PET texture features compared to CT, although CT features were also affected by the removal of slices. Still, the effect of only removing image stacks was not studied for CT features. Therefore, whether or not CT features were only affected by the removal of stacks, or if removal of slices also altered the feature distributions, was not established.

5.3.2 Model Performances with Artefact Corrected Features

Results from classifying patients before and after image artefact correction suggests that the effect of removing artefacts was negligible in terms of model performances. Both classification experiments based on the standard and artefact corrected feature matrices (Section 3.5) gave approximately 67 % wAUC as the highest scores. A similar observation was reported by Leijenaar et al. (2018) when predicting HPV status from radiomics features [138]. Leijenaar et al. (2018) [138] found that the AUC scores obtained from including and excluding CT artefacts were not significantly different according to *DeLong's* test [139].

5.3.3 Changes in Feature Selection From Removing Artefacts

The two most selected texture features prior to artefact removal were 32 bins PET *Informal Measure of Correlation 1* (Imc1) and 64 bins CT *Gray Level Non-Uniformity* (GLNU) (Figure 4.12). According to the definitions of these features (Appendix E, Section E), Imc1 compares the probability distributions between two intensity values using mutual information, while GLNU measures the variability in image intensity values [16]. Contrary to Imc1, GLNU may be sensitive to removal of CT slices, while also being affected by the removal of image stacks. However, missing CT slices also had an impact on the 32 bins PET and CT Busyness features, which became the most selected texture features after artefact correction. Busyness quantifies the rate of change in image intensities within a neighbourhood of voxels [16] and has been recognised as prognostic of clinical endpoints in lung cancer [140], [141].

Figure 4.13 shows that CT Busyness and GLNU were correlated with the ROI size, which indicates that these features primarily described tumour volume. Sphericity, on the other hand, was not considered to be associated with the ROI. Recall that artefact correction had only a direct effect on first-order and texture features (Section 3.4.4). Still, Sphericity was not among the 10 most selected features after the correction (Figure 4.12). Sphericity quantifies the ROI roundness relative to a sphere [16], and was included in a radiomics signature proposed by Aerts et al. (2014) [14].

Artefact correction was not found to increase the selection of first-order features (Figure 4.12), although these features were identified as the most stable towards artefact removal (Table 4.3). These observations indicate that characteristics of the ROI intensity distribution were not recognised as prognostic of treatment outcomes in this data set.

A general explanation for why some features were no longer selected as often after artefact correction as before could be that the selection of other features increased. If Busyness were selected more often, these features may have surpassed Imc1 and GLNU among the 10 most selected features (Figure 4.12).

5.4 Exploiting Feature Redundancy

Due to a large number of features typically studied in radiomics, the *curse of dimensionality* [142] is a frequently encountered challenge. For instance, Wu et al. (2016) found that almost 75 % of the 440 studied radiomics features were significantly correlated [122]. By removing the correlated features, the dimensionality of radiomics data sets can be reduced without significant loss of information.

5.4.1 Modifying Intra-Correlated Features

This thesis successfully applied the methodology proposed by Hassan et al. (2018) to PET and CT features extracted from images containing artefacts. By adjusting the feature definitions, the dependency between CT texture features and the number of image intensity bins was reduced. Recall that the level of image discretisation is determined by the number of intensity bins. Although Hassan et al. (2018) only modified features extracted from lung cancer CT images, this thesis demonstrated the applicability of the proposed modifications to both PET and CT features in head and neck cancer.

This thesis adopted the *Intraclass Correlation Coefficient* (ICC) [31] (Section) used by Hassan et al. (2018) to quantify intra-feature correlation [119]. Considering the ICC scores obtained prior to feature adjustments (Figure 4.14), the features studied in this thesis appeared to be relatively stable towards varying image discretisations. It was found that approximately half of the texture features in Figure 4.14 associated ICC scores exceeding 0.5. According to Hassan et al. (2018), features of about 0.5 ICC were considered as stable [119]. One possible explanation for feature stability observed in this thesis could be that images were only discretised into 32, 64 and 128 bins, while, Hassan et al. (2018) [119] used 8, 16, 32, 64 and 128 of bins.

All the features extracted from artefact corrected images were determined to be invariant to image discretisation, with ICC scores surpassing 0.9 before modification (Figure 4.15). Furthermore, modifying these features increased the correlation image discretisation level. The only difference between obtaining features dependent and independent to image discretisation was correction of image artefacts. This implies that image textures, differentiating between features extracted from differently discretised images, were destroyed with the correction operation. However, this thesis did not further investigate these results.

Modifying the features extracted from images containing artefacts reduced their dependency to the number of image intensity bins for all but CT GLNU (Figure 4.14). Contrary to features from artefact corrected images, these features appeared to capture texture variations across image discretisations. Similar to Hassan et al. (2018) [119], all modified features associated at least 0.9 ICC. However, an increase in dependency between intensity discretisation and CT GLNU was not reported by Hassan et al. (2018) although their definition of GLNU was the same as in this thesis (Appendix E, Section E.7).

5.4.2 The Relationship Between PET Parameters and Radiomics Features.

Removal of inter-correlated features was carried out by excluding one of the features in a correlated pair from the data set (Section 3.8.3). A threshold of 0.95 *Spearman's Rank Correlation* (SCC) [34] was used to quantify the degree of correlation. However, since the feature selected for removal was arbitrary, the procedure may had led to suboptimal results.

For instance, all the PET parameters included in this thesis (Section 3.3.3) were removed. There parameters have been recognised by Moan et al. (2019) as prognostic of disease-free survival in HPV unrelated cancers. The SCC between feature pairs revealed that the *SUV peak*, *Total Lesion Glycolysis* (TLG) and *Metabolic Tumour Volume*, described in Section 3.3.3, were correlated with the 128 bins PET GLSZM *High Gray Level Zone Emphasis* (HGLZE), PET first-order *Energy* and *voxel volume* radiomics features, respectively. This indicates that information similar to what is captured by the PET parameters were also described by these radiomics features. The PET Energy feature, correlated to TLG and ROI (Figure 4.17), has also shown to be correlated to clinical outcomes in lung cancer [14].

The PET parameters SUV max and SUV mean have been found to correlate with aggressive tumour behaviour and poor prognosis [143], [144]. However, these parameters are incapable of describing the heterogeneous distribution of PET intensities [145], [14]. On the contrary, PET HGLZE (Appendix E, Section E.4) captures such information by considering the distribution of neighbouring voxels with the same intensity [16].

A drawback of this study was not to account for correlations between radiomics features and the ROI. An association between features and the number of voxels in each ROI was demonstrated by Hassan et al. (2018) [119]. The authors proposed adjustments to feature definitions to relax such correlations, but such corrections were not performed in this thesis. Adjusting feature definitions to account for varying ROI sizes could have contributed to reducing the correlation between features.

5.4.3 Model Performances After Removal of Correlated Features

Removal of intra-correlated features from the standard feature matrix (Sectino 3.5) by thresholding reduced the number of features from 513 to 341 (Section 3.8.3). Yet, the same maximum wAUC score was obtained in classification experiments using either the subset of 341 features, or the 513 features in the standard feature matrix. This result demonstrates that removal of 172 did not lead to a significant loss of predictive information related to clinical outcomes.

Removing both intra- and inter-feature correlations (Section 3.8.3) gave a subset of 152 features. Compared to using the standard feature matrix (Figure 4.8 b)), the classification performance was improved for 80 % of the models (Figure 4.18). The highest performance obtained when classifying patients after removal of intra- and inter-feature correlations was 68 % wAUC, and, although classification was not significantly improved, this result demonstrates the amount of redundancy among the features extracted in this thesis.

5.5 Preliminary Feature Relevance

Ranking of features after removal of intra- and inter-correlations from the standard feature matrix using Fisher Score (Section 3.8.3) showed that the *Major Axis Length* shape feature was considered as the most prognostic of disease-free survival. This feature measures the longest axis in the ROI and was the most selected in classification experiments based on the standard and artefact corrected feature matrices (Figure 4.12). These results indicate that Major Axis Length was predictive of disease-free survival, as well as robust towards the removal of image stacks (Section 3.4.4).

Further, Major Axis Length, Sphericity and CT Busyness were the only features found to be relevant to LGBM predictions (Figure 4.20). However, both MAL and CT Busyness were strongly correlated with ROI, implying that these features were potentially only relaying information on ROI size. Sphericity, on the other hand, was not correlated to ROI size but appeared to be sensitive to the removal of image stacks (Figure 4.11).

5.6 HPV Subgroup Analyses

Rather than predicting HPV status, which has been the goal of previous studies[146], [18], [147], HPV status was used in this thesis to study subgroups of patients. Recall from Section 3.8.3 that the original patient cohort was divided into two subgroups referred to as *HPV related* and *HPV unrelated*.

5.6.1 Classifying the HPV Unrelated Cohort

Using *Ridge Classification* model without prior feature selection gave in 75.5 ± 15.1 wAUC with the standard feature matrix. However, comparing the training and validation scores revealed that this model was severely over-fitted. Over-fitting has been described by Parmar et al. (2018) as a common pitfall in biomarker identification [117]. According to Figure 4.23, the model appeared to have memorised all the variations in the training data, which could negatively impact the model when applied to independent test data.

Classifying the HPV unrelated cohort after removing intra- and inter-feature correlations from the standard feature matrix gave 76.4 ± 13.2 wAUC as the highest score. This result points to a strong association between disease-free survival and patients not related to HPV, as was also recognised for this cohort by Moan et al. (2019) [100]. The wAUC score was obtained by using *MultiSURF* feature selection with *Extreme Gradient Boosting* (XGB), as the overall highest in classification experiments. The *no information rate* (Section 3.8.1) in experiments with the HPV unrelated cohort, which represents the majority outcome, was 55 % wAUC. Thereby, the combined MultiSURF and XGB model had gained statistical power to predict treatment outcomes from the features available after filtering and thresholding to remove redundancy.

The *ReliefF* feature selection algorithm, which resembles MultiSURF, has shown superior performance in radiomics [122]. Although ReliefF appears inferior to MultiSURF, the combination of ReliefF and XGB gave almost 72 % wAUC. Recall that ReliefF and MultiSURF are both multivariate filter methods capable of detecting pairs interacting of features (Sections 2.4.2 and 2.4.2). One of the main differences between these algorithms is the definition of the parameter K, which is related to the selection of observations that are used to estimate the relevancy of features (Section 2.4.2). In MultiSURF, this parameter is incorporated into the algorithm, while K is a hyperparameter in ReliefF.

5.6.2 Classifying the HPV Related Cohort

The highest performance in classification of the HPV related cohort, of about 69 % wAUC, was obtained with *Wilcoxon Rank Sum* feature selection (Section 2.4.1) and *Decision Tree* classification. Contrary to classifying HPV unrelated cancers, the highest performance in this subgroup was achieved by using the all available features.

The Wilcoxon Rank Sum algorithm has been reported as superior together with the *Random Forest* classifier in a radiomics lung cancer study [17]. In their study, Parmar et al. (2015) achieved 66 % AUC from classification of patient survival [17]. However, the described superior performance of the Wilcoxon Rank Sum method was not recognised in this thesis. Contrary to MultiSURF, Wilcoxon Rank Sum is a univariate filter method incapable of capturing feature interactions.

The no information rate in the HPV related classification experiment was 73 % and neither of the models achieved a wAUC score surpassing that threshold. This over-representation of only one treatment outcome could have led to model over-fitting (Section 5.1.2). An empirical study by Provost and Weiss (2011) demonstrated that the AUC metric was affected by class imbalance [134], although the purpose of the weighting scheme described in Section 3.8.1 was to account for such class imbalance.

5.7 The Variance in Model Performance Estimates

An increased variation in wAUC scores was observed by comparing results from classifying the HPV unrelated cohort (Section 4.2.5) to results from classifying the complete cohort (Section 4.2.5). The reduced cohort size, as well as the increased number of CV folds from five to 10, are possible explanations for this variability. Still, the relatively small differences between training and validation scores in Figure 4.22 indicate that the general performance of MultiSURF and XGB was well captured, despite such fluctuations. Moreover, a standard deviation of less than 15 % in validation scores indicates that 40 repeats of the nested CV procedure sufficed to capture random variations in wAUC scores.

5.8 Selecting a Model to Infer Feature Relevance

The MultiSURF and Extreme Gradient Boosting (XGB) algorithms were selected to estimate feature relevance of disease-free survival in head and neck cancer. These algorithms achieved the highest performance in classification experiments from classifying the HPV unrelated cohort (Section 4.24). However, results from previous experiments shows that this model was only superior in the HPV unrelated experiment. By dividing the cohort into subgroups to classify HPV unrelated cancers, the model achieving the highest wAUC scores shifted from a combination of Fisher Score and LGBM to MultiSURF and XGB. An explanation for this could be that more complex feature relations became apparent in the HPV subgroup, which was not recognised by the univariate Fisher Score algorithm. Note, however, that both MultiSURF and Fisher Score considers the spatial distance between samples to determine the importance of features.

Again, the main differences between LGBM and XGB are the Gradient-based One-Side Sampling (GOSS) and *Exclusive Feature Bundling* (EFB) (Section 2.5.8), as previously described. Since the LGBM ensemble is built from subset of observations with GOSS, LGBM can handle sample noise, but is also prone to loss of information since only a subset is used to train each model. Further, a drawback of combining features with EFB is that features are prevented from interacting with each other to produce relevant information.

