Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Science and Technology

Predicting Patient Outcome Using Radioclinical Features Selected With RENT for Patients With Colorectal Cancer

Lars Jetmund Svartis Engesæth
Data Science

Abstract

Colorectal cancer remains a problem in medicine, costing countless lives each year. The growing amount of data available about these patients have piqued the interest of researchers, as they try to use machine learning to aid diagnosis, decision making, and treatment for these patients. Unfortunately, as the data sets grow, the risk of creating unstable and non-generalizable models increase.

The research in this thesis has aimed at investigating how to implement a novel technique called RENT (Repeated Elastic Net Technique) for feature selection. The predictive problem was a binary classification problem on colorectal cancer patients to predict overall survival. The analysis applied repeated stratified k-fold cross-validation with four folds and five repeats to reduce the risk of random subsets causing non-generalizable results. Further, the analysis created 25 000 different RENT models to search through the hyperparameters to find high performance parameter combinations. Each of the 25 000 models were trained with six different Random Forest [RF] hyperparameter combinations and twelve logistic regression hyperparameter combinations, resulting in 450 000 different models.

A high performing group of models was collected for one unique combination of hyperparameters. These models had the highest average test performance: accuracy 0.76 ± 0.07, MCC 0.47 ± 0.16, F1 positive class 0.57 ± 0.13, F1 negative class 0.83 ± 0.05, and AUC 0.69 ± 0.08. The results have also shown that the generalization error is lower for a RENT based RF model than non-RENT based RF model. The RENT analysis revealed that patients that died was overrepresented in a group of patients that were the most frequently predicted incorrectly. Finally, the RENT analysis has resulted in a distribution of features that were most frequently selected for high predictive ability. Most of the clinical features in this group has previously been reported as relevant by medical literature.

The research and the corresponding framework show promising results to implement a brute-force approach to the RENT analysis, to ensure low generalization error and predictive interpretability. Further research with this framework can support medicine in validating feature importance for patient outcome. The framework could also prove useful in other research fields than medicine, given predictive problems with similar challenges.

Sammendrag

Tykktarmskreft er fortsatt et problem innen medisin, og koster utallige liv hvert år. Den økende mengden data som er tilgjengelig om disse pasientene har vekket interessen til forskerne, der flere prøver å bruke maskinlæring for å hjelpe diagnostisering, beslutningstaking og behandling for disse pasientene. Dessverre, ettersom datasettene vokser, øker også risikoen for å lage ustabile og ikke-generaliserbare modeller.

Forskningen i denne oppgaven har tatt sikte på å undersøke hvordan man implementerer en ny teknikk kalt RENT (*Repeated Elastic Net Technique*) for variabel seleksjon. Det prediktive problemet var et binært klassifiseringsproblem på pasienter med tykk- og endetarmskreft for å forutsi samlet overlevelse. Analysen brukte gjentatt stratifisert k-foldet kryssvalidering med fire folder og fem repetisjoner for å redusere risikoen for at tilfeldige undergrupper av data fører til ikke-generaliserbare resultater. Videre beregnet analysen 25 000 forskjellige RENT-modeller for å søke gjennom hyperparametrene for å finne høyytelsesparameterkombinasjoner. Hver av de 25 000 modellene ble trent med seks forskjellige hyperparameterkombinasjoner for *Random Forest [RF]* og tolv hyperparameterkombinasjoner for logistisk regresjons, noe som resulterte i totalt 450 000 forskjellige modeller.

En høytytende gruppe modeller ble samlet inn for én unik kombinasjon av hyperparametre. Disse modellene hadde den høyeste gjennomsnittlige testytelsen: «*accuracy*» 0,76 ± 0,07, MCC 0,47 ± 0,16, F1 positiv klasse 0,57 ± 0,13, F1 negativ klasse 0,83 ± 0,05 og AUC 0,69 ± 0,08. Resultatene har også vist at generaliseringsfeilen er lavere for en RENT-basert RF-modell enn ikke-RENT-basert RF-modell. RENT-analysen avdekket at pasienter som døde var overrepresentert i en pasientgruppe som oftest ble predikert feil. Til slutt har RENT-analysen resultert i en fordeling av variabler som oftest ble valgt for høy prediksjonsevne. De fleste av de kliniske trekkene i denne gruppen er tidligere rapportert som relevante av medisinsk litteratur.

Forskningen og det tilhørende rammeverket viser lovende resultater for å implementere en *brute-force*-tilnærming til RENT-analysen, for å sikre lav generaliseringsfeil og prediktiv tolkbarhet. Ytterligere forskning med dette rammeverket kan bistå medisin i å validere variablers betydning for pasienters prognose. Rammeverket kan også vise seg nyttig innenfor andre forskningsfelt enn medisin, gitt prediktive problemer med lignende utfordringer.

Acknowledgements

This thesis marks the end of my 2-year study in Data Science at Norwegian University of Life Sciences. The research in this thesis has been extremely useful to develop my skills as an independent analyst, as well as a programmer. The final script was just shy of 4000 lines before it was compressed for better readability and usefulness to others.

The work the reader will find here would not be possible without my excellent supervisors, Professor Oliver Tomic and Professor Cecilia Marie Futsæther. Oliver has contributed as the main supervisor, and his genuine interest and enthusiasm to my research and my personal development is inspiring in its own. He has allowed me to learn independently while simultaneously helping me build 8-dimensional data structures on Friday afternoons. Cecilia helped me specify the research and her cunning wit has been invaluable for the quality of the thesis.

Furthermore, I would like to personally thank Professor Kathrine Røe Redalen for assisting me with the technical aspects of medicine outside of my speciality, as well as answering all my questions regarding the research that has served as the source of the data.

I would also like to thank my parents for their undying support, despite my continued refusal to follow a standard path through life. A heartfelt thanks is also directed to my friends and peers at NMBU, that have made my years at this university joyful and unforgettable, both in the study halls and during extracurricular activities.

Kanskje verden er litt stri,
men når det gråner skal du si,
at «du har hatt en bra studentertid!»
- «Studentenes Kall», 1933

Ås, 14.06.2022

Lars Jetmund Svartis Engesæth

Table of contents

List of figures

List of tables

Chapter 1 Introduction

1.1 Background and motivation

Cancer is non-specific term for a collection of diseases where cells grow uncontrollably in the body. It can affect any part of the body, and malignant cancers can spread to nearby tissue and metastasize (WHO 2021). In 2020, almost 20 million new cases of cancer was reported worldwide, while almost 10 million new deaths were reported (WHO 2020). The number of new deaths reported makes cancer the second leading cause of death worldwide (Our World in Data 2018).

Cancer can be classified by the type of cells they affect or from which organ they originate from. The third leading type of cancer is colorectal cancer (CRC), which includes cancer in the bowel, colon, and the rectum. These types of cancers are sometimes grouped together due to their similarities (American Cancer Society 2020). Almost 2 million cases of CRC and more than 900 000 deaths from CRC was reported in 2020 (WHO 2021). Higher age, obesity, cigarette smoking, drinking, and family history of CRC is correlated with the risk of developing CRC (National Cancer Institute 2021). The relative survival rate of colon cancer and rectal cancer is presented in table 1 below:

Table 1: Table with relative survival rates for patients with colon cancer and rectal cancer with differentiation based on SEER (Surveillance, Epidemiology, and End Results) status. The numbers in table 1 were gathered from the American Cancer Society-webpage, with clarifications about the definitions regarding SEER stage (American Cancer Society 2022).

	5-year relative survival rate	
SEER stage	**Colon cancer**	**Rectal cancer**
Localized	91 %	90 %
Regional	72 %	73 %
Distant	14 %	17 %
All SEER stages combined	64 %	67 %

Table 1 highlights how survival rate decreases significantly as the cancer spreads to other parts of the body, which supports the importance of diagnosing and treating CRC as soon as possible.

Treatment of cancer depends on the tumour, but patients must usually undergo tomographic photography using either CT, PET, or MRI, or all of these imaging types (Gillies, Kinahan et al. 2016). These images are based on different technologies which in turn yield different tissue characteristics (Höhne, Fuchs et al. 1990). Due to the abundance of data available in these images, research into radiomics and data-supported decision making has exploded in the past decades. As the tomographic imaging has increased, so has the need for radiologists to interpret these images. Thus, machine learning (ML) has become increasingly used since processing power has become cheaper and more available.

Lately, ML has moved into deep learning to overcome the massive amounts of data available (Kleppe, Skrede et al. 2021). However, as deep learning tends to lead to black box models, where the internal logic of the model is unknown, research is still investigating the use of more transparent and interpretable models (LeCun, Bengio et al. 2015, Wiemken and Kelley 2020). Unfortunately, as the data sets increase in width with added feature columns and the number of patients is often small, ML algorithms are prone to overfitting to the training data and resulting in poor generalization to new data. Therefore, feature selection has been investigated as a method to rectify overfitting by limiting the number of features that are available for the models.

Furthermore, great progress has been accomplished in the research into radiomics as an aid in medical practice. Radiomics has been used increasingly in predictive modelling in medicine, especially for cancer research (Aerts, Velazquez et al. 2014, Parekh and Jacobs 2017, Parekh and Jacobs 2019, Li, Zhu et al. 2020). Some studies find that the radiomics features improve predictive ability. However, most research must perform some sort of feature selection to reduce the features space to improve model generalization. In this thesis, the feature selection will be performed by using RENT (Repeated Elastic Net Technique), that has been introduced by researchers from Norwegian University of Life Sciences (Jenul, Schrunner et al. 2021, Jenul, Schrunner et al. 2021).

1.2 Previous work

RENT has previously been applied by master students from NMBU to investigate its application and performance. Sana used RENT for feature selection and then investigated the explained variance of the target with the selected features (Sana 2021). Mohammadi implemented RENT for feature selection and then predicted medication class for patients treated for ADHD and found improved predictive performance using RENT (Mohammadi 2021). Olofsson used RENT for feature selection and repeated stratified k-fold cross validation to identify biomarkers in Alzheimer's patients (Olofsson 2021). Søvdsnes implemented RENT for feature selection on clinical and radiomic features to improve predictive ability of radiomic features applied to patients with colorectal cancer (Søvdsnes 2021).

1.3 Goal and research questions

Given the background presented in the previous sections, there are several reasons to further explore the possibilities related to medical data, cancer research, and machine learning.

This thesis aims to explore the following questions:

1. How accurately can machine learning models predict OS in colorectal cancer patients when RENT performs feature selection?
2. Which features are the most important in this data set to predict OS in patients with colorectal cancer?

Considering previous theses has used RENT as normal, this thesis was motivated to instead try a thorough brute-force approach to search through the parameters that RENT normally chooses for the user.

1.4 Outline

Chapter 2 will present relevant theoretical background to introduce relevant topics in order to understand the thesis. The reader is expected to have some background information in statistics, programming, and machine learning.

Chapter 3 will introduce the methods applied in the analysis. CRISP-DM will be introduced, and then data preparation, treatment, and modelling will be discussed. A thorough explanation of the implementation of RENT will help the reader understand the novel feature selection technique.

Chapter 4 will present the results, both in terms of average model performance as well as the performance for the best models. Additionally, there will be presented results for a baseline model without the assistance from RENT. Lastly, this chapter will present findings from the RENT analysis.

Chapter 5 will discuss the results, the implication of the results, and give some interpretation of the results. The chapter will relate the selected features to published literature to validate the findings.

Chapter 6 will present the concluding remarks for the thesis. This chapter will also give suggestions for further work that have sprung forward as a result from working on the thesis.

Chapter 2 Theory

This chapter will present the theoretical background needed to understand the research carried out in this thesis, the models created and how to evaluate these models. Certain topics are not presented in-depth, as this master thesis is rooted in data science. Therefore, certain details are outside the scope of the thesis. However, whenever applicable, there will be presented sources if the reader wishes to pursue these topics.