Another limitation of this study was not to record multiple performance metrics to obtain a broader view of model performances. For instance, if the studied outcome is underrepresented, the *precision* and *recall* metrics can be used to obtained probability estimates of correctly classifying the minority class. Combining these metrics with the *harmonic* mean gives the *F-score* [148] as another alternative to measure model performance.

5.8.1 Choosing Hyperparameters

Model over-fitting can, to some degree, be relaxed through hyperparameter optimisation by imposing generalisation on the model [149]. Comparing the training and validation performance of Ridge Classification in the HPV unrelated experiment (Figure 4.21 a)), the SMAC protocol could not prevent the model from overfitting by increasing the regularising. On the contrary, the relatively small difference between the training and validation performance of the combined MultiSURF and XGB model (Figure 4.22) indicates that SMAC selected appropriate hyperparameters for this model. Since the model included both feature selection and classification, SMAC proved capable to handle joint optimisation of two connected algorithms.

The selected hyperparameter combination for MultiSURF and XGB was the average of the configurations recorded in the HPV unrelated classification experiment. Alternative methods to determine a final configuration is to perform a CV on all the training data, or to average configuration weighted by the number of times each setting was selected during the experiment.

Note from Table 4.4, that the average XGB ensemble consisted of 120 tree base models, each tree with an average depth of 230 levels. This illustrates the model complexity, which deviates from what is typically associated with weak learners [73]. Still, the α and λ regularising coefficients (Section 2.5.8) may have compensated for the model complexity by increasing the regularisation.

5.8.2 Interpretation of the Learning Curve

Observing a relatively close relationship between training and validation scores in the learning curve in Figure 4.26 for the combined MultiSURF and XGB model implies that the model was capable of generalising to the data. Lack of over-fitting supports the hypothesis from the previous section that the strong regularisation of XGB compensated for the complexity of the model. However, the standard deviation of the validation scores in Figure 4.26 suggests that 10 CV folds led to significant variations in performance estimates. Such variations were also observed for the model in the HPV unrelated classification experiment (Figure 4.22). Thus, less than 10 folds should therefore have been used in classification epxeriments (Section 3.8.3) to reduce the variability model performance estimates.

5.8.3 Ranking Features with MultiSURF

About 42 % of the 26 features selected by MultiSURF (Section 4.4) as the most prognostic to disease-free survival originated from the CT texture category (Figure 4.27). Moreover, 23 % originated from the CT first-order category. This means that CT features were recognised as more predictive of clinical outcome, compared to PET. The superiority of CT features could be explained by loss of information in PET images due to post-reconstruction filtering (Section 3.3.1).

Among the features that MultiSURF selected were *T Stage*, *Stage*, *Naxogin days* and *Eastern Cooperative Oncology Group* performance status. Recognising stage features as related to clinical outcome coincides with the use of tumour stage for clinical treatment selection in head and neck cancer [3]. The feature ranked highest of all these 26 was the Major Axis Length shape feature, which was also recognised as the most predictive of patients response to treatment in the original cohort (Section 5.5). Studies have indicated that benign tumours are more spherical than malignant tumours [150], and shape features have shown the capacity to distinguish between malignancy and treatment response [151].

The second and fourth highest ranked features were the 32 bins CT *Dependence Variance* (DV), and the 64 bins CT *Large Dependence High Gray Level Emphasis* (LDHGLE). The third highest ranked feature was the median image intensity. Thus, three of the four highest ranked features originated from CT. Both the LDHGLE and DV features measures intratumour heterogeneity which has demonstrated predictive value in radiomics [14], [19]. Noting that the DV and LDHGLE features were determined to be important in 32 and 64 bins images demonstrates that intratumour heterogeneity is expressed at different intensity scales [14].

The CT *Run Length Non-Uniformity* (RLNUN) texture feature was determined by Aerts et al. (2014) as the most prognostic both in head and neck and lung cancers. According to MultiSURF, this feature was the 10^{th} most important for disease-free survival.

5.8.4 Biomarker Identification with Shapley Additive Explanations

Both MultiSURF and SHAP values did rank the same four features as the most predictive of disease-free survival (Section 4.4). The consensus between these methods could be interpreted in the direction that XGB may have adapted to the features selected by MultiSURF and that this gave suboptimal results. However, experimental results (Figure 4.21) showed that the XGB model performed poorer without prior feature selection. This would most likely not have been the situation if MultiSURF did not select features relevant to XGB.

Among the features selected using SHAP values was only Major Axis Length by definition dependent of the ROI (Appendix E, Section E.2). The CT median, Dependence Variance and 64 bins LDHGLE, were all less than 0.5 SCC correlated to ROI size (Figure 4.29). This means that features with relatively weak associations to ROI size were identified as relevant of patient treatment outcome.

5.9 A Software Ecosystem for Radiomics Research

Curated open-source software may serve as standard references for radiomics researchers. An example of such is the PyRadiomics package [16]. To increase the reliability in radiomics software, functionality can be developed to automatically update the configuration files for PyRadiomics feature extraction [99]. This mechanism will relieve the user from manually maintaining these settings, which can increase efficiency and reduce the potential for mistakes. Using material from the *feature_extraction* folder (Section 3.1) in biorad [25], a Python decorator function [152] can be written to automatically update the fields in these files. For instance, if a function is used to calculate some parameters, the decorator will ensure that the appropriate field in the configuration file is update each time a call is made to this function.

Furthermore, the code used to perform classification experiments in this thesis can be extended to include wrapper algorithms for feature selection and to record multiple performance metrics. The current implementations are located in the *experiments* folder (Section 3.1) of the package, and relies heavily on the Scikit-Learn *application programming interface* [153]. To increase the methodological transparency in radiomics, procedures to perform such classification experiments should be readily available.

An implementation of the *Bootstrap Bias Corrected CV* (BBC-CV) procedure, described by Tsamardinos, Greasidou, and Borboudakis (2018) [154], can be used to compare the performance of this method to the nested CV. Tsamardinos, Greasidou, and Borboudakis (2018) proposed BBC-CV as an alternative to nested CV, claiming that the protocol associated smaller variance and bias, in addition to being computationally more efficient [154]. Performing classification experiments to compare these protocols will reveal which is the most computationally efficient and is the least biased. The result will not only have applications to radiomics but to all practitioners of model comparison experiments.

5.10 Suggested Topics to Progress Radiomics Research

The prognostic value of the potential biomarkers identified in this thesis (Section 4.4) can be assessed using an independent HPV unrelated cohort. Quantification of the predictive information in these features will contribute to elucidate the relevance of tumour volume and intratumour heterogeneity in head and neck cancer treatment.

Beam-hardening and streak artefacts, from bone and dental fillings, are common in head and neck cancer imaging [115]. Therefore, the impact of such artefacts on radiomics features should be studied to develop methods for correcting and evaluating the quality of the features. For instance, comparing the distributions of CT features before and after the removal of only image stacks will reveal if this has an equally strong impact on CT features as was observed for PET features in this thesis (Section 5.3.1). Moreover, discarding image stacks if more than 50 % of the ROI was influenced by artefacts is a threshold suggested by Ger et al. (2018) that was not explored in this thesis.

Whether or not interactions between clinical factors, PET parameters, and radiomics features contributed to increasing the performance of classification models was not investigated in this thesis. Classification experiments with only radiomics features or PET parameters would reveal the prognostic information exclusively contained in these features. Moreover, multi-block methods, such as *Sequential Orthogonalised Partial Least Squares* [155] or *Response-Oriented Sequential Alternation* [156], as well as Group LASSO [118], can be used to study the prognostic information in feature subsets with respect to clinical endpoints.

Feature redundancy facilitates an opportunity to reduce the curse of dimensionality [142]. Smaller sets can be obtained by removing correlated features from radiomics analyses while monitoring the amount of predictive information retained in the data. Derivation of additional adjustments to texture features, as proposed by Hassan et al. (2018) [119], reduces the need for texture optimisation, which was demonstrated in this thesis (Section 4.2.3).

This thesis only investigated the relationship between pairs of features. However, Welch et al. (2019) recommend to consider multicollinearity in the analysis of radiomics features [157]. One method to detect multicollinearity is by the *condition number test*, which identifies the correlated variables instead of producing a score value to quantify correlation [158].

Chapter 6

Conclusion

Routine acquisition of medical images facilitates an opportunity for using image-based personalised cancer treatment in clinical practice [4]. Imaging, contrary to tissue biopsy, describes the entire tumour volume and reflects intratumour heterogeneity [6], [10].

Radiomics quantifies physiological tumour characteristics using image descriptors [12]. These descriptors are capable of capturing the intratumour heterogeneity that is hypothesised to have implications for cancer therapeutics and biomarker discovery [8].

6.1 Software for Radiomics Data Analysis

The initial goal of this thesis was to implement methodologies for radiomics data analysis using the *Python*^{TM} [97] programming language. Procedures for radiomics feature extraction and classification of clinical outcomes were based on open-source libraries, such as *PyRadiomics* [16] and *Scikit-Learn* [67]. Moreover, recommendations from the *Image Biomarker Standardisation Initiative* [108] were followed. The code material for all implemented protocols is publicly available via the *GitHub*© web-based hosting service as a package named *biorad* [25]. Development of open-source software may contribute to increase the methodological transparency and reliability of results in radiomics research.

6.2 Searching for Potential Biomarkers

The second goal of this thesis was to explore clinical factors, PET parameters and radiomics features from PET and CT in search of biomarkers prognostic of *disease-free survival*. In a cohort of 198 head and neck cancer patients, disease-free survival was the studied clinical outcome of radiotherapy treatment.

Removal of image stacks and CT slices to account for bone structures and streak artefacts indicated loss of textural information in the artefact corrected images. In order to elucidate these results, further testing of the guidelines proposed by Ger et al. (2018) [115] is encouraged.

Studies of intra- and inter-feature correlations identified 361 of the 513 original features as redundant. This result demonstrates the need for feature refinement to remove superfluous information in radiomics. Modification of radiomics texture features, after Hassan et al. (2018) [119], successfully reduced the correlation between the adjusted features and the levels of image discretisation. Modification of feature definitions can contribute to reducing the need for image texture optimisation.

Dividing the patients into two subgroups by relation to HPV gave 76.4 ± 13.2 % AUC as the highest performance in classification experiments, using a combination of *MultiSURF* feature selection and *Extreme Gradient Boosting* to classify the HPV unrelated patients. The HPV related cohort included 53 % cases of disease-free survival, which demonstrates the potential for identification of prognostic factors in patient subgroup.

Four features were identified as potentially prognostic of disease-free survival. Among these were two CT features quantifying characteristics of intratumour heterogeneity, and the CT median intensity. One feature quantified tumour shape characteristics and was, contrary to the CT features, significantly correlated with tumour volume. This shape feature was also considered the most reliable indicator of disease-free survival. The fourth feature was the CT median intensity. Determining the prognostic value of disease-free survival in these features, using an independent HPV unrelated cohort, will elucidate the relevance of tumour volume and intratumour heterogeneity in treatment of head and neck cancers.

References

[1] World Health Organisation. *IARC Cancer Fact Sheet*. 2018. url: https://gco.iarc.fr/today/data/factsheets/cancers/6-Oesophagus-fact-sheet.pdf.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians*, 68 (6) (2018), pp. 394–424.

[3] Francesca De Felice, Antonella Polimeni, Valentino Valentini, Orlando Brugnoletti, Andrea Cassoni, Antonio Greco, Marco de Vincentiis, and Vincenzo Tombolini. "Radiotherapy controversies and prospective in head and neck cancer: a literature-based critical review". In: *Neoplasia*, 20 (3) (2018), pp. 227–232.

[4] Jimmy J Caudell, Javier F Torres-Roca, Robert J Gillies, Heiko Enderling, Sungjune Kim, Anupam Rishi, Eduardo G Moros, and Louis B Harrison. "The future of personalised radiotherapy for head and neck cancer". In: *The Lancet Oncology*, 18 (5) (2017), e266–e273.

[5] Rajamanickam Baskar, Kuo Ann Lee, Richard Yeo, and Kheng-Wei Yeoh. "Cancer and radiation therapy: current advances and future directions". In: *International journal of medical sciences*, 9 (3) (2012), p. 193.

[6] Patrick Grossmann, Olya Stringfield, Nehme El-Hachem, Marilyn M Bui, Emmanuel Rios Velazquez, Chintan Parmar, Ralph TH Leijenaar, Benjamin Haibe-Kains, Philippe Lambin, Robert J Gillies, et al. "Defining the biological basis of radiomic phenotypes in lung cancer". In: *Elife*, 6 (2017), e23421.

[7] Andrea Sottoriva, Inmaculada Spiteri, Sara GM Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavare. "Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics". In: *Proceedings of the National Academy of Sciences*, 110 (10) (2013), pp. 4009–4014.

[8] R Fisher, L Pusztai, and C Swanton. "Cancer heterogeneity: implications for targeted therapeutics". In: *British journal of cancer*, 108 (3) (2013), p. 479.

[9] Suzanne Kane. *Introduction to Physics in Modern Medicine*. Jan. 2009. doi: `10.1201/9781420023619`.

[10] Jon Cacicedo, Arturo Navarro, Olga Del Hoyo, Alfonso Gomez-Iturriaga, Filippo Alongi, Jose A Medina, Olgun Elicin, Andrea Skanjeti, Francesco Giammarile, Pedro Bilbao, et al. "Role of fluorine-18 fluorodeoxyglucose PET/CT in head and neck oncology: the point of view of the radiation oncologist". In: *The British journal of radiology*, 89 (1067) (2016), p. 20160217.

[11] Leroy Hood and Stephen H Friend. "Predictive, personalized, preventive, participatory (P4) cancer medicine". In: *Nature reviews Clinical oncology*, 8 (3) (2011), p. 184.

[12] Vishwa Parekh and Michael A Jacobs. "Radiomics: a new application from established techniques". In: *Expert review of precision medicine and drug development*, 1 (2) (2016), pp. 207–226.

[13] R.J. Gillies, A.R. Anderson, R.A. Gatenby, and D.L. Morse. "The biology underlying molecular imaging in oncology: from genome to anatome and back

again". In: *Clinical Radiology*, 65 (7) (2010), pp. 517–521. doi: https://doi.org/10.1016/j.crad.2010.04.005. url: http://www.sciencedirect.com/science/article/pii/S0009926010001820.

[14] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Rene Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nature communications*, 5 (2014), p. 4006.

[15] E-Ryung Choi, Ho Yun Lee, Ji Yun Jeong, Yoon-La Choi, Jhingook Kim, Jungmin Bae, Kyung Soo Lee, and Young Mog Shim. "Quantitative image variables reflect the intratumoral pathologic heterogeneity of lung adenocarcinoma". In: *Oncotarget*, 7 (41) (2016), p. 67302.

[16] Griethuysen, Parmar Fedorov, Aucoin Hosny, Beets–Tan Narayan, Pieper Fillon–Robin, and Aerts. "Computational Radiomics System to Decode the Radiographic Phenotype." In: *Cancer Research*, 72 (21) (Nov. 2017). url: https://doi.org/10.1158/0008-5472.CAN-17-0339.

[17] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. "Machine learning methods for quantitative radiomic biomarkers". In: *Scientific reports*, 5 (2015), p. 13087.

[18] Martin Vallieres, Emily Kay-Rivest, Leo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo JWL Aerts, Nader Khaouam, Phuc Felix Nguyen-Tan, Chang-Shu Wang, Khalil Sultanem, et al. "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer". In: *Scientific reports*, 7 (1) (2017), p. 10117.

[19] Marta Bogowicz, Oliver Riesterer, Kristian Ikenberg, Sonja Stieb, Holger Moch, Gabriela Studer, Matthias Guckenberger, and Stephanie Tanadini-Lang. "Computed tomography radiomics predicts HPV status and local tumor control

after definitive radiochemotherapy in head and neck squamous cell carcinoma”. In: *International Journal of Radiation Oncology* Biology* Physics*, 99 (4) (2017), pp. 921–928.

[20] Kyle Strimbu and Jorge A Tavel. “What are biomarkers?” In: *Current Opinion in HIV and AIDS*, 5 (6) (2010), p. 463.

[21] National Cancer Institute. *NCI Dictionary of Cancer Terms disease-free survival*. 2018. url: `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/disease-free-survival` (visited on 08/02/2019).

[22] Yucheng Zhang, Anastasia Oikonomou, Alexander Wong, Masoom A Haider, and Farzad Khalvati. “Radiomics-based prognosis analysis for non-small cell lung cancer”. In: *Scientific reports*, 7 (2017), p. 46349.

[23] David H Wolpert, William G Macready, et al. “No free lunch theorems for optimization”. In: *IEEE transactions on evolutionary computation*, 1 (1) (1997), pp. 67–82.

[24] R.A. Day. “The origins of the scientific paper: The IMRAD format”. In: *AMWA Journal*, 4 (Jan. 1989), pp. 16–18.

[25] G. S. R. E. Langberg. *BioRad*. `https://github.com/gsel9/biorad`. 2019.

[26] Richard Lowry. “Concepts and applications of inferential statistics”. In: (2014).

[27] Frank Wilcoxon. “Individual comparisons by ranking methods”. In: *Biometrics bulletin*, 1 (6) (1945), pp. 80–83.

[28] Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Vol. 7. Perthes et Besser, 1809.

[29] Samuel Sanford Shapiro and Martin B Wilk. “An analysis of variance test for normality (complete samples)”. In: *Biometrika*, 52 (3/4) (1965), pp. 591–611.

[30] Ralph B D’Agostino. “Transformation to normality of the null distribution of g1”. In: *Biometrika* (1970), pp. 679–681.

[31] Patrick E Shrout and Joseph L Fleiss. “Intraclass correlations: uses in assessing rater reliability.” In: *Psychological bulletin*, 86 (2) (1979), p. 420.

[32] Terry K Koo and Mae Y Li. “A guideline of selecting and reporting intraclass correlation coefficients for reliability research”. In: *Journal of chiropractic medicine*, 15 (2) (2016), pp. 155–163.

[33] Ronald Aylmer Fisher. “Statistical methods for research workers”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 66–70.

[34] Philip Sedgwick. “Spearman’s rank correlation coefficient”. In: *Bmj*, 349 (2014), g7327.

[35] Geoffrey E Hinton, Terrence Joseph Sejnowski, and Tomaso A Poggio. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[36] Anil K Jain. “Data clustering: 50 years beyond K-means”. In: *Pattern recognition letters*, 31 (8) (2010), pp. 651–666.

[37] David Arthur and Sergei Vassilvitskii. “K-means++: the advantages of careful seeding”. In: *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*. 2007.

[38] Inderjit S Dhillon. “Co-clustering documents and words using bipartite spectral graph partitioning”. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2001, pp. 269–274.

[39] Yizong Cheng and George M Church. "Biclustering of expression data." In: *Ismb*. Vol. 8. 2000. 2000, pp. 93–103.

[40] V. Klema and A. Laub. "The singular value decomposition: Its computation and some applications". In: *IEEE Transactions on Automatic Control*, 25 (2) (1980), pp. 164–176.

[41] Beatriz Pontes, Ral Girldez, and Jess S Aguilar-Ruiz. "Quality measures for gene expression biclusters". In: *PloS one*, 10 (3) (2015), e0115497.

[42] Isabelle Guyon and Andre Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research*, 3 (Mar) (2003), pp. 1157–1182.

[43] Yvan Saeys, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics". In: *bioinformatics*, 23 (19) (2007), pp. 2507–2517.

[44] Jiliang Tang, Salem Alelyani, and Huan Liu. "Feature Selection for Classification: A Review." In: *Data Classification: Algorithms and Applications*. Ed. by Charu C. Aggarwal. CRC Press, 2014, pp. 37–64. isbn: 978-1-4665-8674-1. url: http://dblp.uni-trier.de/db/books/collections/aggarwal2014.html#TangAL14.

[45] Sebastian Raschka. *Python machine learning*. Packt Publishing Ltd, 2015.

[46] Karl Pearson. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50 (302) (1900), pp. 157–175.

[47] Brian C Ross. “Mutual information between discrete and continuous data set s”. In: *PloS one*, 9 (2) (2014), e87357.

[48] Alexander Kraskov, Harald Stogbauer, and Peter Grassberger. “Estimating mutual information”. In: *Physical review E*, 69 (6) (2004), p. 066138.

[49] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. ninth Dover printing, tenth GPO printing. Dover, 1964.

[50] Henry B Mann and Donald R Whitney. “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics* (1947), pp. 50–60.

[51] Michael P Fay and Michael A Proschan. “Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules”. In: *Statistics surveys*, 4 (2010), p. 1.

[52] NICHOLAS T. LONGFORD. “A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects”. In: *Biometrika*, 74 (4) (Dec. 1987), pp. 817–827. issn: 0006-3444. doi: `10.1093/biomet/74.4.817`. eprint: `http://oup.prod.sis.lan/biomet/article-pdf/74/4/817/786386/74-4-817.pdf`. url: `https://doi.org/10.1093/biomet/74.4.817`.

[53] 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* (2018).

[54] Ryan J Urbanowicz, Randal S Olson, Peter Schmitt, Melissa Meeker, and Jason H Moore. “Benchmarking relief-based feature selection methods for bioinformatics data mining”. In: *Journal of biomedical informatics*, 85 (2018), pp. 168–188.

[55] Roshan Kumari and Saurabh Srivastava. “Machine Learning: A Review on Binary Classification”. In: *International Journal of Computer Applications*, 160 (Feb. 2017), pp. 11–15. doi: `10.5120/ijca2017913083`.

[56] Peter A Lachenbruch and M Goldstein. “Discriminant analysis”. In: *Biometrics* (1979), pp. 69–85.

[57] Bradley Efron. “Bayes’ theorem in the 21st century”. In: *Science*, 340 (6137) (2013), pp. 1177–1178.

[58] Jerome H Friedman. “Regularized discriminant analysis”. In: *Journal of the American statistical association*, 84 (405) (1989), pp. 165–175.

[59] Olivier Ledoit and Michael Wolf. “Honey, I shrunk the sample covariance matrix”. In: *UPF economics and business working paper*, (691) (2003).

[60] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning*, 20 (3) (1995), pp. 273–297.

[61] R Tyrrell Rockafellar. “Lagrange multipliers and optimality”. In: *SIAM review*, 35 (2) (1993), pp. 183–238.

[62] Isabelle Guyon, B Boser, and Vladimir Vapnik. “Automatic capacity tuning of very large VC-dimension classifiers”. In: *Advances in neural information processing systems*. 1993, pp. 147–155.

[63] Mark A Aizerman. “Theoretical foundations of the potential function method in pattern recognition learning”. In: *Automation and remote control*, 25 (1964), pp. 821–837.

[64] Strother H Walker and David B Duncan. “Estimation of the probability of an event as a function of several independent variables”. In: *Biometrika*, 54 (1-2) (1967), pp. 167–179.

[65] Andrei Nikolaevitch Tikhonov, AV Goncharsky, VV Stepanov, and Anatoly G Yagola. *Numerical methods for the solution of ill-posed problems*. Vol. 328. Springer Science & Business Media, 2013.

[66] George AF Seber and Alan J Lee. *Linear regression analysis*. Vol. 329. John Wiley & Sons, 2012.

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research*, 12 (2011), pp. 2825–2830.

[68] Naomi S Altman. “An introduction to kernel and nearest-neighbor nonparametric regression”. In: *The American Statistician*, 46 (3) (1992), pp. 175–185.

[69] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[70] David Opitz and Richard Maclin. “Popular Ensemble Methods: An Empirical Study”. In: *J. Artif. Int. Res.* 11 (1) (July 1999), pp. 169–198. issn: 1076-9757. url: `http://dl.acm.org/citation.cfm?id=3013545.3013549`.

[71] Leo Breiman. “Random forests”. In: *Machine learning*, 45 (1) (2001), pp. 5–32.

[72] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: *Machine learning*, 63 (1) (2006), pp. 3–42.

[73] Robert E Schapire. “The strength of weak learnability”. In: *Machine learning*, 5 (2) (1990), pp. 197–227.

[74] Jerome H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics*, 29 (2000), pp. 1189–1232.

[75] Leo Breiman. “Bias, variance, and arcing classifiers”. In: (1996).

[76] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[77] 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: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 3146–3154. url: `http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf`.

[78] Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, and Tie-Yan Liu. “A Communication-Efficient Parallel Algorithm for Decision Tree”. In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., 2016, pp. 1279–1287. url: `http://papers.nips.cc/paper/6381-a-communication-efficient-parallel-algorithm-for-decision-tree.pdf`.

[79] Stefan Romberg, Moritz August, Christian X. Ries, and Rainer Lienhart. “Robust Feature Bundling”. In: *In LNCS*. 2012.

[80] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC bioinformatics*, 8 (1) (2007), p. 25.

[81] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. “Consistent individualized feature attribution for tree ensembles”. In: *arXiv preprint arXiv:1802.03888* (2018).

[82] Marc Claesen and Bart De Moor. “Hyperparameter search in machine learning”. In: *arXiv preprint arXiv:1502.02127* (2015).

[83] Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*. Tech. rep. Cambridge University Press, 2008.

[84] Katharina Eggensperger, Matthias Feurer, Frank Hutter, James Bergstra, Jasper Snoek, Holger Hoos, and Kevin Leyton-Brown. “Towards an empirical foundation for assessing bayesian optimization of hyperparameters”. In: *NIPS workshop on Bayesian Optimization in Theory and Practice*. Vol. 10. 2013, p. 3.

[85] James Bergstra and Yoshua Bengio. “Random Search for Hyper-parameter Optimization”. In: *J. Mach. Learn. Res.* 13 (Feb. 2012), pp. 281–305. issn: 1532-4435. url: `http://dl.acm.org/citation.cfm?id=2188385.2188395`.

[86] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. “The application of Bayesian methods for seeking the extremum”. In: *Towards global optimization*, 2 (117-129) (1978), p. 2.

[87] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. “Auto-WEKA: Automated selection and hyper-parameter optimization of classification algorithms”. In: *CoRR, abs/1208.3719* (2012).

[88] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. “Sequential model-based optimization for general algorithm configuration”. In: *International Conference on Learning and Intelligent Optimization*. Springer. 2011, pp. 507–523.

[89] Carl Edward Rasmussen. “Gaussian processes for machine learning”. In: MIT Press, 2006.