2.1 Feature selection

As data collection techniques continuously improve, there is a greater access to data for predictive modelling. As data collection increases in breadth by logging more features, there is a rising number of situations where the number of features in the data set is greater than the number of samples. This imbalance between the features and the number of samples is sometimes referred to as "*The Curse of Dimensionality*" as coined by Richard Bellman in 1957 in his book "Dynamic Programming" (Bellman, Bellman et al. 1957). The term refers to the problem that the feature space increases so fast that the volume of data becomes sparse, which negatively affects predictive performance. Feature selection methods allow a researcher to identify a subset of features to that may be used for further analysis, which may lead to a number of benefits as elaborated by Guyon and Elisseeff (Guyon and Elisseeff 2003):

- Improve data visualization and data understanding,
- Reducing measurement and data storage requirements,
- Reduced training time,
- Ease the *curse of dimensionality* to improve predictive performance.

The optimal number of features for a data set is debated, but various sources claim that the feature volume should be 10 times the sample volume. In 1993, Hush and Horne found that a Multi-layer Perceptron classifier should have 10 times as many samples as weights (Hush and Horne 1993). Gillies et al. and Sollini et al. echo that their personal preference is having at least 10 samples per one feature (Gillies, Kinahan et al. 2016, Sollini, Antunovic et al. 2019). Others have suggested that there should be a 10 to 1 ratio between positive outcomes/events and features (Harrell, Lee et al. 1996, Peduzzi, Concato et al. 1996).

Provided that the training data contains relevant information, as the features space increases, the test performance of a classifier will usually first increase, and then as the feature space increases further the test performance will usually decrease (Sima and Dougherty 2008). This effect is illustrated below in Figure 1. This predictive performance peak was originally demonstrated by G. Hughes in 1968 (therefore referred to as "*Hughes Phenomenon*" in literature) on discrete classifiers, but it can affect all classifiers (Hughes 1968).

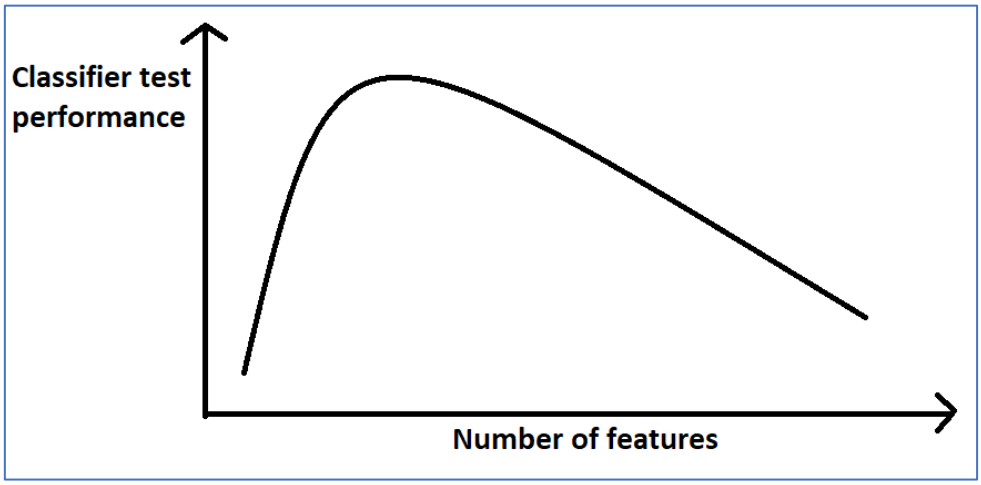


Figure 1: Illustration of Hughes phenomenon. Drawn with inspiration from: (Decourselle, Simon et al. 2012)

In contrast to Hush and Horne, it has been shown that the optimal feature size relative to sample size depend strongly on the classifier in question, and that generalizing a rule-of-thumb should be done warily (Hua, Xiong et al. 2004). Furthermore, the results from Hua et al. indicate that the correlation between the features changes the optimal ratio between sample size and feature size (Hua, Xiong et al. 2004). van Smeden et al. directly opposed that positive events per feature held any rationale for logistic regression (van Smeden, de Groot et al. 2016). Theodoridis and Koutroumbas support the findings of Hau and Xiong et al. that the effect of ratio between features and samples cannot be generalized to all classifiers (Theodoridis and Koutroumbas 2008).

In essence, these findings indicate that there is some sort of global maximum for the model performance as a function of the number of features included. This motivates the research and analysis performed on features selection in this thesis. A model consisting of too few features result in poor performance (underfitting), since crucial information is missing. Contrastingly, models consisting of too many features also result in poor performance (overfitting) since the there is a mismatch between the number of samples and features (Theodoridis and Koutroumbas 2008).

2.2 Imaging in medicine

Medicine has, since the discovery of the X-ray in 1895 by W.C. Röntgen, implemented more and more imaging techniques to aid decision making (Smithsonian Institution 1946, Tubiana 1996, Udupa and Herman 1999). Udupa and Herman define the purpose of 3D imaging as:

> *"The purpose of 3D imaging is: given a set of multidimensional images pertaining to an object/object system,* to output *qualitative/quantitative information about the object/object system under study."* (Udupa and Herman 1999)

There are numerous image types, and they are usually 2D slices of certain body parts. They can be represented in 3D as a number of 2D slices to create a volume of tomographic images (Udupa and Herman 1999). When these images are evaluated over time, the time dimension adds another dimensionality to the images, which means some image types are 4-dimensional. These 4th dimension is an approximation, since it is always possible to sample an image at a shorter time interval (Udupa and Herman 1999). The precision of this dimension is improved by having shorter time between each sample.

2.2.1 Medical image types in this thesis

The most commonly used images aside from the mentioned X-rays are Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Position Emission Tomography (PET) (COCIR s.a.). This thesis analysed Diffusion Weighted Images (DWI), Fast Field Echo (FFE) Images, and T1- and T2-weighted MRI (referred to in this thesis as *T1T2Sense*).

DWI was introduced in 1986 as a variation of MRI, where diffusion of water molecules created the contrast in the images (Merboldt, Hanicke et al. 1985, Le Bihan, Breton et al. 1986, Posse, Cuenod et al. 1993).

FFE is an imaging technique that is based on gradient echo sequences with reduced flip angle excitation pulses, and has been used as an image variant in medicine to examine body tissue (van der Meulen, Groen et al. 1988, Preziosi, Orlacchio et al. 2003, Lee, Shin et al. 2017).

T1T2Sense are MRI images that use both T1-weighting and T2-weighting. T1- and T2-weighting refers to different relaxation times in tissue (Rinck 2021). T1- and T2-weighting are sensitive to different types of bodily tissue.

However, the medical details regarding these imaging techniques are left outside the scope of this thesis. Discussions surrounding the features extracted from these images will be based on a data science perspective, and not in a medical science perspective. As described in chapter 3.2.2 *Medical Images*, the descriptive features of these images will be extracted by an algorithm, which results in feature names that are based in image analysis instead of medicine.

2.3 Radiomics

Radiomics is a study emerging from the combination of radiology and oncology (Parekh and Jacobs 2017). The goal is to extract the numerical values from an image and feed these datasets into algorithms to gain insights beyond the perception of human eyes. Thus, personalized diagnosis can be given to each patient (Parekh and Jacobs 2019). Furthermore, large imaging data sets of body parts may be swiftly triaged to identify which cases requires attention first (Parekh and Jacobs 2019). Research has been motivated to focus on cancer, since almost every cancer patient undergoes tomographic imaging (e.g.: CT, PET, MRI) at least once during their diagnosis and treatment (Gillies, Kinahan et al. 2016).

Radiomic images contain different types of information that can be divided into first order-, second order-, and higher order statistics (Gillies, Kinahan et al. 2016). These three subtypes refer to different feature types. The combination of these distinctions lead to a high number of features, up to 440 (Aerts, Velazquez et al. 2014) and 636 (Grossmann, Stringfield et al. 2017) features. As the number of available features grow, there is a greater risk of overfitting a model to the training data. Therefore, there is a need to reduce the dimensionality of the feature space. Gillies et al. suggests the following methods to reduce the feature space (Gillies, Kinahan et al. 2016):

- Remove highly redundant features, such as highly correlated features,
- Test-retest data can be applied to investigate if the results are reproducible,
- Build models with the highest-priority features with features representing each agnostic and semantic class (shape, size, and first-, second-, and higher-order textures).

Balagurunathan et al. found that most CT features are highly reproducible in a test-retest analysis, and suggest that these features can be combined with clinical features to improve predictability (Balagurunathan, Kumar et al. 2014). A test-retest analysis measures the same characteristic twice to evaluate if the measures are unchanged (Vilagut 2014).

Radiomics has been used to enable diagnosis of prostate cancer (Wibmer, Hricak et al. 2015), prognosis for head- and neck-cancer (Aerts, Velazquez et al. 2014), and predict outcome for lung cancer patients (Kadir and Gleeson 2018). Research has also shown promising results for selecting individualize treatment based on radiomics features (Teruel, Heldahl et al. 2014, Lv, Xin et al. 2022). Parekh and Jacobs have presented a novel multiparametric radiomics (mpRAD) approach to improve predictive performance (Parekh and Jacobs 2017, Parekh and Jacobs 2019, Parekh and Jacobs 2020). Alternatively, several researchers have published articles on multiparametric MRI which indicate promising predictive results (Barentsz, Richenberg et al. 2012, Yoo, Kim et al. 2015, Demirel and Davis 2018).

All in all, radiomics is an emerging field with possibilities and challenges. Gillies et al. highlight the following challenges for radiomics:

- Reproducibility: As a young discipline, radiomics might suffer the same slow progress that molecular biology diagnostics underwent,
- Big Data: There are such vast amounts of medical data available that radiomics might lead to further complications and bottlenecks in research,
- Data sharing: Data sharing can be a challenge due to its sensitive nature and overcoming these challenges can lead to more robust research (Gillies, Kinahan et al. 2016).

In the context of this thesis, the data sharing could be enabled due to the data set being anonymized. Data could only be shared if it the patients agreed to it. The big data aspect is also relevant in this thesis, considering there were in total 191 images and only 3 of these were extracted. Details regarding this process in discussed in 3.2.2 *Medical images*. As discussed in chapter 1.1 *Background and motivation*, radiomics and medicine is prone to having too many features, so extracting more features from the other image echoes would introduce even more potentially redundant features.

2.4 Overfitting and underfitting

Overfitting is a problem that occurs in ML when a model performs well on the training data, but it does not generalize to the unseen test data (Raschka and Mirjalili 2019). Figure 2 below shows an example of how underfit, well fit, and overfit models could separate binary training samples.

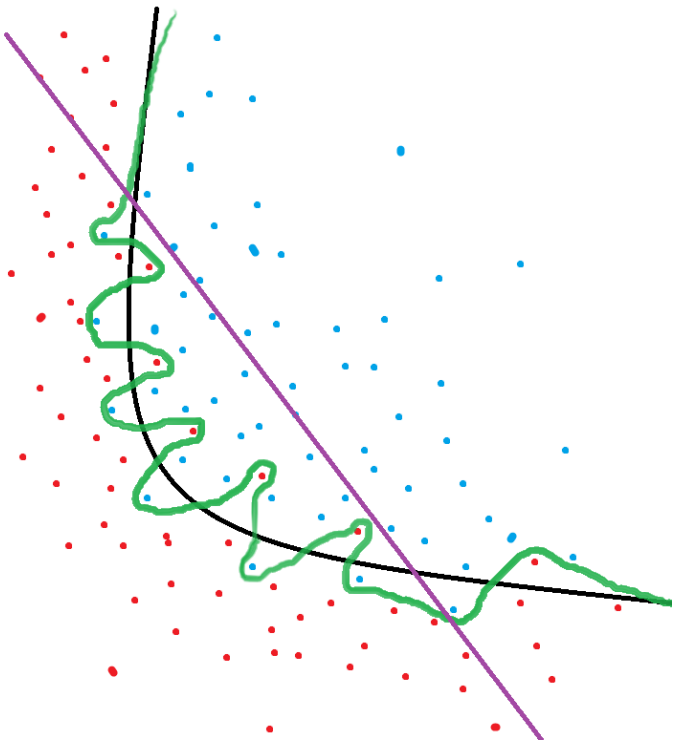


Figure 2: This figure illustrates an underfit model (purple), an overfit model (green), and better fit model (black) to the data. Drawn with inspiration from (Raschka and Mirjalili 2019).