[90] James Bergstra, Remi Bardenet, Yoshua Bengio, and Balazs Kegl. "Algorithms for Hyper-parameter Optimization". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Granada, Spain: Curran Associates Inc., 2011, pp. 2546–2554. isbn: 978-1-61839-599-3. url: `http://dl.acm.org/citation.cfm?id=2986459.2986743`.

[91] Donald R. Jones, Matthias Schonlau, and William J. Welch. "Efficient Global Optimization of Expensive Black-Box Functions". In: *J. of Global Optimization*, 13 (4) (Dec. 1998), pp. 455–492. issn: 0925-5001. doi: `10.1023/A:1008306431147`.url:`https://doi.org/10.1023/A:1008306431147`.

[92] Ron Kohavi et al. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. 2. Montreal, Canada. 1995, pp. 1137–1145.

[93] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[94] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics*, 7 (2006), p. 91.

[95] Gavin C. Cawley and Nicola L.C. Talbot. "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation". In: *J. Mach. Learn. Res.* 11 (Aug. 2010), pp. 2079–2107. issn: 1532-4435. url: `http://dl.acm.org/citation.cfm?id=1756006.1859921`.

[96] FK Wang. "Confidence interval for the mean of non-normal data". In: *Quality and Reliability Engineering International*, 17 (4) (2001), pp. 257–267.

[97] G. van Rossum. *Python tutorial*. Tech. rep. CS-R9526. Amsterdam: Centrum voor Wiskunde en Informatica (CWI), May 1995.

[98] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. "Jupyter Notebooks-a publishing format for reproducible computational workflows." In: *ELPUB*. 2016, pp. 87–90.

[99] Thibaud P Coroller, Vishesh Agrawal, Vivek Narayan, Ying Hou, Patrick Grossmann, Stephanie W Lee, Raymond H Mak, and Hugo JWL Aerts. *exampleSettings*. 2015. url: `https://github.com/Radiomics/pyradiomics/tree/master/examples/exampleSettings` (visited on 06/17/2019).

[100] Jon Magne Moan, Cecilie Delphin Amdal, Eirik Malinen, Jørund Graadal Svestad, Trond Velde Bogsrud, and Einar Dale. "The prognostic role of 18F-fluorodeoxyglucose PET in head and neck cancer depends on HPV status". In: *Radiotherapy and Oncology*, 140 (2019), pp. 54–61.

[101] *A Dictionary of Astronomy. The Science of Microfabrication*. 2nd ed. Oxford University Press, 2012.

[102] K. Greenway D. Bell. *Hounsfield unit*. 2019. url: `https://radiopaedia.org/articles/hounsfield-unit` (visited on 08/02/2019).

[103] Ki Yap Daniel J Bell. *Standard uptake value*. 2019. url: `https://radiopaedia.org/articles/standard-uptake-value?lang=us` (visited on 08/02/2019).

[104] PJ Julyan, JH Taylor, DL Hastings, HA Williams, and J Zweit. "SUVpeak: a new parameter for quantification of uptake in FDG PET". In: *Nuclear Medicine Communications*, 25 (4) (2004), p. 407.

[105] Kenneth R Zasadny, Paul V Kison, Isaac R Francis, and Richard L Wahl. "FDG-PET determination of metabolically active tumor volume and comparison with CT". In: *Clinical positron imaging*, 1 (2) (1998), pp. 123–129.

[106] El Naqa, Apte Grigsby, Donnelly Kidd, Chaudhari Khullar, Schmitt Yang, Laforest, and Deasy Thorstad. "Exploring feature-based approaches in PET images for predicting cancer treatment outcomes". English (US). In: *Pattern Recognition*, 42 (6) (June 2009). issn: 0031-3203. doi: `10.1016/j.patcog.2008.08.011`.

[107] Florent Tixier, Catherine Rest, Mathieu Hatt, Nidal M. Albarghach, O Pradier, Jean philippe Metges, Laurent Corcos, and Dimitris Visvikis. "Intratumor Heterogeneity Characterized by Textural Features on Baseline F-18-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer". In: *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 52 (Feb. 2011), pp. 369–78. doi: `10.2967/jnumed.110.082404`.

[108] Alex Zwanenburg, Stefan Leger, Martin Vallieres, Steffen Lock, et al. "Image Biomarker Standardisation Initiative". In: *arXiv preprint arXiv:1612.07003* (2016).

[109] Ralph TH Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter Jc Van Elmpt, Esther GC Troost, Ronald Boellaard, Hugo JWL Aerts, Robert J Gillies, and Philippe Lambin. "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis". In: *Scientific reports*, 5 (2015), p. 11075.

[110] Robert M Haralick, Karthikeyan Shanmugam, et al. "Textural features for image classification". In: *IEEE Transactions on systems, man, and cybernetics*, (6) (1973), pp. 610–621.

[111] Mathieu Hatt, Florent Tixier, Larry Pierce, Paul Kinahan, Catherine Cheze Le Rest, and Dimitris Visvikis. "Characterization of PET/CT images using texture analysis: the past, the present... any future?" In: *European Journal of Nuclear Medicine and Molecular Imaging*, 44 (June 2016). doi: `10.1007/s00259-016-3427-0`.

[112] Griethuysen, Parmar Fedorov, Aucoin Hosny, Beets–Tan Narayan, Pieper Fillon–Robin, and Aerts. *Welcome to pyradiomics documentation!* 2016. url: `https://pyradiomics.readthedocs.io/en/latest/index.html` (visited on 07/15/2019).

[113] Guillaume Thibault, Bernard FERTIL, Claire Navarro, Sandrine Pereira, Nicolas Levy, Jean SEQUEIRA, and Jean-Luc MARI. “Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification”. In: Nov. 2009.

[114] Dong-Hui Xu, Arati S Kurani, Jacob D Furst, and Daniela S Raicu. “Run-length encoding for volumetric texture”. In: *Heart*, 27 (25) (2004), pp. 452–458.

[115] Rachel B Ger, Daniel F Craft, Dennis S Mackin, Shouhao Zhou, Rick R Layman, A Kyle Jones, Hesham Elhalawani, Clifton D Fuller, Rebecca M Howell, Heng Li, et al. “Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis”. In: *Computerized Medical Imaging and Graphics*, 69 (2018), pp. 134–139.

[116] C Bonferroni. “Teoria statistica delle classi e calcolo delle probabilita”. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8 (1936), pp. 3–62.

[117] Chintan Parmar, Joseph D. Barry, Ahmed Hosny, John Quackenbush, and Hugo J.W.L. Aerts. “Data Analysis Strategies in Medical Imaging”. In: *Clinical Cancer Research*, 24 (15) (2018), pp. 3492–3499. issn: 1078-0432. doi: `10.1158/1078-0432.CCR-18-0385`. eprint: `http://clincancerres.aacrjournals.org/content/24/15/3492.full.pdf`. url: `http://clincancerres.aacrjournals.org/content/24/15/3492`.

[118] Robert Tibshirani. “The lasso method for variable selection in the Cox model”. In: *Statistics in medicine*, 16 (4) (1997), pp. 385–395.

[119] M Hassan, Kujtim Latifi, Geoffrey Zhang, Ghanim Ullah, Robert Gillies, and Eduardo Moros. “Voxel size and gray level normalization of CT radiomic features in lung cancer”. In: *Scientific Reports*, 8 (July 2018). doi: `10.1038/s41598-018-28895-9`.

[120] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern recognition letters*, 27 (8) (2006), pp. 861–874.

[121] Parnian Afshar, Arash Mohammadi, Konstantinos N Plataniotis, Anastasia Oikonomou, and Habib Benali. “From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities”. In: *arXiv preprint arXiv:1808.07954* (2018).

[122] Weimiao Wu, Chintan Parmar, Patrick Grossmann, John Quackenbush, Philippe Lambin, Johan Bussink, Raymond Mak, and Hugo JWL Aerts. “Exploratory study to identify radiomics classifiers for lung cancer histology”. In: *Frontiers in oncology*, 6 (2016), p. 71.

[123] M. Lindauer, K. Eggensperger M. Feurer, A. Biedenkapp J. Marben, S. Falkner A. Klein, and F. Hutter. *SMAC Documentation*. 2019. url: `https://automl.github.io/SMAC3/master/installation` (visited on 08/02/2019).

[124] Payam Refaeilzadeh, Lei Tang, and Huan Liu. “On comparison of feature selection algorithms”. In: *Proceedings of AAAI workshop on evaluation methods for machine learning II*. Vol. 3. 4. 2007, p. 5.

[125] Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. “Cross-validation pitfalls when selecting and assessing regression and classification models”. In: *Journal of cheminformatics*, 6 (1) (2014), p. 10.

[126] Ioannis Tsamardinos, Amin Rakhshani, and Vincenzo Lagani. “Performance-estimation properties of cross-validation-based protocols with simultaneous

hyper-parameter optimization". In: *International Journal on Artificial Intelligence Tools*, 24 (05) (2015), p. 1540023.

[127] Ranjbar, Zwart Ning, Weindling Wood, Mitchell Wu, and Hoxworth JM Li. "Computed Tomography-Based Texture Analysis to Determine Human Papillomavirus Status of Oropharyngeal Squamous Cell Carcinoma". In: *Comput Assist Tomogr*, 42 (2) (2018). issn: 2219-679X.

[128] Andrew J. Wong, Aasheesh Kanwar, Abdallah S. Mohamed, and Clifton D. Fuller. "Radiomics in head and neck cancer: from exploration to application". In: *Translational Cancer Research*, 5 (4) (2016), pp. 24–35. issn: 2219-6803. doi: 10.1056/NEJMoa0912217. url: http://tcr.amegroups.com/article/view/8805.

[129] Carole Fakhry, William Westra, Sigui Li, Anthony Cmelak, John Ridge, Harlan Pinto, Arlene Forastiere, and Maura Gillison. "Improved Survival of Patients With Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma in a Prospective Clinical Trial". In: *Journal of the National Cancer Institute*, 100 (4) (Feb. 2008), pp. 261–269. issn: 0027-8874. doi: 10.1093/jnci/djn011. url: https://doi.org/10.1093/jnci/djn011.

[130] K Kian Ang, Jonathan Harris, Richard Wheeler, Randal Weber, David I Rosenthal, P Nguyen-Tan, William H Westra, Christine H Chung, Richard Jordan, Charles Lu, Harold Kim, Rita Axelrod, C Craig Silverman, Kevin Redmond, and Maura L Gillison. "Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer". In: *The New England journal of medicine*, 363 (July 2010), pp. 24–35. doi: 10.1056/NEJMoa0912217.

[131] Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. "Prediction error estimation: a comparison of resampling methods". In: *Bioinformatics*, 21 (15) (2005), pp. 3301–3307.

[132] Jacques Wainer and Gavin Cawley. “Nested cross-validation when selecting classifiers is overzealous for most practical applications”. In: *arXiv preprint arXiv:1809.09446* (2018).

[133] Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles. *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May, 15-17, 2014, Proceedings*. Vol. 8445. Springer, 2014.

[134] Provost and Weiss. “Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction”. In: *CoRR*, abs/1106.4557 (2011). arXiv: `1106.4557`. url: `http://arxiv.org/abs/1106.4557`.

[135] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical bayesian optimization of machine learning algorithms”. In: *Advances in neural information processing systems*. 2012, pp. 2951–2959.

[136] Katharina Eggensperger, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. “Surrogate Benchmarks for Hyperparameter Optimization.” In: *MetaSel@ ECAI*. 2014, pp. 24–31.

[137] Hesham Elhalawani, Aasheesh Kanwar, Abdallah S.R. Mohamed, Aubrey White, James Zafereo, Andrew Wong, Joel Berends, Shady Abohashem, Bowman Williams, Jeremy M. Aymard, Subha Perni, Jay Messer, Ben Warren, Bassem Youssef, Pei Yang, Mohamed A.M. Meheissen, Mona Kamal, Baher Elgohari, Rachel B. Ger, Carlos E. Cardenas, Xenia Fave, Lifei Zhang, Dennis Mackin, G. Elisabeta Marai, David M. Vock, Guadalupe M. Canahuate, Stephen Y. Lai, G. Brandon Gunn, Adam S. Garden, David I. Rosenthal, Laurence Court, and Clifton D. Fuller. “Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients”. English (US). In: *Scientific Reports*, 8 (1) (Dec. 2018). issn: 2045-2322. doi: `10.1038/s41598-017-14687-0`.

[138] Ralph TH Leijenaar, Marta Bogowicz, Arthur Jochems, Frank JP Hoebers, Frederik WR Wesseling, Sophie H Huang, Biu Chan, John N Waldron, Brian O'Sullivan, Derek Rietveld, C Rene Leemans, Ruud H Brakenhoff, Oliver Riesterer, Stephanie Tanadini-Lang, Matthias Guckenberger, Kristian Ikenberg, and Philippe Lambin. "Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study". In: *The British Journal of Radiology*, 91 (1086) (2018). PMID: 29451412, p. 20170498. doi: `10.1259/bjr.20170498`. eprint: `https://doi.org/10.1259/bjr.20170498`. url: `https://doi.org/10.1259/bjr.20170498`.

[139] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." In: *Biometrics*, 44 (3) (1988), pp. 837–845.

[140] Gary JR Cook, Connie Yip, Muhammad Siddique, Vicky Goh, Sugama Chicklore, Arunabha Roy, Paul Marsden, Shahreen Ahmad, and David Landau. "Are pretreatment 18F-FDG PET tumor textural features in non–small cell lung cancer associated with response and survival after chemoradiotherapy?" In: *Journal of nuclear medicine*, 54 (1) (2013), pp. 19–26.

[141] David V Fried, Susan L Tucker, Shouhao Zhou, Zhongxing Liao, Osama Mawlawi, Geoffrey Ibbott, and Laurence E Court. "Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer". In: *International Journal of Radiation Oncology* Biology* Physics*, 90 (4) (2014), pp. 834–842.

[142] Richard E Bellman. *Adaptive control processes: a guided tour*. Vol. 2045. Princeton university press, 2015.

[143] Jolinta Lin, Seth Kligerman, Rakhi Goel, Payam Sajedi, Mohan Suntharalingam, and Michael D. Chuong. "State-of-the-art molecular imaging in esophageal cancer management: implications for diagnosis, prognosis, and treatment".

In: *Journal of Gastrointestinal Oncology*, 6 (1) (2014). issn: 2219-679X. url: http://jgo.amegroups.com/article/view/2909.

[144] John Cuaron, Mark Dunphy, and Andreas Rimner. “Role of FDG-PET scans in staging, response assessment, and follow-up care for non-small cell lung cancer”. In: *Frontiers in Oncology*, 2 (2013), p. 208. issn: 2234-943X. doi: 10.3389/fonc.2012.00208. url: https://www.frontiersin.org/article/10.3389/fonc.2012.00208.

[145] Floris H. P. van Velden, Patsuree Cheebsumon, Maqsood Yaqub, Egbert F. Smit, Otto S. Hoekstra, Adriaan A. Lammertsma, and Ronald Boellaard. “Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies”. In: *European Journal of Nuclear Medicine and Molecular Imaging*, 38 (9) (Oct. 2011), pp. 1636–1647. issn: 1619-7089. doi: 10.1007/s00259-011-1845-6. url: https://doi.org/10.1007/s00259-011-1845-6.

[146] K Buch, A Fujita, B Li, Y Kawashima, MM Qureshi, and O Sakai. “Using texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinomas on CT”. In: *American Journal of Neuroradiology*, 36 (7) (2015), pp. 1343–1348.

[147] Akifumi Fujita, Karen Buch, Baojun Li, Yusuke Kawashima, Muhammad M Qureshi, and Osamu Sakai. “Difference between HPV-positive and HPV-negative non-oropharyngeal head and neck cancer: texture analysis features on CT”. In: *Journal of computer assisted tomography*, 40 (1) (2016), pp. 43–47.

[148] David Martin Powers. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. In: (2011).

[149] Peter Bühlmann, Torsten Hothorn, et al. “Boosting algorithms: Regularization, prediction and model fitting”. In: *Statistical Science*, 22 (4) (2007), pp. 477–505.

[150] Thibaud P Coroller, Vishesh Agrawal, Vivek Narayan, Ying Hou, Patrick Grossmann, Stephanie W Lee, Raymond H Mak, and Hugo JWL Aerts. "Radiomic phenotype features predict pathological response in non-small cell lung cancer". In: *Radiotherapy and Oncology*, 119 (3) (2016), pp. 480–486.

[151] Rajat Thawani, Michael McLane, Niha Beig, Soumya Ghose, Prateek Prasanna, Vamsidhar Velcheti, and Anant Madabhushi. "Radiomics and radiogenomics in lung cancer: a review for the clinician". In: *Lung Cancer*, 115 (2018), pp. 34–41.

[152] K. Smith, S. Montanaro J. Jewett, and A. Baxter. *PEP 318 – Decorators for Functions and Methods*. 2019. url: `https://www.python.org/dev/peps/pep-0318/` (visited on 08/04/2019).

[153] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gael Varoquaux. "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[154] Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation". In: *Machine Learning*, 107 (12) (Dec. 2018), pp. 1895–1922. issn: 1573-0565. doi: `10.1007/s10994-018-5714-4`. url: `https://doi.org/10.1007/s10994-018-5714-4`.

[155] Elena Menichelli, Trygve Almoy, Oliver Tomic, Nina Veflen Olsen, and Tormod Naes. "SO-PLS as an exploratory tool for path modelling". In: *Food quality and preference*, 36 (2014), pp. 122–134.

[156] Kristian Hovde Liland, Tormod Naes, and Ulf G Indahl. "ROSA-a fast extension of partial least squares regression for multiblock data analysis". In: *Journal of Chemometrics*, 30 (11) (2016), pp. 651–662.

[157] Mattea L Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G Purdie, Brian O'Sullivan, Hugo JWL Aerts, et al. "Vulnerabilities of radiomic signature development: The need for safeguards". In: *Radiotherapy and Oncology*, 130 (2019), pp. 2–9.

[158] David A Belsley. "A guide to using the collinearity diagnostics". In: *Computer Science in Economics and Management*, 4 (1) (1991), pp. 33–50.

[159] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. [Online; accessed <today>]. 2001. url: `http://www.scipy.org/`.

[160] Travis Oliphant. *NumPy: A guide to NumPy*. USA: Trelgol Publishing. [Online; accessed <today>]. 2006. url: `http://www.numpy.org/`.

Appendix A

Feature Exploration

A.1 Clinical Factors

Table A.1 summarizes the characteristics of the 198 patients included in the data set.

Table A.1: A summary of pre-treatment and tumor characteristics referred to as *clinical factors* of the patient cohort.

Factor		Description
Total number of patients		198
Age (years)		60 (40, 80)$^{(1)}$
Gender	Male	50 (25 %)
	Female	148 (74 %)
Tumour stage	T1/T2	96.0 (48 %
	T3/T4	102 (52 %)
Pack years		22 (0, 128)*
Naxogin (days)		39 (0, 45)*
Cisplatin (treatments)	0	44 (22 %)

Table A.1: A summary of pre-treatment and tumor characteristics referred to as *clinical factors* of the patient cohort.

Factor		Description
	1-3	19 (10 %)
	4-6	135 (68)
Stage	0	1 (0 %)
	I	2 (1 %)
	II	17 (8 %)
	III	39 (19 %)
	IV	138 (69 %)
Degree of spread	N0	121 (61 %)
	N1	47 (24)
	N2	23 (12 %)
	N3	
Tumour site	Oral cavity	17.0 (9 %)
	Oropharynx	144 (73 %)
	Hypopharynx	16.0 (8 %)
	Larynx	21.0 (10 %)
Tumour volume (cm ^{3})		14.7 (0.800, 285)(1)
HPV status	Positive	83 (42 %)
	Negative	18 (9 %)
	Unknown	97 (49 %)

Table A.1: A summary of pre-treatment and tumor characteristics referred to as *clinical factors* of the patient cohort.

Factor		Description
ICD-10	C01	34 (17 %)
	C02	8 (4.04 %)
	C03	1 (0 %)
	C04	3 (1 %)
	C05	4 (2 %)
	C06	1 (0 %)
	C09	74 (37 %)
	C10	36 (18 %)
	C12	6 (3 %)
	C13	10 (5 %)
	C32	21 (11 %)
Histology	0	138 (70 %)
	1	50 (26 %)
	2	9 (5 %)
	3	1 (0 %)
ECOG performance status	0	128 (65 %)
	1	65 (33 %)
	2	5 (2 %)
Charlson Comorbidity Index	0	130 (66 %)

Table A.1: A summary of pre-treatment and tumor characteristics referred to as *clinical factors* of the patient cohort.

Factor	Description
1	45 (23 %)
2	15 (8 %)
3	4 (2 %)
4	3 (2 %)
5	1 (0 %)

(1): median, (minimum, maximum)

A.2 The Distribution of Radiomics Features

Figure A.1 shows the distribution of shape, PET and CT radiomics features per patient. Each distinct colour in Figure A.1 represents a different feature. However, the aim with Figure A.1, is to visualise the distributions of feature values to identify observations deviating from the mean behavior.

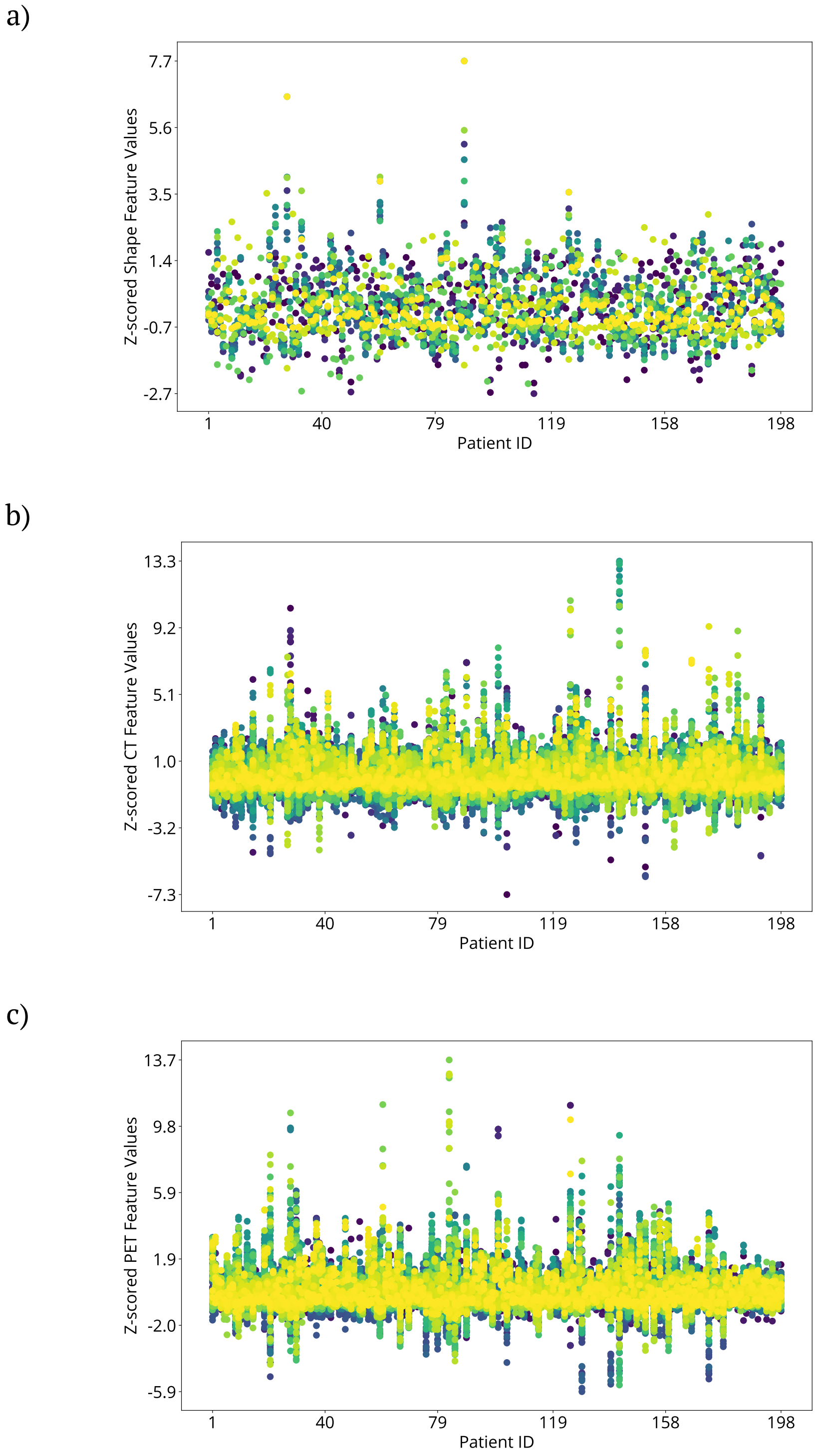


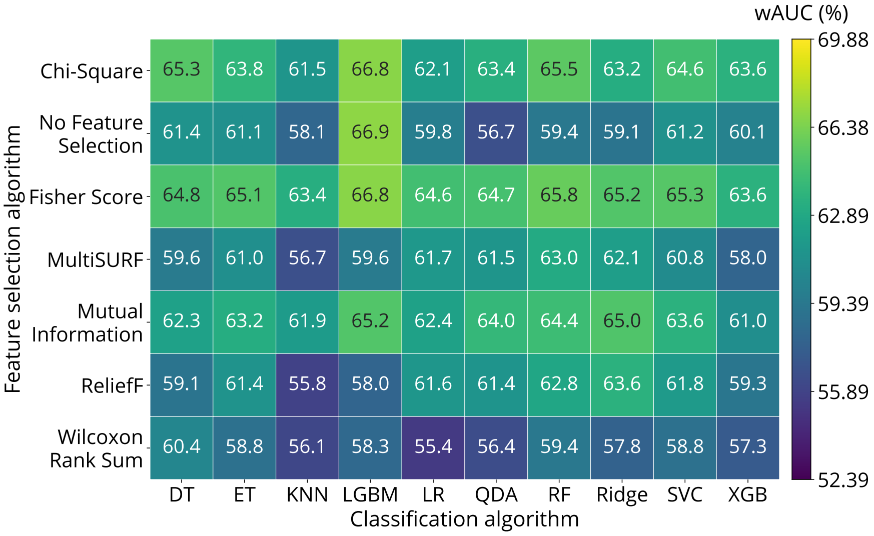
Figure A.1: Scatter plots of radiomics feature values (vertical axis) for each of the 198 patients (horizontal axis). Each color label represents a different feature.