Figure 2 illustrates an example of three different models attempting to classify blue and red samples, with one feature on the x-axis and one features on the y-axis. The true relationship in the data is nonlinear. The model drawing the green line is thus overfitted, since it classifies all training samples correctly, but it would likely perform poorly on test data. Likewise, the purple model is underfitted and would perform poorly on both test data and train data. The black line is probably the most appropriate curve, despite the performance not being 100% accurate.

Overfitting and underfitting is often explained using the following two terms: *bias* and *variance*. Bias refers to the component of error resulting from a too simple learning algorithm (James, Witten et al. 2013). In Figure 2 above, the purple model has high bias, since it tries to fit a linear model to a non-linear relationship. Variance refers to the degree which a learning algorithm would change if the samples in the training subset was replaced with another subset of the data set (James, Witten et al. 2013). The green squiggly line in Figure 2 above follows each point precisely. If any of the red points it circles would have been replaced, the model would probably change significantly. Thus, this model has high variance. Generally, high bias is related to underfit models, while high variance is related to overfit models (Raschka and Mirjalili 2019). A technique to attempt to minimize both bias and variance is explained in the next chapter.

2.4.1 Regularization

Since both high variance and high bias is unwanted, data scientists often attempt to find a compromise called the *bias-variance trade-off*. When bias is reduced, variance grows, and vice versa. This effect is illustrated in Figure 3 below. The total error due to bias and variance has a minimum point that is a balance between minimized variance and minimized bias.

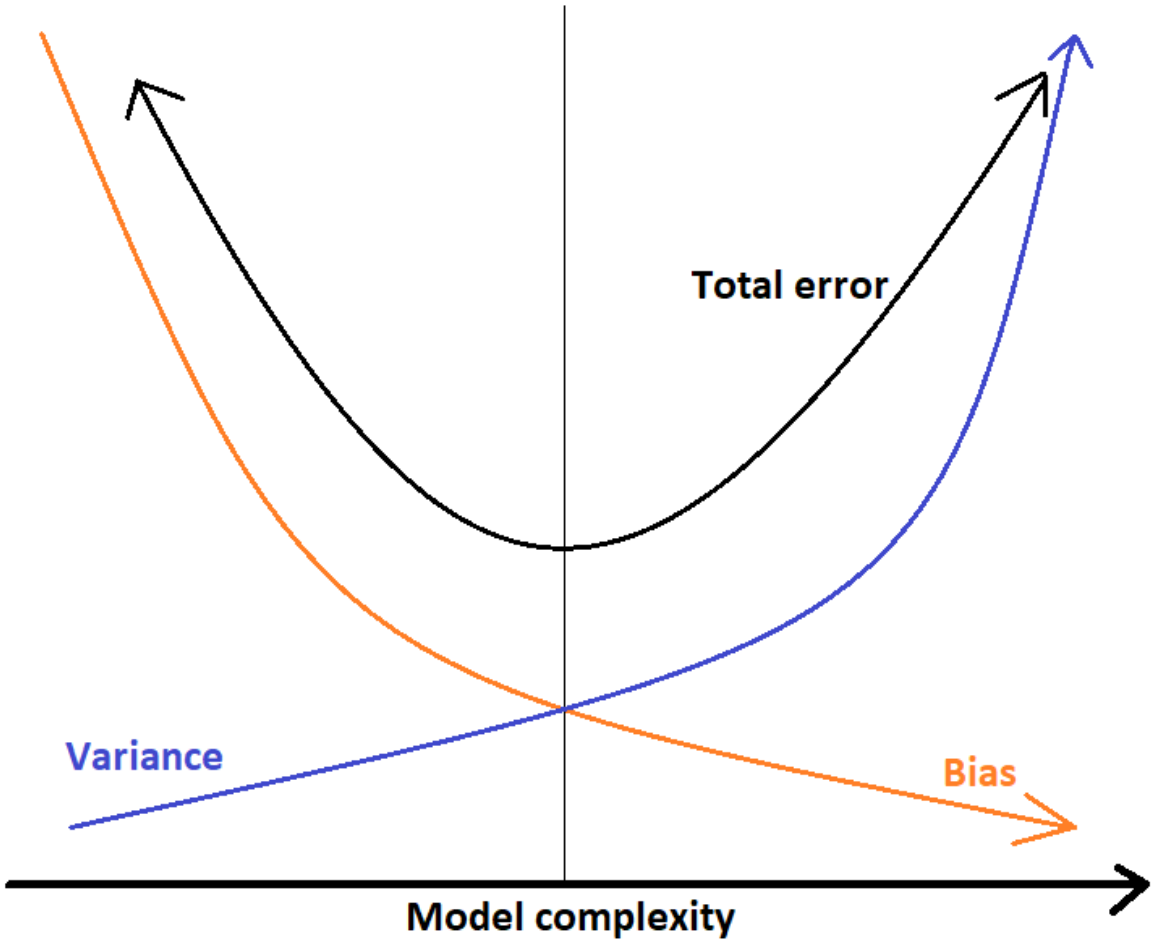


Figure 3: Illustration of bias-variance-trade-off. As the model complexity increases, the bias increases, and the variance decreases. The thin and vertical line marks the optimal balance between bias and variance. Drawn with inspiration from (Huilgol 2020).

One technique to find this optimal point is using regularization (Raschka and Mirjalili 2019). Regularization introduces additional information to penalize extreme parameter values. It is an effective tool to take care of collinearity and filtering out noise, both of which ultimately lead to less risk of overfitting. Furthermore, Bousquet and Elisseeff have found that regularization implies stability in the models (Bousquet and Elisseeff 2002).

There are two main ways of performing regularization: L1 regularization (also called *LASSO*) and L2 regularization (also called *Ridge*). The L1 or L2 penalty are added to the cost function

of the given learning algorithm (see chapter 2.8.1 *Logistic regression* for information about the cost function for logistic regression). The formula for L1 penalty and L2 penalty is as follows:

$$L1: \delta\|w\|_1 = \delta \sum_{j=1}^{m} |w_j|$$

$$L2: \delta\|w\|_2^2 = \delta \sum_{j=1}^{m} {w_j}^2$$

where *w* is the weight vector, δ is the penalty term, *j* takes the values from 1 to *m*, and *m* is the dimensionality of the number of samples. Both regularization methods shrink weights, which reduces overfitting. However, only L1 has the ability of creating sparse vectors, where the weights are set to 0, which leads to feature selection through feature exclusion. Both L1 and L2 can be more suitable depending on the data set and situation. RENT uses *Elastic Net*, a combination of L1 and L2 regularization, and elastic net is therefore introduced in the next chapter.

2.4.2 Elastic Net

Elastic Net is a compromise between LASSO (L1) regression and Ridge (L2) regression (Raschka and Mirjalili 2019). Elastic net was introduced by Zou and Hastie in 2005, because it could outperform LASSO, while still shrinking some features to 0 (Zou and Hastie 2005). LASSO regression is preferred for the ability to remove features by selecting the most important ones. As described in chapter 2.1 Feature selection, this will lead to faster and more stable models. This advantage is combined with Ridge regression, which is able to shrink weights and group similar features. Ridge regression is used when the dataset includes highly correlated features (Hilt and Seegrist 1977).

Elastic Net is tailored to the data set by adjusting the L1-ratio between L1 and L2 regularization.

$$ElasticNet = J(w) + \delta\left[(\alpha * L1) + \left((1-\alpha) * L2\right)\right]$$

Where *J(w)* is the cost function of the classifier, δ is the penalty term, and α is the L1-ratio. When the L1-ratio is set to 0, there is only L2-regularization and no L1-regularization, and vice versa for when the L1-ratio is set to 1 (Nogueira, Sechidis et al. 2018).

Furthermore, Nogueira et al. showed experimentally that by using Elastic Net they could achieve high stability despite similar loss as LASSO regression alone, which in turn recovers the "true" set of relevant features (Nogueira, Sechidis et al. 2018). In the research for this thesis, it was implemented a novel method called RENT (Repeated Elastic Net Technique), which is essentially Elastic Net with repeats to further improve stability and generalization (Jenul, Schrunner et al. 2021, Jenul, Schrunner et al. 2021). RENT is introduced in chapter **Feil! Fant i kke referansekilden.** ***Feil! Fant ikke referansekilden.***, and its implementation is described in chapter 3.4.2 *Implementation of RENT*.

2.5 RENT introduction

In this thesis, the data set includes several sets of highly correlated features. For example: The type of cancer stage is highly correlated with the treatment the patient receives. Thus, the idea is to include both L2 and L1 regularization, as discussed in chapter 2.4.2 *Elastic Net*, to simultaneously remove redundant features and find appropriate weights for the remaining correlated features. Bøvelstad et al. found that learning should be repeated on subsets, and that feature selection methods implementing feature shrinkage or linear combinations outperform univariate and stepwise feature selection, further supporting the motivation to use RENT (Bøvelstad, Nygård et al. 2007). The idea behind RENT is to use an ensemble of generalized linear models on subsets of the data with elastic net regularization for feature selection (Jenul, Schrunner et al. 2021, Jenul, Schrunner et al. 2021). Figure 4 below shows an overview of the RENT pipeline. A detailed description of the implementation of RENT is presented in chapter 3.4.2 *Implementation of RENT*.

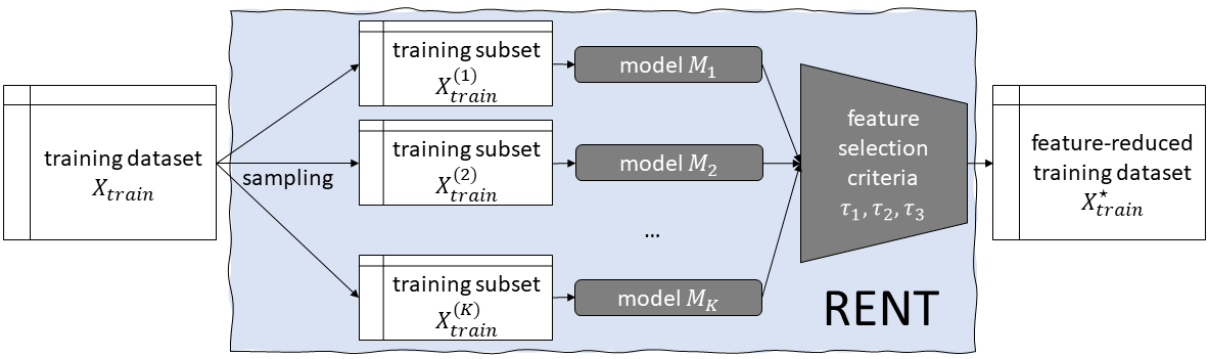


Figure 4: (Jenul, Schrunner et al. 2021). RENT feature selection pipeline. RENT models with different subsets are trained k *times on the data set to create an ensemble of feature selections.*

The developers of RENT has found that implementation of RENT can achieve stable and reproducible results that also keep predictive performance at a high level (Jenul, Schrunner et al. 2021). When RENT is compared to other feature selectors, RENT is almost always competitive. Whenever RENT scores slightly worse than other feature selectors, it has significantly reduced the dimensionality of the problem and thus reduced runtime and computational complexity and strain (Jenul, Schrunner et al. 2021).