Appendix B

Classification Experiments

B.1 Removing Intra-Feature Correlations

Figure B.1 shows the result of classifying disease-free survival after removal of intra-correlated features (Experiment 4, Section 3.8.3). The no information rate (Section 3.10) in the classified cohort was 67 % disease-free survival.



	DT	ET	KNN	LGBM	LR	QDA	RF	Ridge	SVC	XGB
Chi-Square	65.3	63.8	61.5	66.8	62.1	63.4	65.5	63.2	64.6	63.6
No Feature Selection	61.4	61.1	58.1	66.9	59.8	56.7	59.4	59.1	61.2	60.1
Fisher Score	64.8	65.1	63.4	66.8	64.6	64.7	65.8	65.2	65.3	63.6
MultiSURF	59.6	61.0	56.7	59.6	61.7	61.5	63.0	62.1	60.8	58.0
Mutual Information	62.3	63.2	61.9	65.2	62.4	64.0	64.4	65.0	63.6	61.0
ReliefF	59.1	61.4	55.8	58.0	61.6	61.4	62.8	63.6	61.8	59.3
Wilcoxon Rank Sum	60.4	58.8	56.1	58.3	55.4	56.4	59.4	57.8	58.8	57.3

Figure B.1: Average wAUC (%) from including features retained after removal of intra-feature correlations (Section 3.8.3) to classify disease-free survival with combinations of feature selection (vertical axis) and classification (horizontal axis) algorithms. The wAUC (%) score was averaged over 40 repeated experiments. The colour bar shows that a higher score corresponds to more correct classifications.

B.2 Removing Both Intra- and Inter-Feature Correlations

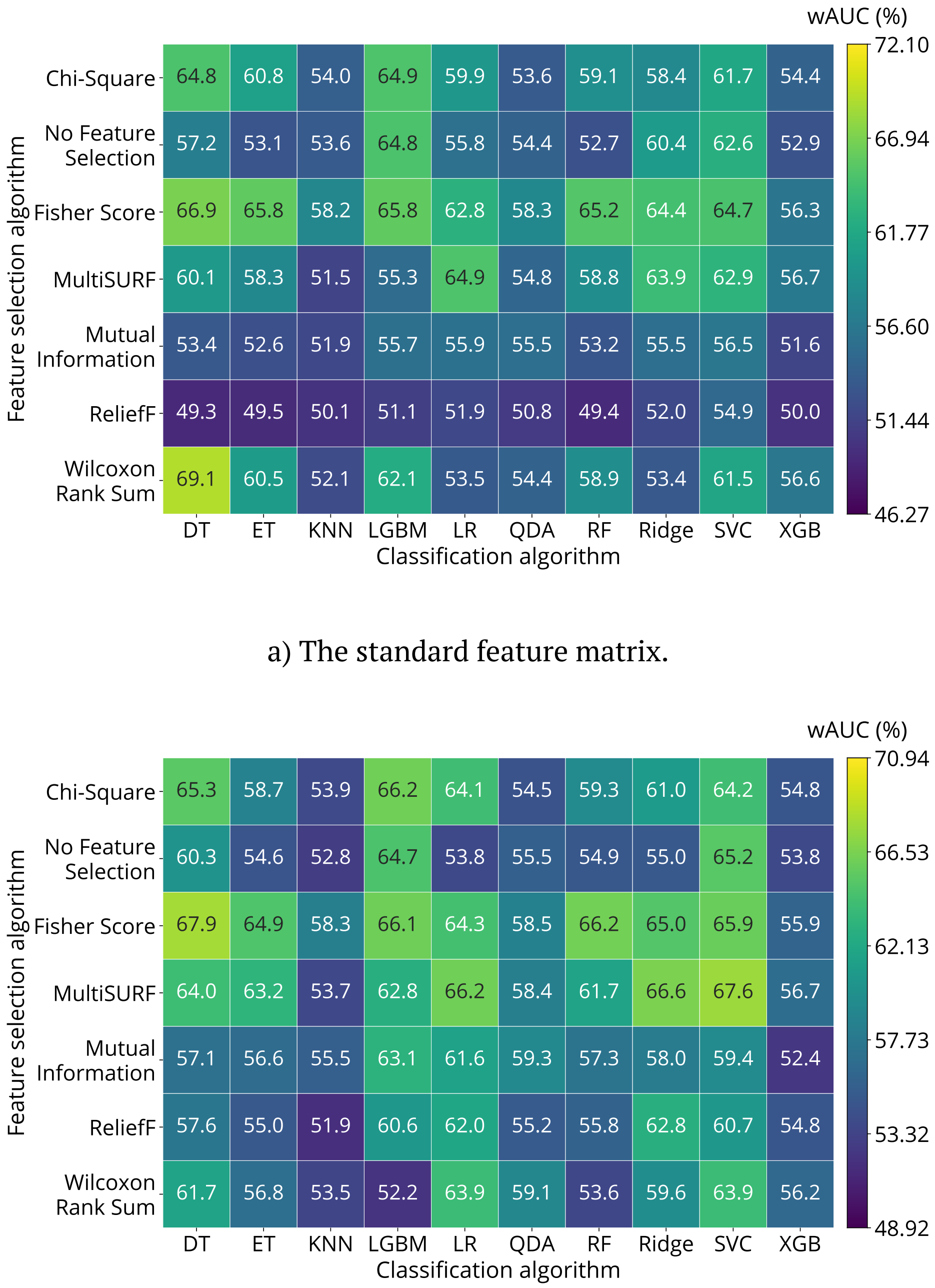
Table B.1 shows the average hyperparameter configuration selected for the model combining Fisher Score feature selection and LGBM classification in Experiment 5 (Section 3.8.3). This classification experiment included the standard feature matrix (Section 3.5) subjected to filtering and removal intra- and inter-correlated features, as described in Sections 3.8.3 and 3.8.3.

Table B.1: The mean hyper-parameter configuration for the model combining Fisher Score feature selection and LGBM classification calculated from configurations selected in the classification experiment based on the standard feature matrix subjected to filtering and removal intra- and inter-correlated features (Experiment 5, Section 3.8.3)

Hyper-Parameter	Mean
Learning Rate	8.50
Maximum tree depth	165
Minimum samples in leaves	4
Number of trees	101
α-regularization	11.2
λ-regularization	49.3

B.3 Classifying Disease-Free Survival in the HPV Related Cohort

Figure B.2 shows the results of classifying patients in the HPV related cohort (Experiments 8 % 9, Section 3.8.3). The no information rate in the HPV related experiments was 73 % disease-free survival.



a) The standard feature matrix.

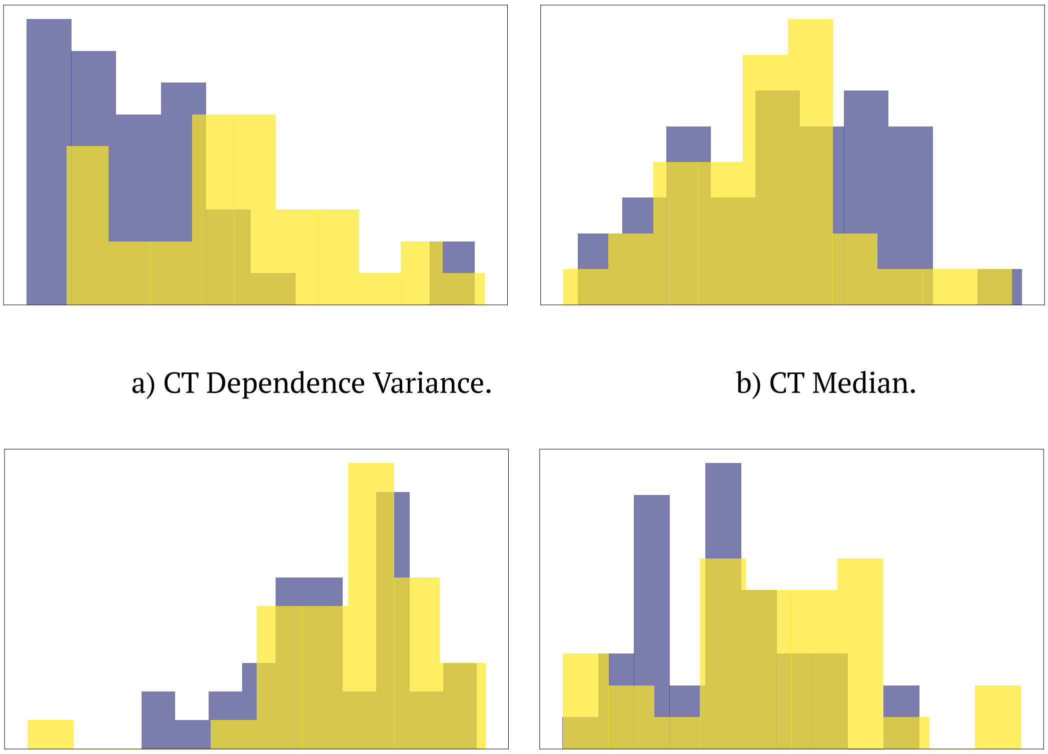
b) Removal of intra- and inter-correlated features.

Figure B.2: Average wAUC (%) from including (a) standard features, and (b) features from removal of intra- and inter-correlations (Section 3.8.3) to classify disease-free survival in the HPV related cohort with combinations of feature selection (vertical axis) and classification (horizontal axis) algorithms. The colour bar shows that a higher score corresponds to more correct classifications.

Appendix C

Potential Biomarkers

Figure C.1 shows the distribution of each feature selected as potential biomarkers grouped and hued according to clinical endpoints in the HPV unrelated cohort.



a) CT Dependence Variance.

b) CT Median.

a) CT LDHGLE 64 bins.

b) Major Axis Length.

Figure C.1: The distribution of features selected as potential biomarkers grouped according to clinical endpoint for patients in the HPV unrelated cohort. Purple histograms signifies disease-free survival, whereas yellow represents alternative outcomes. Abbreviation: *Large Dependence Low Gray Level Emphasis* (LDHGLE).

Figure C.2 shows the relationships between the features selected as potential biomarkers grouped according to clinical endpoints in the HPV unrelated cohort.

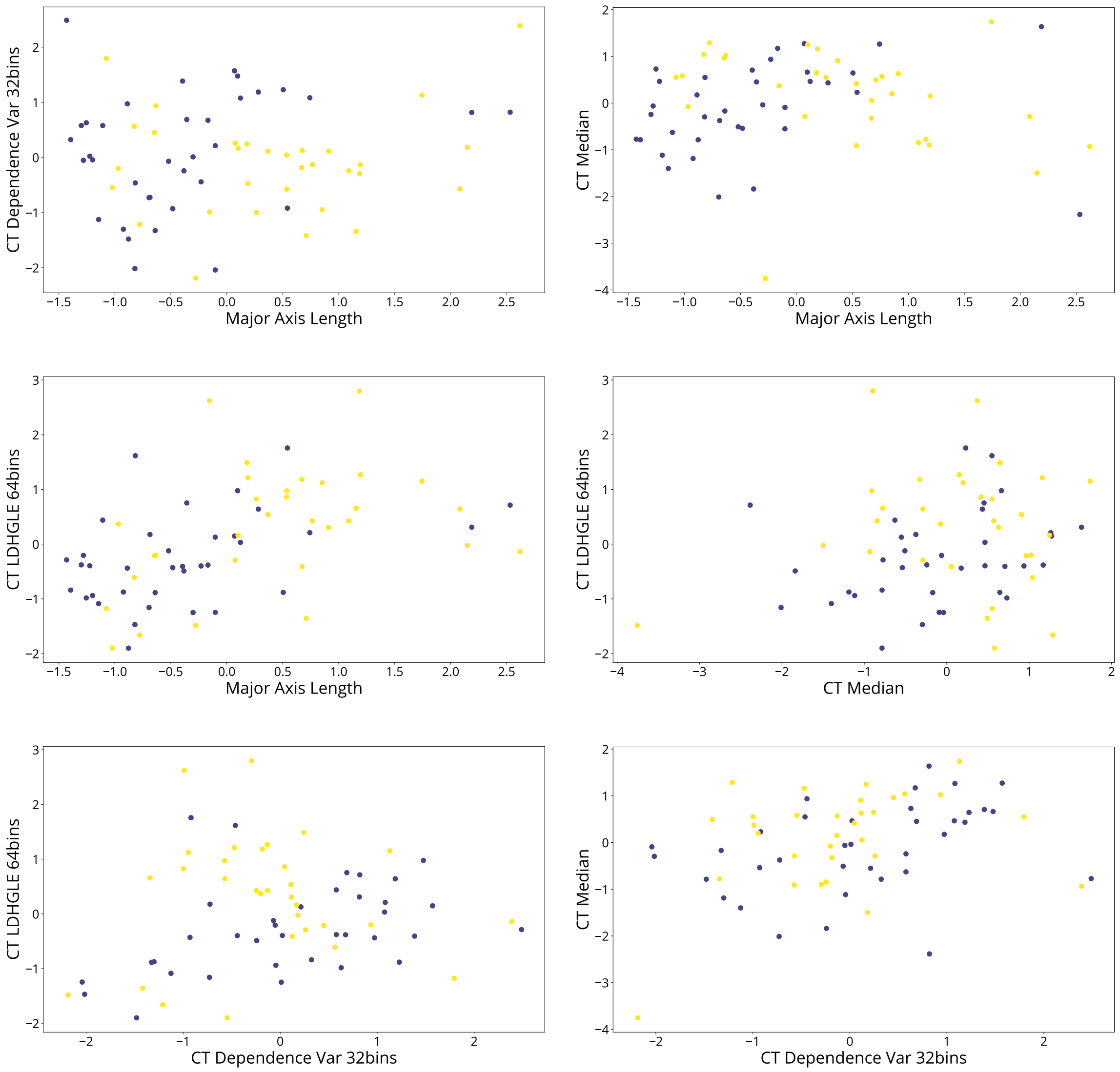


Figure C.2: The relationships between features selected as potential biomarkers grouped according to clinical endpoints in the HPV unrelated cohort. Purple data points signifies disease-free survival, whereas yellow data points represents alternative outcomes. Abbreviation: *Large Dependence Low Gray Level Emphasis* (LDHGLE).