2.6 Stability and generalization

An important measure in machine learning is the performance of a model. Performance metrics are explained in the next chapter. However, performance is generally calculated on the basis of a stable performance, which is affected by randomness. Two sources lead to randomness in the model: The sampling between the training set and the test set, and the noise in the data points (Bousquet and Elisseeff 2002). RENT and this thesis will use the measure of stability for a feature selector introduced by Nogueira et al.:

"An algorithm is 'unstable' if a small change in data leads to large changes in the chosen feature subset." (Nogueira, Sechidis et al. 2018)

Previous work by Guyon and Elisseeff, and Brown et al. has been performed to attempt to define techniques to find meaningful subsets of features (Guyon and Elisseeff 2003) (Brown, Pocock et al. 2012). Nogueira et al. showed that Elastic Net was able to find the "true set" of relevant features from a synthetic data set (Nogueira, Sechidis et al. 2018).

2.7 Performance metrics

There are numerous methods for evaluating ML models and their performance. A "confusion matrix" is commonly used to explain and visualize the different methods for binary classification. This thesis predicts a binary classification problem, so the confusion matrix is appropriate. Other methods exist to evaluate multiclass problems, but they will not be covered here. The following Figure 5 show a confusion matrix:

		Predicted class:	
		Positive	Negative
True class:	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 5: A confusion matrix help visualize the relationship between TP, FN, FP, and TN. The correct predictions, TP and TN, are color coded green, while the incorrect predictions are coloured pink. With inspiration from inspired by (James, Witten et al. 2013) and (Raschka and Mirjalili 2019).

TP and TN are predictions that are correctly predicted positive or negative. FP and FN are incorrect predictions. The naming of a positive event is arbitrary and does not necessarily reflect a desired outcome. For this thesis, a positive outcome (1) is a patient that has died, while a negative outcome (0) is a patient that survived. The simplest method to evaluate a model to is to calculate "accuracy". Accuracy is calculated by computing the amount of correct predictions (TP and TN) against the number of total predictions (James, Witten et al. 2013):

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

However, accuracy is rarely reported alone, and sometimes disregarded in comparison to more robust performance metrics. For data sets with where the class distribution is far from 50/50, accuracy can give misleading results. For example, if a data set has 90% positive samples, the model can achieve a 90% accuracy by predicting all samples as positive. A model like this has likely not learned anything, but the accuracy parameter falsely gives that impression. For this reason, several performance metrics is usually used to give higher scientific weight to results.

Two other commonly used performance metrics are "precision" and "recall". Precision and recall are more sensitive to false positives and false negatives, respectively. The formula for **precision** is:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

which calculates how many of the positive predictions are correctly predicted (Raschka and Mirjalili 2019).

Recall uses the following formula:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

which takes the fraction of true positive predictions against the total number of positive samples (Raschka and Mirjalili 2019).

Precision and recall both have weaknesses and strengths for different data sets. To optimize for this difference, it is possible to use the **F1-score** (sometimes called F-score), which uses both recall and precision in its formula (Raschka and Mirjalili 2019):

$$F1\ score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Additionally, **Matthews Correlation Coefficient** (MCC) is another performance metric that is especially relevant for unbalanced data sets. The formula below is based on the confusion matrix and was introduced in 1975 by B.W. Matthews (Matthews 1975):

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Finally, the **Receiver Operating Characteristic** (ROC) and its related Area Under Curve (AUC) is often computed to benchmark a model (Hanley and McNeil 1983). The ROC is a graphical plot that visualizes how the predictive precision of a model changes as the false positive rate is changed (Fawcett 2006). The integral of the area under this curve is the AUC, and this integral should be higher than 0.5 to beat a model that is guessing randomly. Figure 6 below shows an artificial ROC-plot.

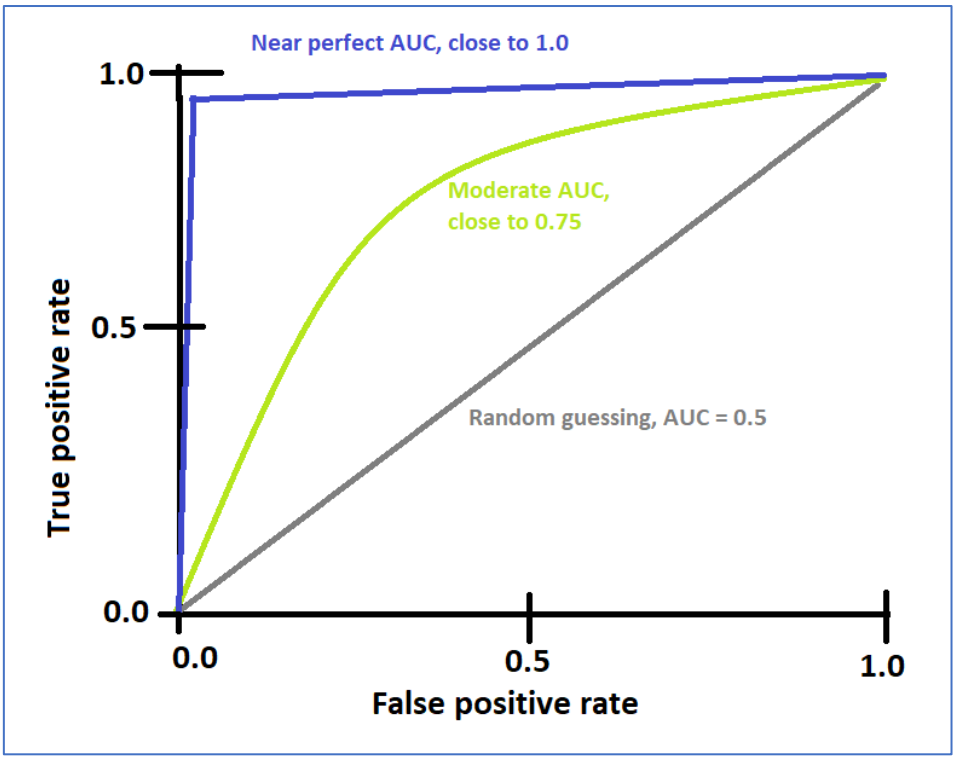


Figure 6: Illustration of an artificially drawn ROC. The definition of a "moderate" test performance is dependent on the data set it is tested on. The figure was drawn with inspiration from: (Juliani and Ellefmo 2019)

All these performance metrics can be automatically calculated by *sklearn* (Pedregosa, Varoquaux et al. 2011).

Accuracy and F1-score take values between 0 and 1. MCC takes values from -1 to 1, where -1 equals 100% incorrect predictions, 0 equals true random guessing, and 1 means 100% correct predictions. AUC technically takes values from 0 to 1, although an AUC-value of 0.5 implies that the model is equally accurate as randomly guessing. AUC lower than 0.5 implies worse than random guessing, and AUC equal 1 is 100% accurate predictions for any true positive and false positive rates.

2.8 Machine learning algorithms

Machine learning has been around since the 1950s when Alan Turing discussed *learning machines* in his paper on “The imitation game” (TURING 1950). Since then, several machine learning algorithms have been developed to learning problems. Machine learning algorithms are usually divided into supervised and unsupervised algorithms, as well as reinforced learning

(Raschka and Mirjalili 2019). Supervised learning algorithms have a label for each sample, which means that the algorithm predicts an outcome for that sample. Unsupervised learning algorithms do not have this label, which means that these algorithms try to find hidden patterns or structures in the data without the guidance of sample labels.

In the following chapters, Logistic Regression and Random Forest will be introduced, as these algorithms were used in this thesis. There are, however, numerous alternative machine learning algorithms available.

2.8.1 Logistic regression

Logistic regression was introduced by Pierre-Francois Verhulst in three papers between 1838 and 1847 after inspiration from Thomas Robert Malthus' book "An Essay on the Principle of Population" from 1798 (Cramer 2002). Verhulst's works were only discovered in 1920, and the logistic function has since then been applied in biology, biomathematics, chemistry, statistics, and data science, to only name a few of the scientific areas. The logistic expression is as follows (James, Witten et al. 2013):

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

where p is the probability of the sample belonging to class 1, e is the base of the natural logarithm, β_0 and β_1 are the parameters of the model. β_0 is also referred to as the intercept of the model, while β_1 is the slope of the model.

The activation function for logistic regression defined as the following expression which is based on the *logit* function (Raschka and Mirjalili 2019):

$$logit(p) = \log\frac{p}{(1-p)}$$

where log is the natural logarithm, and p is the probability of the positive event.

The inverse of the logit function, also called logistic sigmoid function, is defined as follows:

$$\varphi(z) = \frac{1}{(1+e^{-z})}$$

where z is a combination of the net input, the linear combinations, and the inputs.

Figure 7 below is an example of how the s-curve from logistic function plots. The graph never outputs a value lower than 0, or a value higher than 1 for any x-value. Thus, any class prediction of a sample is always contained inside the [0,1]-interval.

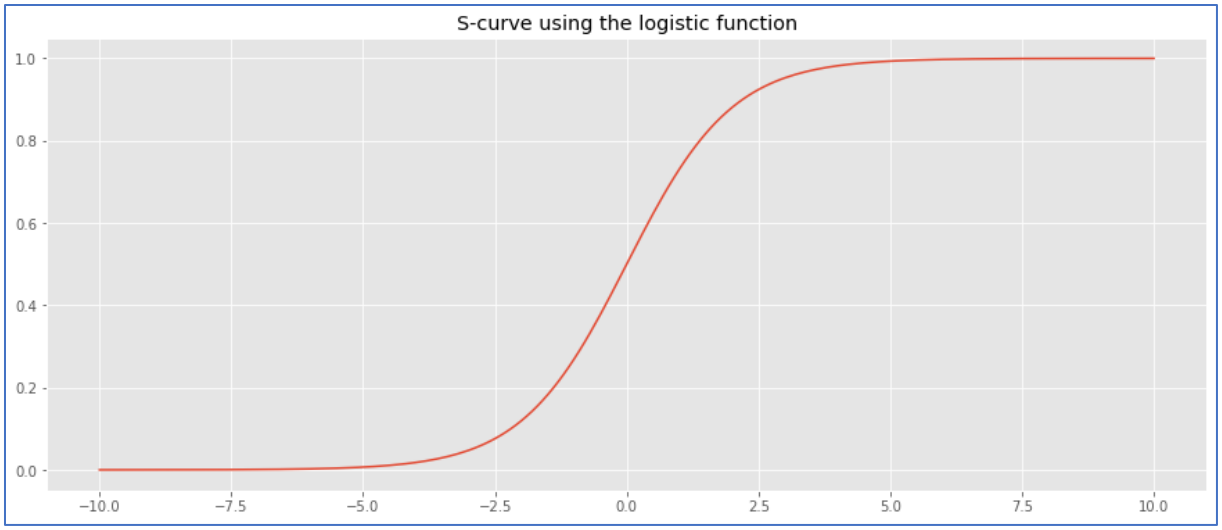


Figure 7: The plot of the logistic function is often called a Sigmoid curve due to the shape. The curve is constrained in the [0,1]-interval, and increasing the x-parameters to higher or lower than +/-10 will not change this constraint. The figure is plotted using the logistic function included in the Python library "scipy.special".

Furthermore, the logistic regression presented above had only one explanatory variable, β_1. This model can be generalized to include any number of β up to β_m. The model remains linear regardless of the number of β (Hosmer Jr., Lemeshow et al. 2013, James, Witten et al. 2013).