Appendix D

Performing a Classification Experiment

The code used to perform model comparison simulations in this thesis is available from the `https://github.com/gsel9/biorad` in a folder named *bioard* [25]. Scripts including the relevant code for such experiments are contained in the *experiments* folder. An experimental setup is available in the *main.py* file. The file *comparison_schemes.py* contains an implementation of nested stratified cross-validation (CV) [94] with SMAC [136] hyper-parameter optimization. Parallel execution of classification experiment was managed with the *joblib* package [159]. The following code sections illustrate a model comparison experiment.

Code Section D.1 shows a declaration of experimental parameters.

```python
MAX_EVALS = 80 # The number (integer) of objective evaluations with SMAC.
CV = 10 # The number of folds in the nested cross-validation protocol.
NUM_REPS = 40 # The number of experimental repeats.
```

Code Section D.1: Configure experiment parameters.

In Code Section D.1, the parameter *MAX_EVALS* specifies the upper number of objective function evaluations performed by SMAC. This parameter corresponds to the variable T in Algorithm 1. Moreover, *NUM_REPS* specifies the number of random seed values to sample, shown in Code Listing D.2.

```python
import numpy as np

# To ensure reproducibility of randomly generated numbers.
```

```
np.random.seed(seed=0)
random_states = np.random.choice(1000, size=NUM_REPS)
```

Code Section D.2: Collect random seed values.

The the *NUM_REPS* parameter use to sample random seed values, as shown in Code Listing D.2, also corresponds to the number of times the nested CV procedure is repeated.

Code Section D.3 illustrates an example of how the objective for the classification experiment can be defined.

```
from sklearn.metrics import roc_auc_score

def balanced_roc_auc(y_true, y_pred):
    """Calculate the weighted AUC optimization metric."""
    return roc_auc_score(y_true, y_pred, average=’weighted’)
```

Code Section D.3: Optmisation objective.

Note that the function in Code Section D.3 accepts two arguments. These arguments corresponds to the ground truths and the predicted values. Moreover, the function in Code Section D.3 specifies the method for quantifying the performance of models in a classification experiment.

Further, the algorithms currently implemented in biorad [25] are based on the *scikit-learn* API [67]. These algorithms combines each scikit-learn implementation with a hyper-parameter domain, included in the *SMAC* package [88]. To prepare a set of models, consisting of combination of feature selection and classification algorithms, Code Section D.4 illustrates the required format.

```
from algorithms import feature_selection
from algorithms import classification

# Instantiate classification algorithms.
estimators = [
    classification.QuadraticDiscriminantEstimator(),
    classification.SVCEstimator(),
]

# Instantiate feature selection algorithms.
```

```python
selectors = [
    feature_selection.WilcoxonSelection(),
    feature_selection.MultiSURFSelection(),
]
```

Code Section D.4: Configure experiment models.

As shown in Code Section D.4, classification and feature selection algorithms are imported and organized separately in two lists. A function that combines these algorithms into a format that is compatible with a scikit-learn *Pipeline* object is available in the *main.py* file.

A call to the function shown in Code Listing D.5 executes the model comparison experiment.

```python
from model_comparison import model_comparison_fn

model_comparison_fn(
    experiments=prep_pipeline(estimators, selectors),
    path_final_results='results_experiment1.csv',
    output_dir='parameter_search',
    random_states=random_states,
    score_func=balanced_roc_auc,
    max_evals=MAX_EVALS,
    cv=CV,
   X=X,
    y=y
)
```

Code Section D.5: Simulation example.

The first parameter, *experiments*, given to the function in Code Listing D.5 is the function included in *main.py* that handles the formatting of the classification and feature selection algorithms from Code Listing D.4. Furthermore, *path_final_results* and *output_dir* are references to the experimental results and a folder containing the outputs produced by SMAC. The *random_states*, *score_func* and *max_evals* parameters are defined in Code Listings D.2, D.3 and D.1, respectively. The data matrix, $\mathbf{X} \in \mathbb{R}^{n \times p}$, containing n observations and p features have to be in the format of a *NumPy* [160] array. Moreover, the vector of ground truths, $\mathbf{y} \in \mathbb{R}^{n}$, is required to be of the same format.

A folder is created once the experimented is initiated to store temporary results. If the experiment is aborted, the experiment can be continued from the last store temporary results. Once the experiment is complete, the folder with the temporary results is removed, and the results are contained in a single file.

Appendix E

Definitions of Radiomics Features

The following feature definitions are derived from the *PyRadiomics*, version 2.1.2, *Python*$^{\text{TM}}$ package [16], [97].

The tumor volume is referred to as the ROI.

E.1 First-Order Features

Let

- X be a set of N_p voxels contained in the ROI ,
- $P(i)$ denote a histogram of N_g unique intensity values,
- and $p(i) = P(i)/N_p$ be the histogram scaled by the number voxels in the ROI

Table E.1: Definitions of extracted the first-order features.

Description	**Definition**
Energy: Measures the magnitude of image intensities. The arbitrary constant $c \in \mathbb{R}$ prevents negative intensities.	$\sum_{i=1}^{N_p}(X(i)+c)^2$

Total Energy: Scales the *Energy* feature value by the voxel volume in mm^3.	$V \cdot \sum_{i=1}^{N_p}(X(i)+c)^2$
Entropy: Quantifies randomness in image values. An arbitrary small parameter $\epsilon \in \mathbb{R}$ prevents $\log 0$.	$-\sum_{i=1}^{N_p} p(i) \log_2(p(i)+\epsilon)$
Minimum: The global image minimum intensity.	$\min(X)$
10th Percentile: The 10th percentile of image intensity values.	
90th Percentile: The 90th percentile of image intensity values.	
Maximum: The global image maximum intensity.	$\max(X)$
Mean: The average image value.	$\frac{1}{N_p}\sum_{i=1}^{N_p} X(i)$
Median: The image median value.	
Interquartile Range:	

The difference between the 75th and 25th percentiles.	
Range: The difference between the maximum and minimum intensity values.	$\max(X) - \min(X)$
Mean Absolute Deviation: The mean distance between image values and the image mean, $\bar{X}$.	$\frac{1}{N_p}\sum_{i=1}^{N_p} \lvert X(i) - \overline{X} \rvert$
Robust Mean Absolute Deviation: The mean distance between image values and the intensities ranging from the 10^{10} to the 90th percentiles.	$\frac{1}{N_{10-90}}\sum_{i=1}^{N_{10-90}} \lvert X_{10-90}(i) - \overline{X}_{10-90} \rvert$
Root Mean Squared: The square–root of averaged squared image values as a measure of image value magnitude. The arbitrary constant $c \in \mathbb{R}$ prevents negative intensities.	$\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(X(i) + c)^2}$
Skewness: Quantifies asymmetry in the intensity distribution about the mean image value.	$\frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(X(i)-\overline{X})^3}{(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(X(i)-\overline{X})^2})^3}$
Kurtosis:	$\frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(X(i)-\overline{X})^4}{(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(X(i)-\overline{X})^2})^2}$

Et mål på hvor spiss distribusjonen av intensitetsverdier er. Høye kurtosisverdier tilsier at distribusjonen er konsentrert mer rundt "halene" til et histogram enn rundt $\bar{X}$.	
Uniformity: Quantifies the sum of the squares of each intensity value representing homogeneity of the image array. Greater uniformity implies a greater homogeneity of intensity values.	$\sum_{i=1}^{N_g} p(i)^2$

E.2 Shape Features

Consider a triangle *mesh* with vertices positioned midway on edges between voxels. Each triangle is defined by three adjacent vertices, and share edges by exactly one other triangle.

- N_v be the number of voxels in the ROI,
- N_f represent the number of triangles defining the *mesh*,
- V be the mesh volume measured in mm^3,
- and A be the mesh surface area measured in mm^2.

Table E.2: Definitions of extracted shape features.

Feature	**Ligning**
Mesh Volume:	$\sum_{i=1}^{N_f} \frac{V_{X_i}\cdot(V_{Y_i}\times V_{Z_i})}{6}$

Volume calculated with a triangle mesh, where each face, i, in the mesh is defined by the vertices V_{X_i} V_{Y_i} V_{Z_i}.	
Voxel Volume: Volume calculated as the number of voxels in the ROI multiplied by the size of each voxel, V_i.	$\sum_{i=1}^{N_v} V_i$
Surface Area: Calculates the ROI surface area using a triangle mesh.	$\frac{1}{2}\sum_{i=1}^{N_f} \lvert V_{Z_i}V_{Y_i} \times V_{Z_i}V_{X_i} \rvert$
Surface Area to Volume Ratio: Scales the surface area, A, by the ROI mesh volume, V.	$\frac{A}{V}$
Sphericity: Measures the roundness of the ROI shape relative to a sphere.	$\frac{\sqrt[3]{3\pi V^2}}{A}$
Maximum Three-Dimensional Diameter: The largest pairwise Euclidean distance between ROI surface mesh vertices.	
Maximum Two-Dimensional Diameter:	

The largest pairwise Euclidean distance between ROI surface mesh vertices along the axial, coronal or sagittal plans.	
Major Axis Length: The length of the largest axis derived from *Principal Component Analysis*. applied to the ROI coordinates.	$4\sqrt{\lambda_M}$
Minor Axis Length: The length of the second-largest axis derived from *Principal Component Analysis*. applied to the ROI coordinates.	$4\sqrt{\lambda_m}$
Least Axis Length: The length of the smallest axis derived from *Principal Component Analysis*. applied to the ROI coordinates.	$4\sqrt{\lambda_l}$
Elongation: The ratio between the two largest principal components derived from the ROI coordinates.	$\sqrt{\frac{\lambda_m}{\lambda_M}}$
Flatness The ratio between the smallest and the largest principal components derived from the ROI coordinates.	$\sqrt{\frac{\lambda_l}{\lambda_M}}$

E.3 Gray Level Co-Occurrence Matrix Features

Let

- $\epsilon \in \mathbb{R}$ be an arbitrary small number,
- P(i,j) denote the *Gray Level Co-Occurrence Matrix* (GLCM) with a neighbourhood radius δ along an angle θ,
- $p(i, j) = p(i, j)/\sum P(i, j)$ be the GLCM scaled by the number of elements in the GLCM,
- N_g be the number of image intensity values,
- $p_x(i) = \sum_{j=1}^{N_g} p(i, j)$ and $p_y(j) = \sum_{j=1}^{N_g} p(i, j)$ represent the marginal row and column probabilities, respectively,
- μ_x and μ_y be the mean value p_x, and p_y, with σ_x and σ_y as the corresponding standard deviations,
- $p_{x+y}(k) = \sum_{i=1}^{N_g} sum_{j=1}^{N_g} p(i, j)$ for $i + j = k \mid k \in [2, 2N_g]$,
- $p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j)$ for $|i - j| = k \mid k \in [0, N_g - 1]$,
- $HX = -\sum_{i=1}^{N_g} p_x(i) \log_2 \big(p_x(i) + \epsilon\big)$ as the p_x entropy,
- $HY = -\sum_{j=1}^{N_g} p_y(j) \log_2 \big(p_y(j) + \epsilon\big)$ as the p_y entropy,
- $HXY = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \log_2 \big(p(i, j) + \epsilon\big)$ as the p(i,j) antropy
- $HXY1 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \log_2 \big(p_x(i)p_y(j) + \epsilon\big)$.
- $HXY2 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i)p_y(j) \log_2 \big(p_x(i)p_y(j) + \epsilon\big)$.

Table E.3: Definitions of extracted *Gray Level Co-Occurrence Matrix* features.