The optimal β_m are found such that they maximize the log-likelihood, l, presented below (Raschka and Mirjalili 2019):

$$l(\mathrm{w}) = \log L(\mathrm{w}) = \sum_{i=1}^{n} \left[y^{(i)} \log\left(\varphi\left(z^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - \varphi\left(z^{(i)}\right)\right) \right]$$

where log is the natural logarithm, y is the sample label, and $\varphi(z)$ is the activation function for logistic regression.

By rewriting this expression as a cost function, gradient descent can be applied to minimize the new expression:

$$\mathrm{J}(\mathrm{w}) = \sum_{i=1}^{n} \left[-y^{(i)} \log\left(\varphi\left(\mathrm{z}^{(i)}\right)\right) - \left(1 - y^{(i)}\right) \log\left(1 - \varphi\left(z^{(i)}\right)\right) \right]$$

where $\varphi(z)$ is the activation for LR. The activation function for LR is the sigmoid function defined previously in this chapter. The y is the label for the ith sample. If the first term has y=0, that term will be reduced to zero. Likewise, the second term will be reduced to zero if the y=1.

2.8.2 Random forest

Random Forest is an extension to decision trees that was introduced by Tin Kam Ho in 1995, where a collection of decision trees based on pseudo-randomly generated subspaces are combined to make a combined classification model (Ho 1995). Decision trees are prone to overfitting to the training data if they are allowed to be deep enough (Ho 1995). Ho later showed that random forests could improve generalization performance, without reducing training performance (Ho 1998).

Decision trees were first introduced by Morgan and Sonquist in 1963 (Loh 2015) or Hunt in 1966 (Stefanowski 2008), depending on which source is to be believed. However, decision trees would be improved upon since then to be included in classification problems (Stefanowski 2008, Loh 2015). The algorithm tries to mimic human decision making, as visualized in Figure 8 below:

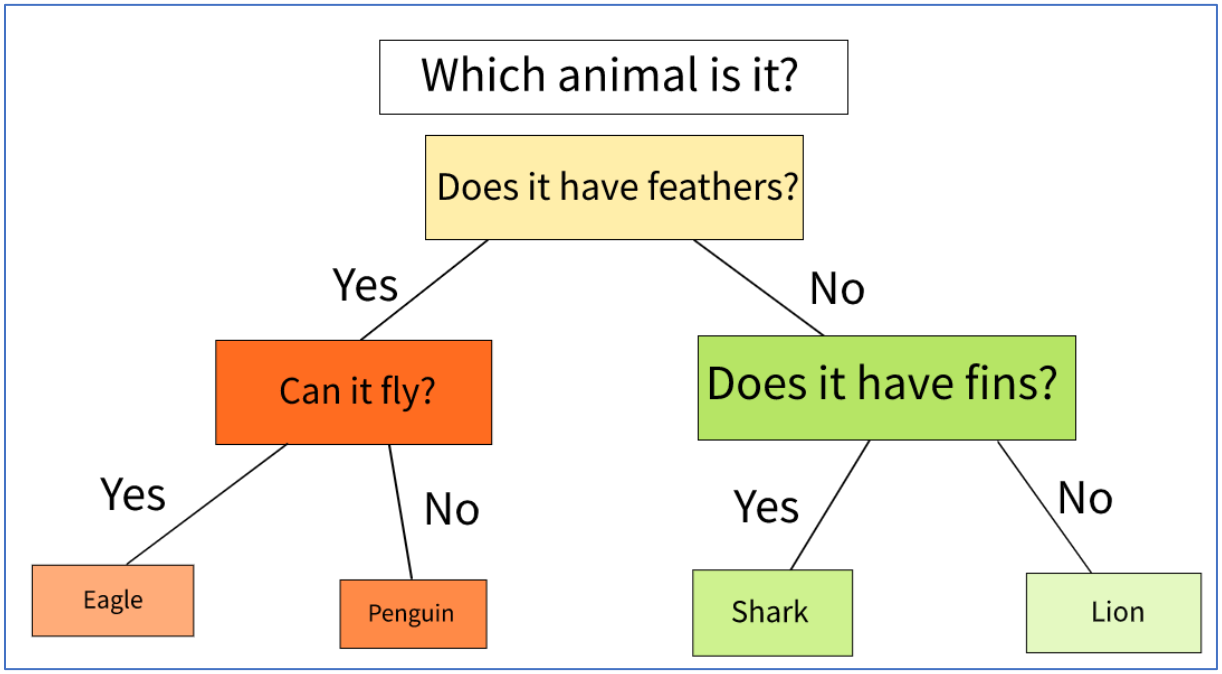


Figure 8: Simplified version of a decision tree to decide which animal a person sees. There are obivously more options than the four presented animals.

Decision trees incrementally split the data points into smaller divisions by analysing the data values, similar to how Figure 8 above exemplifies the decision process. At any node, the algorithm tries to divide the remaining samples into two groups with as little *error* (also called *impurity*) as possible. This is called Information Gain (IG), and the IG expression is presented here:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

where f is the feature used to split the data set, D_p is the dataset of the parent node, D_j is the dataset of the jth child node, I is the impurity measure, N_p is the total number of training samples in the parent node, and N_j is the total number of samples in the jth child node (Raschka and Mirjalili 2019). The IG is the difference between the impurity of the parent node and the child nodes. By minimizing the impurity of the child nodes, the IG of that split is maximized. Since most decision tree implementations use binary splitting, the summation above can be simplified to a left and right node:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

The impurity measure, I, can be either "*Gini impurity*", "*entropy*", or "*classification error*". The impurity measures to which degree each sample in a node belong to the same class. In this thesis Gini impurity and entropy will be used, and thus presented below.

Gini impurity is used to minimize the probability of misclassification (Raschka and Mirjalili 2019):

$$I_g(t) = \sum_{i=1}^{c} p(i|t)\big(1 - p(i|t)\big) = 1 - \sum_{i=1}^{c} p(i|t)^{\,2}$$

where $p(i|t)$ is the proportion of samples from class i in a node, t. The maximum I_g in a binary classification problem will be 0.5.

Entropy is used to maximize the mutual information in the decision tree and can be calculated with this expression (Raschka and Mirjalili 2019):

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t)$$

where $p(i|t)$ is the proportion of samples from class *i* in a node, *t*. The maximum I_g in a binary classification problem will be 1.0.

Random Forests also utilize “Bootstrap Aggregation”, a method for making multiple versions of a decision tree predictor and combining these decision trees to make an aggregated prediction (Breiman 1994). Breiman coined the term “Bagging” from bootstrap aggregation, and suggested that bagging can improve unstable models at the cost of interpretability (Breiman 1996). By using subsamples of the entire dataset, different models can be created from the same dataset. Random Forest takes this approach one step further by only considering a subsample of features when creating a decision tree. Therefore, random forests are less likely to overfit to the training data since the individual trees only see parts of the training set (Breiman 1996).

While random forest can overcome many of the challenges related to overfitting in decisions trees, they are less interpretable. Recent studies have presented heuristics and methods to improve the interpretability of random forest by creating a decision tree based on a decision forest (Sagi and Rokach 2020, Vidal and Schiffer 2020).

Chapter 3 Methodology

This chapter will cover the method used to produce the results found in the research and simulation performed in this thesis. There methodology will first be presented generally, and then each of the points in the general methodology will be explained more in-depth. Thus, the reader should be able to see the point of the different chapters.

3.1 Overall structure of methodology and CRISP-DM

Initially, the overall structure of the methods used in this thesis will be presented. Thereafter, the finer details of each element will be explained. The following Figure 9 shows a rough outline of the method:

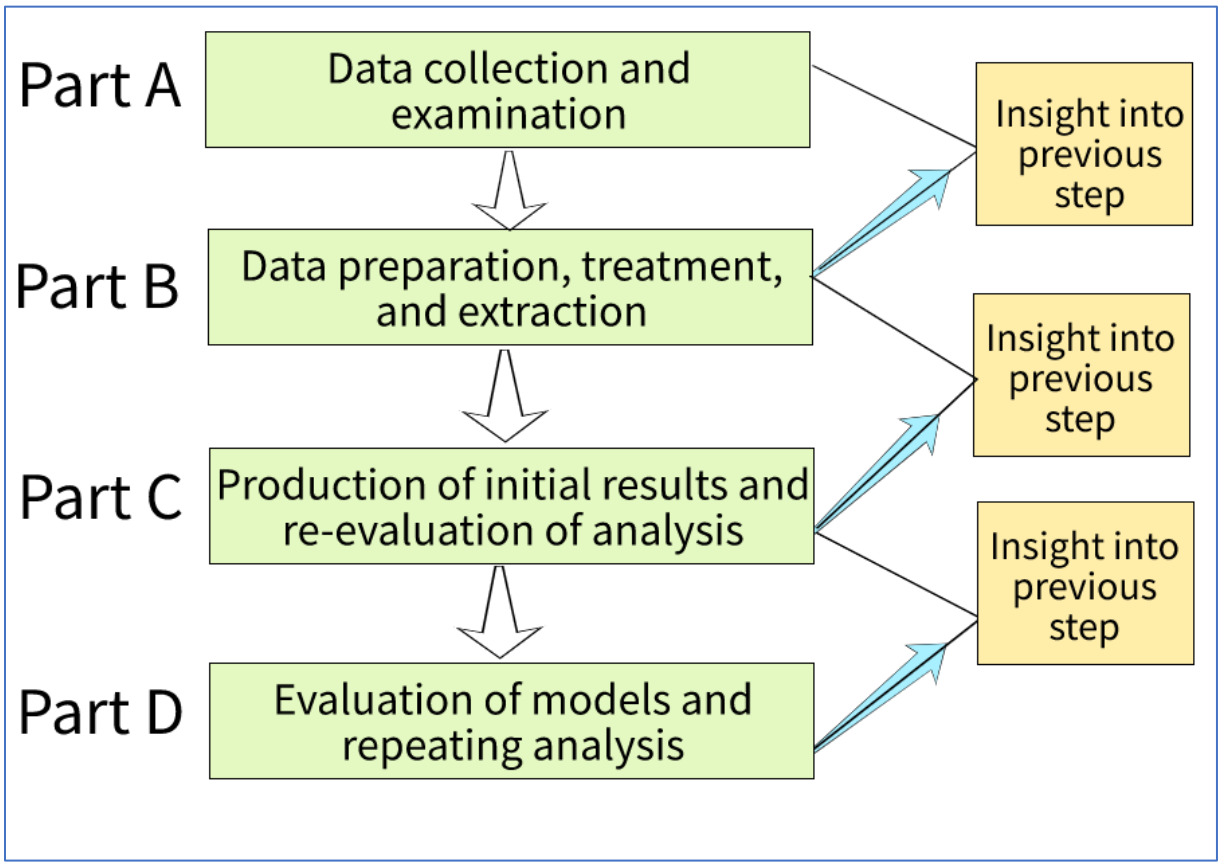


Figure 9: Rough outline of the method used in the analysis.

As the work progresses from step A through D, the insight into the models and analysis increases. Thus, this insight allows for making changes in the previous steps to improve the previous steps. As such, the methods is somewhat circular.

The methodology is partly inspired by CRISP-DM (*Cross-industry standard process for data mining*), which is an open standard for data mining (Chapman, Clinton et al. 2000). CRISP-DM is recognized as the most widely used model for data mining projects, although CRISP-DM has been critiqued for only describing "what to do" and not "how to do" (Mariscal, Marbán et al. 2010). While the bullet list above is divided into 4 parts, CRISP-DM uses the following 6 parts (Chapman, Clinton et al. 2000):

1. Business understanding
2. **Data understanding**
3. **Data preparation**
4. **Modelling**
5. **Evaluation**
6. Deployment

CRISP-DM's highlighted steps 2-5 are especially similar to steps A-D in the figure above. The following description is based on the following sources if the reader wants deeper explanation (Chapman, Clinton et al. 2000, Shearer 2000).

Business understanding refers to the preliminary work of getting into the project objectives and planning how to achieve the objectives. While this step is omitted from the methodology, it is a normal step in any thesis or similar long projects.

Data understanding includes examination of the data set, discovering data problems and identifying necessary steps towards preparing the data. The data understanding details relevant for this thesis is presented in chapter 3.2 *Part A: Data Collection and examination*.

Data preparation refers to all actions to prepare the data for modelling. The actions performed in this step is highly specific to the data set and the models created in the next step. The data preparation details performed in this analysis is covered in chapter 3.3 *Part B: Data pre-processing.*

Modelling includes creating the appropriate models, as well as tweaking model parameters to optimize them. Since the models sometimes require very specific data types or data structures, it is common to go back to "data preparation" to adjust these tasks. The ML models used in this thesis are largely detailed in theory chapter 2.8 *Machine learning algorithms*. However, RENT models are detailed in chapter 3.4.2 *Implementation of RENT*.

Evaluation of the models and their output, and therefore indirectly the data preparation and data understanding, may lead to changes in any of the previous steps. This means that CRISP-DM will often lead to a circular approach where the user will improve and tailor their process after achieving new insights into the data.

Deployment is the final step of CRISP-DM, and this step is also not covered in the methodology in this thesis. However, this thesis, in itself, is a deployment of the analysis and research that has been performed.

3.1.1 Programming language and extensions

Python 3.7 was used for the programming part of the research, with notable extensions: *NumPy* (Harris, Millman et al. 2020), *pandas* (The pandas development team 2022), *scikit-learn* (Buitinck, Louppe et al. 2011) are essential tools to assist ML in Python. Moreover, NMBU-created packages like *RENT*[1], *Hoggorm*[2], and *Imskaper*[3] were used for feature selection, plotting, and feature extraction, respectively. The code was mostly tested in *Jupyter Lab*, a browser-based interactive computational environment (Kluyver 2016). The reader may, for inspiration or any other reason, look up the code used in the analysis, following this GitHub link [4].

3.2 Part A: Data Collection and examination

The data was collected by the OxyTarget study (NCT number: *NCT01816607*). The study was started in October 2013, and was ended in December 2017 (Redalen 2018). Information about the patients' outcome have been collected through follow-up by general doctors, up to 5 years after the patients' ended treatment. A total of 192 patients were included and 7 of these patients had withdrawn their consent by the time the research in this thesis began. Considering the final admitted patients were recruited in 2017, most patients had either died or had their final 5-year follow-up by the start of 2022 and this thesis. Some patients ended their final follow-up on

[1] https://github.com/NMBU-Data-Science/RENT
[2] https://github.com/olivertomic/hoggorm
[3] https://github.com/NMBU-Data-Science/imskaper
[4] https://github.com/LarsEngeseth/Master-thesis

palliative care, and a few patients had not completed their 5-year follow-up but was registered alive in January of 2022.

3.2.1 Clinical data

This thesis has done research on clinical data and tumour images. In total, 185 patients did not retract their consent to be a part of the study. The inclusion data consists of 103 features. The features consisted of 7 different datatypes (excluding the NaN-type):

1. Integer
2. String
3. Floats
4. Python ‘datetime.datetime’-objects
5. NumPy ‘numpy.int64’-objects
6. NumPy ‘numpy.float64’-objects
7. Pandas ‘pandas._libs.tslibs.timestamps.Timestamp’-objects

Python’s datetime-objects can either be “aware” or “naïve”. Naïve objects do not contain information about their time zone, whereas aware objects contain that information. It is therefore recommended to use aware objects for comparing different datetimes, since their time difference is unambiguous (Python Software Foundation 2022). Dates in OxyTarget were all performed in Norway, and it is thus assumed that the dates are all from the same time zone.

The patients are measured on terms of progression-free survival (PFS). PFS is defined as, in the context of OxyTarget, the moment until the disease of a patient progressed as a local cancer recurrence, metastasis, or leads to death, although the definition of PFS may vary in other studies. The PFS event is registered as a binary event, i.e., there was no progression (0) or there was some progression (1). Time to the PFS event is measured for every patient that has an event before their year-5 check-up after their treatment. PFS is sometimes used as an alternative to Overall Survival (OS). A positive OS event is a patient that has died within a certain time, in this thesis 5 years, after treatment or diagnosis (National Cancer Institute s.a.). Some patients that are OS-positive may have died independently of their initial disease.

3.2.2 Medical images

Three sets of medical image modalities were used in this thesis: DWI, FFE, and T1T2Sense. These image modalities have been introduced in chapter 2.2.1 *Medical image types in this thesis*. The image files were stored as a Nifti (*Neuroimaging Informatics Technology Initiative*) images, a technology introduced in the beginning of the 2000s as an improvement to the "Analyze format" (Larobina and Murino 2014). The features of the images were extracted using "Imskaper", a tool created by previous master students from NMBU to extract features from images. Imskaper is based on the Python package "pyradiomics" (van Griethuysen, Fedorov et al. 2017). A patient cohort of 81 had image files for the three modalities. Each image modality yielded 104 features, which resulted in three datasets with 81 patients and 104 features.

3.3 Part B: Data pre-processing

Pre-processing is almost always needed in predictive analyses, and there are far more techniques available than mentioned in these chapters. These chapters will instead cover techniques applied in the thesis analysis.

3.3.1 Feature selected through pre-processing

As discussed in chapter 2.1 *Feature selection*, feature selection has several benefits. Yet, manually removing features would be suboptimal in this thesis. Medical professionals possess greater knowledge about informative medical measurements and could thus make a more qualified feature selection. The motivation to feature selection in this analysis is to approach the problem from a data science perspective, instead of a medical perspective. Thus, RENT was tuned to select the most optimal features without the need for using medical domain knowledge.

It should be noted that some feature selection has already been performed by the medical professionals, since they select the parameters to register for the research project. Any other measurement that could be relevant in outcome prediction will not be present in the analysis in this thesis. Furthermore, some feature exclusion was performed as part of the pre-processing of the dataset. For example, several features were too sparse to be included. Methods to treat missing data are covered in chapter 3.3.2.2 *Imputing missing data*. Figure 10 below shows a

graphical representation of how the available feature space shrinks for each step in the research process as some features are excluded.

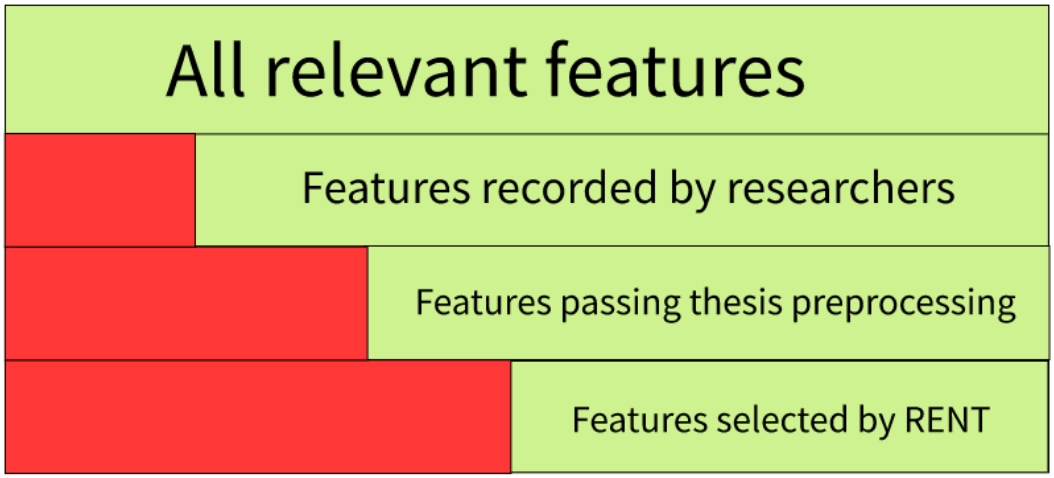


Figure 10: Graphical representation of how the feature space is reduced for each step on the research process. The rectangles are not to scale and should only serve as a tool to highlight that the bottom red rectangle contains potentially important information that are unavailable for the modelling algorithms.

3.3.2 Missing data

3.3.2.1 Types of missingness

When a patient lacks information about a certain feature, this is referred to as missing data. Missing data is problematic for ML because most models cannot be run on datasets with missing data. Missing data is divided into three subgroups:

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR; sometimes called 'NMAR' in literature)

MCAR means that there is no discernible pattern in the missing data, and MNAR suggests that there is a noticeable trend in the way the data is missing (Buuren 2021). Furthermore, MCAR imply that the data set is unbiased and that the reason for the missing values have nothing to do with the data. Although MCAR is convenient, data sets are rarely MCAR (Buuren 2021).

MAR implies that there is randomness in the missing values, but that the missingness is explained by the information available (Carpenter, Bartlett et al. s.a.). An example from the OxyTarget study would be missing surgery locations. The missingness comes from those patients did not have any surgery.

MNAR implies that there is some hidden and definite explanation for the missing values that is unrelated to any feature of the remaining data. The missing values would have explained why the values are missing, but they are not available to give context. MNAR is the most complex case of missing data, and strategies to handle data sets with MNAR is to investigate the cause of the missingness (Buuren 2021).

3.3.2.2 Imputing missing data

Dealing with missing data is generally split into two categories:

1. Removing samples with missing data OR removing features with too many missing samples
2. Imputing the missing data based on the gathered data

Removing missing data is the most basic approach to rectifying missing data. However, the research project this thesis is based on only contains 185 patients. Each removed sample (patient) substantially reduces the size of the data set. Additionally, each patient had at least one missing value, which means that the entire sample size would have been deleted if this method had been implemented. The reason for this is that several features contain information about the individualized treatment of each patient. Some of these treatments or examinations were only performed on a few patients, and the remaining patients have no information about this treatment in their journal. This may indicate that the patients did not receive this treatment, but the information may also be incorrectly not registered.

Yet it is still possible to remove columns from the data set. Columns with most of their values missing or values set to NaN will not be useful, since they cannot be imputed or otherwise kept without removing the patients with missing data. Therefore, columns with more than 25% of their values were removed.

Imputing the data is possible in some cases. This requires that the remaining values in those columns have high quality and are not too sparse. Missing data was imputed by the mean or the median of that column for numerical data or replaced with “MISSING” for categorical data. To use these features, they had to be one-hot-encoded into new columns.
Mean imputation is the most commonly technique used, but it may be inaccurate when the distribution of data in these features are skewed (Raschka and Mirjalili 2019). Thus, in each column where imputation was relevant it was investigated if the feature was skewed or not. If it was skewed, median imputation was chosen. Categorical columns, where the values are

binary, used mean imputation since median imputation would introduce skewness in these cases.

3.3.3 Encoding categorical values

Categorial nominal values, for example gender or tumour type, cannot be given to learning algorithms without mapping the values to dummy *variables* (from hereon: features). In this thesis, the ML standard of using dummy features was implemented (Garavaglia and Asha 1998). Since high dimensionality is a challenge in this thesis, one of the dummy features were in certain cases removed. This does not reduce the information from that feature, since any information in the removed dummy feature is contained in the combination of all the other dummy features (Raschka and Mirjalili 2019). Dropping this redundant information will also reduce multicollinearity in the model. However, the redundant feature was only removed if the category was binary. RENT selects the features it finds most important, but interpretation of the results is challenging when certain feature columns are removed. This challenge can be illustrated with an example: The feature "mucinous" in the dataset takes three values: "Yes", "No", "NA". One-hot-encoding creates 3 columns from this feature. If "No" was dropped and "NA" was selected as important, it would be difficult to interpret the implication of this finding. It could imply that missing information about "mucinous" was good indicator of patient survival. It could also mean that "not-mucinous" was the important information.