Description	Definition
Autocorrelation:	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j)ij$

Quantifies the texture fineness and coarseness of magnitude.	
Joint Average: The mean value of the distribution of i.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)i$
Cluster Prominence: Measures the skewness and asymmetry of the GLCM. Greater asymmetry about the mean is signified by higher feature value.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^4 p(i,j)$
Cluster Shade: Quantifies the skewness and uniformity of the GLCM.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^3 p(i,j)$
Cluster Tendency: Measures voxel regions with similar intensities.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^2 p(i,j)$
Contrast: Measure neighborhood intensity variations, favoring values away from the GLCM diagonal. Higher values signifies disparity in a neighborhood of intensities.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2 p(i,j)$
Correlation:	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)-\mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}$

The linear dependency between image values and the GLCM voexls.	
Difference Average: The relation between pairs of similar intensities and occurrences of pairs with different intensities.	$\sum_{k=0}^{N_g-1} kp_{x-y}(k)$
Difference Entropy: Quantifies randomness in differences between neighborhood intensity values.	$\sum_{k=0}^{N_g-1} p_{x-y}log_2(p_{x-y}(k) + \epsilon$
Difference Variance: Heterogeneity emphasizing pairs of differing intensity values deviating from the mean intensity.	$\sum_{k=0}^{N_g-1}(k - DA)^2 p_{x-y}$
Joint Energy: Quantifies homogeneous image patterns. Increased energy implies a more pairs of neighbouring intensities at higher values.	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g}(p(i,j))^2$
Joint Entropy: Measures variability in neighboring intensity values.	$-\sum_{i=1}^{N_g} \sum_{j=1}^{N_g}(p(i,j))log_2(p(i,j) + \epsilon)$
Informal Measure of Correlation 1:	

$$\frac{HXY - HXY1}{\max\{HX, HY\}}$$

An estimate of the correlation between the probability distributions of i and j based on mutual information to quantify texture complexity.	
Informal Measure of Correlation 2: An alternative measure of texture complexity.	$\sqrt{1-e^{-2(HXY2-HXY)}}$
Inverse Difference Moment Measures the local intensity homogeneity.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p(i,j)}{1+\lvert i-j\rvert^2}$
Maximal Correlation Coefficient: Quantifies the the complexity of image texture.	$\sum_{k=0}^{N_g}\frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$
Inverse Difference: An alternative measure for local image homogeneity.	$\sum_{k=0}^{N_g-1}\frac{p_{x-y}(k)}{1+k}$
Inverse Difference Normalized: An alternative measure of the local homogeneity that normalizes the difference between the neighboring intensity values by dividing by the total number of discrete intensity values.	$\sum_{k=0}^{N_g-1}\frac{p_{x-y}(k)}{1+\frac{k}{N_g}}$
Inverse Variance:	$\sum_{k=1}^{N_g-1}\frac{p_{x-y}(k)}{k^2}$

A homogeneity excluding the GLCM diagonal.	
Maximun Probability: The number of occurrences of the most predominant pair of neighboring image values.	$\max(p(i,j))$
Sum Entropy: The sum of neighborhood intensity differences.	$\sum_{k=2}^{2N_g} p_{x+y} \log_2(p_{x+y}(k) + \epsilon)$
Sum of Squares: Measures the distribution of neighbouring intensity pairs about the mean intensity in the GLCM.	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu_x)^2 p(i,j)$

E.4 Gray Level Size Zone Matrix Features

Let

- N_g denote the number of discrete image intensities,
- N_s denote the number discrete image zone sizes,
- N_p be the number image voxels,
- $N_z = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} P(i,j)$ subject to $1 \leq N_z \leq N_p$,
- $P(i,j)$ represent a GLSZM,
- and $p(i,j) = P(i,j)/N_z$ is a scaled GLSZM.

Table E.4: Definitions of extracted *Gray Level Size Zone Matrix* features.

Description	**Definition**
Small Area Emphasis Measures the distribution of small size zones. A greater value is indicative of more finer textures.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{j^2}}{N_Z}$
Large Area Emphasis Measures the distribution of large area size zones. Larger values indicates coarser textures.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)j^2}{N_Z}$
Gray Level Non–Uniformity The variability of image intensity values in the image. Smaller values points to greater similarity in intensities.	$\frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s}P(i,j)\right)^2}{N_z}$
Gray Level Non–Uniformity Normalized Scaled Gray Level Non–Uniformity.	$\frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s}P(i,j)\right)^2}{N_z^2}$
Size–Zone Non–Uniformity Size zone volume variability in the image. A lower value indicates homogeneity in size zone volumes.	$\frac{\sum_{j=1}^{N_s}\left(\sum_{i=1}^{N_g}P(i,j)\right)^2}{N_z}$
Size–Zone Non–Uniformity Normalized	$\frac{\sum_{j=1}^{N_s}\left(\sum_{i=1}^{N_g}P(i,j)\right)^2}{N_z^2}$

Scaled Size–Zone Non–Uniformity.	
Zone Percentage A measure image texture coarseness of the texture as the ratio of number of zones and number of voxels in the ROI.	$\frac{N_z}{N_p}$
Gray Level Variance The variance in zone intensities.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)(i-\mu)^2$
Zone Variance The variance in zone size volumes.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)(j-\mu)^2$
Zone Entropy The randomness in the distribution of zone sizes and gray levels. More heterogeneneity in texture patterns are signified by higher values.	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)\log_2(p(i,j)+\epsilon)$
Low Gray Level Zone Emphasis Measures the distribution of lower intensity size zones. Higher values indicates more of lower intensity values and size zones in the image.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{i^2}}{N_z}$
High Gray Level Zone Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)i^2}{N_z}$

The distribution of the higher intensities. Larger proportions of higher intensity values and size zones are indicated by higher values.	
Small Area Low Gray Level Emphasis The joint distribution of smaller size zones with lower intensities proportion of the image.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{i^2j^2}}{N_z}$
Small Area High Gray Level Emphasis The joint distribution of smaller size zones with higher intensity values proportion of the image.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)i^2}{j^2}}{N_z}$
Large Area Low Gray Level Emphasis The joint distribution of larger size zones with lower intensity values proportion of the image.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)j^2}{i^2}}{N_z}$
Large Area High Gray Level Emphasis The joint distribution of larger size zones with higher intensity values proportion of the image.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)i^2j^2}{N_z}$

E.5 Gray Level Run Length Matrix Features

Let

- N_g be the number of image intensities,
- N_r denote the number of discrete image run lengths,
- N_p be the number of image voxels,
- $N_z(\theta) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta)$ be the number of image runs along the angle θ, subject to $1 \leq N_z(\theta) \leq N_p$,
- $P(i,j|\theta$ is the GLRLM in direction θ,
- and $p(i,j|\theta) = \frac{P(i,j|\theta).}{N_z(\theta))}$ as the scaled GLRLM matrix.

Table E.5: Definitions of extracted *Gray Level Run Length Matrix* features.

Description	Definition
Short Run Emphasis Measures the distribution of short run lengths. A higher value points to finer textures structural textures.	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j\|\theta)}{j^2}}{N_r(\theta)}$
Long Run Emphasis Measures the distribution of long run lengths. A higher value indicates to coarse structural textures.	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j\|\theta) j^2}{N_r(\theta)}$
Gray Level Non–Uniformity The similarity of intensity values. Lower values correlates with a higher similarity in image values.	$\frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_r} P(i,j\|\theta) \right)^2}{N_r(\theta)}$

Gray Level Non–Uniformity Normalized (GLNN) Scaled Gray Level Non–Uniformity.	$\frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_r} P(i,j\mid\theta)\right)^2}{N_r(\theta)^2}$
Run Length Non–Uniformity The similarity of image run lengths. Lower values describes indicates more similarity among image run lengths.	$\frac{\sum_{j=1}^{N_r}\left(\sum_{i=1}^{N_g} P(i,j\mid\theta)\right)^2}{N_r(\theta)}$
Run Length Non-Uniformity Normalized Scaled Run Length Non–Uniformity.	$\frac{\sum_{j=1}^{N_r}\left(\sum_{i=1}^{N_g} P(i,j\mid\theta)\right)^2}{N_z(\theta)^2}$
Run Percentage Quantifies texture coarseness as the ratio of number of runs and number of voxels in the ROI.	$\frac{N_r(\theta)}{N_p}$
Gray Level Variance The intensity variance in image runs.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\mid\theta)(i-\mu)^2$
Run Variance The variance in runs over image run lengths.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\mid\theta)(j-\mu)^2$
Run Entropy	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\mid\theta)\log_2(p(i,j\mid\theta)+\epsilon)$

Randomness in the joint run lengths and intensity distribution. More randomness indicates more homogeneous texture patterns.	
Low Gray Level Run Emphasis The distribution of low intensity values. A lager concentration of low intensity is indicated by a higher value.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{i^2}}{N_r(\theta)}$
High Gray Level Run Emphasis The distribution of the higher intensities. Larger concentrations of higher intensity values are represented by a higher value.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}P(i,j\|\theta)i^2}{N_r(\theta)}$
Short Run Low Gray Level Emphasis Et mål på den felles distribusjonen av korte rekker med lave intensitetsverdier.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{i^2j^2}}{N_r(\theta)}$
Short Run High Gray Level Emphasis The join distribution of shorter run lengths with low intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)i^2}{j^2}}{N_r(\theta)}$
Long Run Low Gray Level Emphasis The joint distribution of shorter run lengths with higher intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)j^2}{i^2}}{N_r(\theta)}$

Long Run High Gray Level Emphasis The joint distribution of long run lengths with high intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} P(i,j\|\theta) i^2 j^2}{N_r(\theta)}$

E.6 Neighbouring Gray Tone Difference Matrix Features

Let

- n_i denote the number of voxels in the ROI with intensity equal to i,
- N_p be the number of image voxels,
- $p_i = \frac{n_i}{N_v}$ be the probability of an intensity,
- the sum of absolute differences for intensity i, be

$$s_i = \begin{cases} \sum_i |i - \bar{A}_i| & \text{if } n_i \neq 0 \\ 0 & \text{if } n_i = 0 \end{cases}$$

- N_g represent the number intensities
- and $N_{g,p}$ be the number of intensities with $p_i \neq 0$.

Table E.6: Definitions of extracted *Neighbouring Gray Tone Difference Matrix* features.

Description	Definition
Coarseness:	$\frac{1}{\sum_{i=1}^{N_g} p_i s_i}$

The average difference between a center voxel and neighbouring voxels indicating spatial rate of change. A higher value indicates more local uniform texture.	
Contrast: The spatial intensity change dependent on the intensity dynamic range. High contrasts indicate a high dynamic range and the spatial change rate.	$\left(\frac{1}{N_{g,p}(N_{g,p}-1)}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}p_ip_j(i-j)^2\right)\left(\frac{1}{N_p}\sum_{i=1}^{N_g}s_i\right)$
Busyness: Quantifies the change between a center voxel and a neighbourhood. Rapid changes of intensity between voxels and a neighbourhood is indicates by high values.	$\frac{\sum_{i=1}^{N_g}p_is_i}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\|ip_i-jp_j\|}$
Complexity: Measures non-uniformity and rapid intensity changes.	$\frac{1}{N_p}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\|i-j\|\frac{p_is_i+p_js_j}{p_i+p_j}$
Strength: Strength measures slow intensity changes, but more large coarse differences in gray level intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p_i+p_j)(i-j)^2}{\sum_{i=1}^{N_g}s_i}$

E.7 Gray Level Dependence Matrix Features

Let

- N_g be the number of image intensities,
- N_d represent the number of discrete image dependency sizes,
- $N_z = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} P(i,j)$ be the number of image dependency zones,
- $P(i,j)$ represent the dependency matrix
- and $p(i,j) = \frac{\mathbf{P}(i,j)}{N_z}$ be the scaled dependency matrix.

Table E.7: Definitions of extracted *Gray Level Dependence Matrix* features.

Description	**Definition**
Small Dependence Emphasis Measures the small dependency distribution. A is indicative of less homogeneous textures.	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{P(i,j)}{i^2}}{N_z}$
Large Dependence Emphasis Measures the distribution of large dependencies. Greater values indicates more homogeneous textures.	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} P(i,j) j^2}{N_z}$
Gray Level Non-Uniformity The similarity of image intensity values. Lower values points to more similar intensities.	$\frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_d} P(i,j) \right)^2}{N_z}$
Dependence Non-Uniformity	$\frac{\sum_{j=1}^{N_d} \left(\sum_{i=1}^{N_g} P(i,j) \right)^2}{N_z}$

Quantifies the overall similarity of dependencies in the image. Homogeneity among image dependencies are indicated with a lower value.	
Dependence Non-Uniformity Normalized Scaled Dependence Non-Uniformity	$\frac{\sum_{j=1}^{N_d}\left(\sum_{i=1}^{N_g} P(i,j)\right)^2}{N_z^2}$
Gray Level Variance The variance of image intensities.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_d} p(i,j)(i-\mu)^2$
Dependence Variance The variance in image dependencies.	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_d} p(i,j)(j-\mu)^2$
Dependence Entropy The entropy of dependencies.	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_d} p(i,j)\log_2(p(i,j)+\epsilon)$
Low Gray Level Emphasis Measures the distribution of low intensity values. Low values indicates a low concentration of low intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{P(i,j)}{i^2}}{N_z}$
High Gray Level Emphasis Measures the distribution of high intensity values. Low values indicates a low concentration of high intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d} P(i,j)i^2}{N_z}$
Small Dependence Low Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{P(i,j)}{i^2 j^2}}{N_z}$

The joint distribution of small dependencies and small image values.	
Small Dependence High Gray Level Emphasis (SDHGLE) The joint distribution of small dependencies and large image valiues.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{P(i,j)i^2}{j^2}}{N_z}$
Large Dependence Low Gray Level Emphasis (LDLGLE) The joint distribution of large dependencies and small intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{P(i,j)j^2}{i^2}}{N_z}$
Large Dependence High Gray Level Emphasis (LDHGLE) The joint distribution of large dependencies and large intensities.	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}P(i,j)i^2j^2}{N_z}$

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

Postboks 5003
NO-1432 Ås
Norway