3.4 Part C: Making first models

The patients with DWI, T1T2Sense, and FFE images were grouped into a sub cohort of 81 patients. For these patients every image modality was available and could therefore be compared to each other, as well as to the clinical data. Each modality consists of several images taken at different echo times. Echo Time, also called *Time to Echo*, measures the time between the delivery of a pulse and the receival of the signal (Preston 2006). The images will either be T1 or T2 weighted, depending on the echo time. T1 and T2 weighting is sensitive to different tissue types (Preston 2006). For DWI, images from echo time 6 were selected. For T1T2SENSE, images from image time 30 and echo time 2 were used, and for FFE images echo time 5 was used. These echo times were suggested by the owners of the research project.

The cohort has a 36%/64% split between positive and negative outcomes after the patients had competed their 5-year check-up. Thereafter, the image features from all modalities were concatenated to form one combined data set. A predictive RF model for baselining was created for this dataset.

After these analyses had been performed, the dataset was analysed by RENT. As described in chapter 2.5 *RENT introduction*, RENT performs a feature selection. The reduced data set with the features chosen by RENT was given as training data for RF and LR and the performance was compared between all models to investigate the results. Further descriptions regarding the analysis after RENT has selected features can be found in chapter 3.4.2 *Implementation of RENT*.

3.4.1 Test set and train set, cross validation and RSKF

Data sets used to train ML algorithms are usually divided into subsets so that the algorithm can be trained on one subset and the algorithm's predictive performance can be tested on an unseen subset. The goal is to have low generalization error when the model is tested on unseen data. By training and testing on separate data sets, it is ensured that the model has no information about the test set, and the model can therefore give an estimate of the generalization error. However, the data set in this thesis was wide-short, which means that there are far more features than there are samples. Wide-short data sets often lead to unstable models, where the distribution of samples between the training set and the test set substantially influences the model's performance when predicting on the unseen data. A way to overcome the imbalance between train performance and test performance is to use "k-fold cross-validation". K-fold cross-validation" (CV) was introduced and formulated sometime during the 1900s, as explained in-depth by Stone (Stone 1974). K-fold CV has since then become a standard way to validate model performance and has been extensively covered in textbooks for machine learning and statistical learning (Smola and Vishwanathan 2008, James, Witten et al. 2013, Raschka and Mirjalili 2019).

When a model uses k-fold CV, it splits the dataset into *k* folds, and trains a model on *k-1* parts of the dataset and validates the model's performance against the remaining unseen fold. Specifically, for the analyses performed in this thesis, the dataset was split into 4 folds, which means that the model is trained on three folds and tested against the fourth fold, as visualized below in Figure 11. Additionally, using a 4-fold approach also means that the training and

validation happens 4 times with different test set for each split. The average performance across these 4 splits will give a more representative indication of the generalization performance.

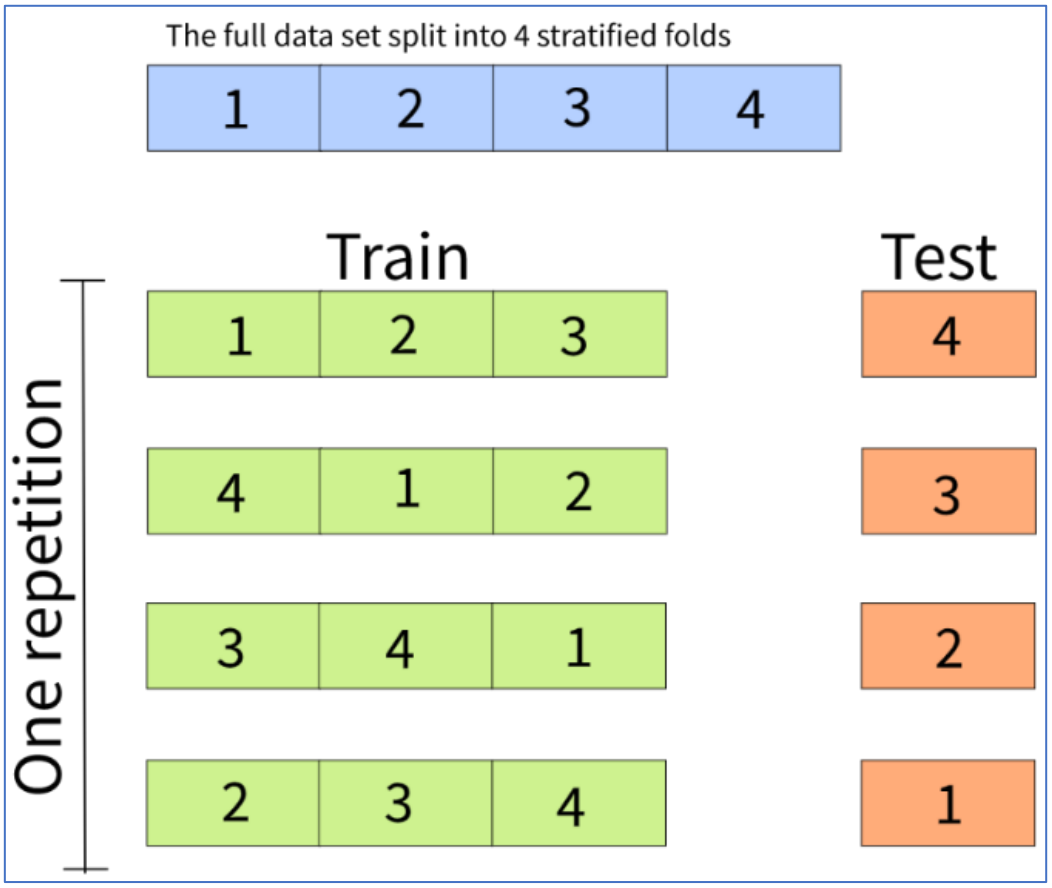


Figure 11: Example visualization of a 4-fold CV approach. Each fold is three times part of the training set and one time the test set. Note that the test-sets are sometimes referred to as validation-sets. Drawn with inspiration from (James, Witten et al. 2013, Raschka and Mirjalili 2019).

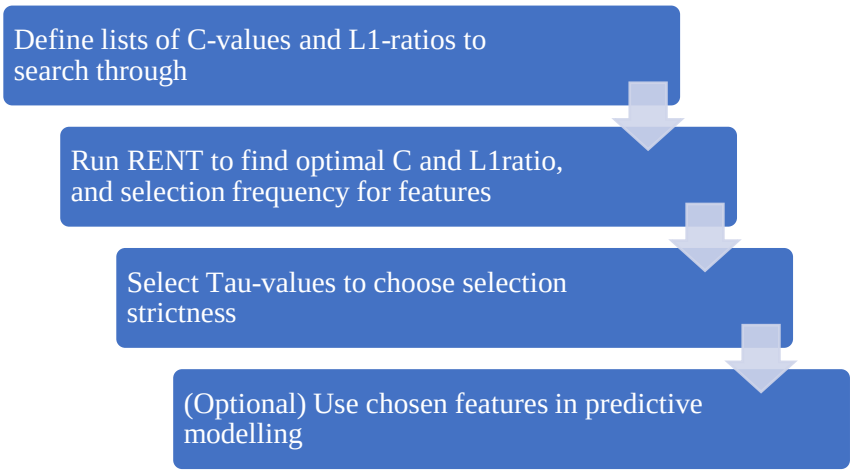
K-fold is slightly improved by ensuring that the folds are stratified, meaning that each fold has equal number of positive and negative samples (Raschka and Mirjalili 2019). Stratified sampling works by sampling each class to the same sub-cohort, so that the ratio between the samples remain intact after sampling (Särndal, Swensson et al. 1992). Stratifying the folds leads to lower bias and variance (Kohavi 1995). Interestingly, Kohavi also suggested that "repeated stratified k-fold", explained in the next paragraph, was a slightly stronger alternative to stratified k-fold (Kohavi 1995).

The k-fold approach can be repeated to ensure that the splitting is not an unrepresentative division of the data set. This approach is commonly referred to as a "repeated stratified k-fold (RSKF). RSKF repeats the stratified k-fold CV with different k-folds in each repetition (Raschka and Mirjalili 2019). These repeats may seem redundant, but graphs produced and

shown in result chapter 4.4 *Visual representation of features selected* shows the degree to which separate folds lead to dissimilar models. By performing a RSKF the model performance is further robust, since it is less likely that the k-fold split was unrepresentative of the actual generalization performance.

3.4.2 Implementation of RENT

The ordinary workflow of RENT can be represented as the following steps:



In addition to choosing C and L1-ratio, RENT also outputs other results if desired, e.g.: selection frequency plots, overview over rate of incorrect predictions for each patient, and overview for each feature's selection frequency. Selection frequency plots display the rate of which each feature is chosen among the *k* models in each RENT analysis, as presented in chapter 4.4 *Visual representation of features selected*. The overview of frequency of incorrect predictions per patient will also be presented chapter 4.5 *Hard-to-predict patients*.

To further provide confidence in the results, the RENT is evaluated with repeated stratified k-fold CV. Previous research involving RENT has applied RSKF CV with 4-folds and 2 repeats, meaning that the CV is computed twice with 4 folds. This thesis uses 4-fold CV with 5 repeats for the RENT analysis. As seen in Figure 11 in the previous chapter, there are 4 folds in each CV. The next repetition would have equal format for train and test, but the data samples in each of the four folds would be different. These repeats with different train-test splits ensures randomness is less likely to lead to extremely fortunate or unfortunate splits between train and test.

3.4.2.1 RENT input parameters

RENT has five model parameters that the user can adjust: γ, α, τ_1, τ_2, and τ_3. γ and α adjust the regularization strength and L1-ratio, while the tau-parameters are cut-offs for strictness of features selection (Jenul, Schrunner et al. 2021). The next paragraphs will explain each of these parameters.

The parameter **γ** refers to the regularization parameter C the learning model. C is the inverse of δ that was introduced in theory chapter 2.4.1 *Regularization*. In other words, **γ** = 1/ δ, and a low C-value will correspond to a strong regularization. RENT expects a list of C-values to find the optimal C, but it is possible to give a singular list-value. The analysis in this thesis forced RENT to evaluate only one C-value for each analysis.

The parameter **α** refers to the parameter L1-ratio in elastic net. The ratio determines the ratio between the L1 penalty and the L2 penalty (Raschka and Mirjalili 2019). RENT also expects a list of L1-ratios, but it is possible to give a singular list-value. The analysis in this thesis forced RENT to evaluate only one L1-ratio for each analysis.

τ_1 takes a value between 0 to 1 and serves as a cut-off value to the strictness of the feature selection. After elastic net regularization, the features are either 0 (i.e.: not selected) or non-zero. **τ_1** dictates that only features that are non-zero more than the selected percentages of times will be included. With **τ_1** set to 0.9, RENT will only output the features that were non-zero in more than in 90% of the *k* models trained on different subsets of the training data.

τ_2 takes a value between 0 and 1 and dictates the minimum proportion of feature weights that must be of the same sign. If a feature has weights that are non-zero and is selected frequently, but the sign of the weights vary within the k models, it could mean that this feature is less important than the features whose weights are always either positive or negative (Jenul, Tomic et al. 2021). The **τ_2** value decides the ratio of the non-zero feature weights that must be of the same sign for the feature to be eligible for selection. If **τ_2** is set equal to **τ_1**, then every feature weight must be of the same sign. This requirement is relaxed by setting **τ_2** smaller than **τ_1**.

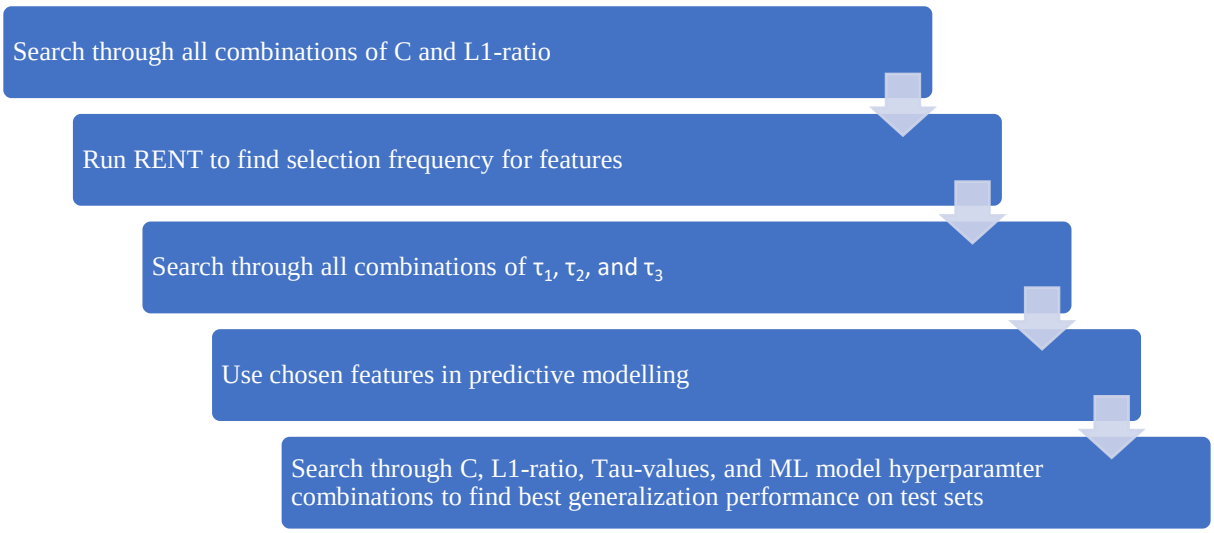
τ_3 is a standard students T-test to set the cut-off point where the null hypothesis is rejected, meaning that the average of the weight is significantly different from zero.

The three **τ** are adjusted to optimize the strictness of the model, i.e., the number of selected features. If the user wants more features, the τ should be reduced to an appropriate level.

3.4.2.2 Search through C and L1

In previous master theses, RENT has been tasked to find the optimal C and L1-ratio (Mohammadi 2021, Olofsson 2021). This thesis takes a different approach to RENT. There is a risk that RENT selects C and L1-ratio corresponding to a local optimum instead of the global optimum. The resulting analyses can lead to generalization errors that reduce the test performance of the results. Instead, this thesis aims to make a manual grid search across different C and L1-ratios to ensure that results from all combinations of these parameters are calculated fully. Five C-values [0.1, 0.5, 1, 5, 10] and five L1-ratios [0.1, 0.25, 0.5, 0.75, 0.9] were chosen, meaning each RENT analysis would be performed 25 times.

Additionally, the analysis in this thesis searched all possible combinations of the different $\boldsymbol{\tau_1}$, $\boldsymbol{\tau_2}$, and $\boldsymbol{\tau_3}$ values to find the most optimal combination of these parameters. Since the different tau values lead to a different selection of features, the resulting ML models will have different information available for training. Thus, the model's performance can be compared based on the selected hyperparameters. The workflow presented in chapter 3.4.2 *Implementation of RENT* has been updated and specified to the following workflow:



There were in total 450.000 different ML models computed in this brute-force analysis, where 300.000 are LR models and 150.000 are RF models. The combinations of 5 C-values, 5 L1-ratios, 5 repeats of 4 RENT models, 5 $\boldsymbol{\tau_1}$, 5 $\boldsymbol{\tau_2}$, and 2 $\boldsymbol{\tau_3}$ gives 25.000 combinations. These 25.000 combinations are each trained on 6 RF parameter combinations and 12 LR parameter combinations. In the first iteration of this analysis, only the average scores from the 6 RF combinations and the average scores for the 12 LR combinations were recorded, meaning that

the resulting dictionary to store all the results had 25.000 keys for storing all the results. This dictionary took 330 MB disk space. The complexity of this computation can also be multiplied with the k number of models in RENT, which was set to the default of k = 100. Additionally, each RF model contains 100 decision trees by default, which means that there are 100 decision trees for each of the 150.000 RF models. This results in a total of 25.000*100*(6*100 RF models) + 25.000*100*(12 LR models) = 1.530.000.000 models.

All of these combinations were computed in one go on a 7 year-old laptop with 16GB RAM and an "*Intel(R) Core(TM) i7-4710HQ CPU @ 2.50GHz*" processor. In total, the modelling finished after 14 hours computation time. Initially, the number of ML models were higher, but the parameter combinations were reduced as much as possible to make the computation finish within a reasonable time.

Heuristics could be used to lower the search space by investigating if any of the combinations of C and L1-ratio did not remove any of the features. In these cases, RENT has not performed any feature selection so computing RENT models for that combination is less likely to produce a desirable result. However, no C and L1-ratio combinations in this thesis resulted in such a situation, and the entire C and L1-ratio space was searched.

3.5 Part D: Evaluation of models and repeating analysis

Several steps went into evaluating the models. Firstly, the performance metrics for both training and testing was registered and compared to investigate indications of overfitting and underfitting, as well as general performance. Secondly, the average performance across different parameters was investigated to determine if any parameters had higher predictive accuracy. Thirdly, the selected features were analysed to both verify that clinical features proven relevant in literature were chosen, in addition to evaluating the selected image features. Fourthly, records were kept on the patients that were frequently predicted incorrectly. Quick evaluations of class distribution were performed to determine if any patterns were discernible for these patients. It was also briefly analysed if any of the features stood out among these patients that were frequently predicted incorrectly by the algorithms.

3.5.1 Making second models

Initially, the performance of both RF and LR was calculated as an average across the different hyperparameters for these classifiers. For the second iteration of this analysis, it was deemed appropriate to investigate the difference between RF models based on the hyperparameters. Registering every combination of parameters for RF meant that the already large dictionary of 25.000 values was increased 6-fold in size to 150.000 values. Thus, the second and more in-depth iteration was only done for RF, and not for LR. However, the computational burden was only reduced from 1530 million models to 1500 million models.

Chapter 4 Results

This chapter will present relevant results from the analysis. Firstly, the features selected from by the RENT analysis will be presented. Secondly, there will be presented results for the performing models. This chapter will also present the results regarding the best performing models. Finally, there will be presented a short analysis of the patients that were the most frequently predicted incorrectly.

4.1 Feature selection frequency from RENT

One of the outputs from the RENT analysis are the most often selected features. The features for each RENT model were counted across all 25 000 RENT parameter combinations. These features have been summarized below in Table 2.

Table 2: Table of the 15 most frequently selected features for all RENT models. The features are color coded: green for clinical, yellow for FFE, and blue for DWI.

Feature name	**Times selected (max: 25 000)**	**Rel. frequency (n= 21 878)**	**Abs. frequency (n= 25 000)**
Susp. metastatic lesions at diagnosis: Yes	20 294	95.6 %	81.2 %
Type of surgery: Hartmann	16 322	74.6 %	65.3 %
Adjuvant treatment: Yes	16 024	73.2 %	64.1 %
CEA (ug/L)	13 488	61.7 %	54.0 %
FFE: LBP_012	13 426	61.2 %	53.7 %
MR distance from anus to tumour	12 255	56.0 %	49.0 %
R classification_0.0	12 074	55.2 %	48.3 %
FFE: glcm_ClusterProminence_d_1	11 654	53.3 %	46.6 %
DWI: glcm_MCC_d_1	11 632	53.2 %	46.5 %
DWI: first_order_InterquartileRange	11 599	53.0 %	46.4 %
FFE: first_order_Median	11 285	51.6 %	45.1 %
p/ypN (TNM ed. 7): 0.0	11 021	50.4 %	44.1 %
Lymphocytes (10^9/L)	10 893	49.8 %	43.6 %
FFE: LBP_102	10 802	49.4 %	43.2 %
p/ypT (TNM ed.7): 4.0	10 581	48.4 %	42.3 %

Feature selection frequency can be considered in two ways: frequency to all RENT models (n=25 000) or frequency to all RENT models that resulted in any features (n=21 878). For certain τ_1 and τ_2 inputs, the RENT model returned zero features, and thus no machine learning was possible. Higher τ_1 and τ_2 input corresponds to stricter feature selection, and higher τ_1 and τ_2 had the highest frequency of "empty" ML models.

Among the 100 most frequently selected features there were 58 clinical features, 25 FFE features, 13 DWI features, and 5 T1T2Sense features. T1T2Sense images were not among the 15 most frequently selected features. The highest ranked T1T2Sense feature was the 68th most selected feature, and it was selected by 5665 of the 25 000 RENT parameter combinations.

The three least frequently chosen features reveal an interesting detail: The two least selected features were never selected in any of the models. The third least selected features were chosen 350 times. Thus, every feature except two in Table 3 below were at least chosen more than 349 times even though these features might not contain any predictive information. The two features that were never chosen contains zeros in each sample.

Table 3: The three least frequently selected features.

Feature name	**Times selected (max: 25 000)**	**Rel. frequency (n = 21 878)**	**Abs. frequency (n = 25 000)**
FFE: gldm_SmallDependenceEmphasis_d_1	350	1.6 %	1.4 %
DWI: first_order_Minimum	0	0.0 %	0.0 %
FFE: first_order_Minimum	0	0.0 %	0.0 %

Similarly, counting the selected features from the RENT analyses if τ_1 and τ_2 is set to 0.75 gives the following feature selection frequency:

Table 4: Feature selection frequency for τ_1 and τ_2 set to 0.75. There are 1000 possible combinations, although only 970 of these models contained at least one feature. The features are color coded: green for clinical, yellow for FFE, and blue for DWI.

Feature name	**Times selected (max: 1000)**	**Rel. frequency (n = 970)**	**Abs. frequency (n = 1000)**
Susp. metastatic lesions at diagnosis: Yes	956	99 %	96 %
Adjuvant treatment: Yes	806	83 %	81 %
Type of surgery: Hartmann	806	83 %	81 %
FFE: LBP_012	604	62 %	60 %
MR distance from anus to tumor	602	62 %	60 %
FFE: glcm_ClusterProminence_d_1	582	60 %	58 %
R classification: 0.0	574	59 %	57 %
CEA (ug/L)	572	59 %	57 %
DWI: first_order_InterquartileRange	550	57 %	55 %
DWI: glcm_MCC_d_1	534	55 %	53 %
FFE_ first_order_Median	530	55 %	53 %
FFE_ glcm_ClusterShade_d_1	508	52 %	51 %
Lymphocytes (10^9/L)	506	52 %	51 %
Blood type: B+	504	52 %	50 %
p/ypT (TNM ed.7): 0.0	504	52 %	50 %

Table 4 is one example of five possible τ_{12} -values. The other tables did not reveal any further insights. They are similar to the presented table, although the frequency of feature selection is lower overall due to stricter feature selection. These tables could also be created for each τ_1 or τ_2 individually, but these 25 tables have not been included.

4.2 Model performance

4.2.1 Average performance for RF and LR

The combined mean test performance scores for each RF and LR models are presented in Table 5 below. There total averages are computed from 150.000 RF models and 300.000 LR models. The RF model averages are computed from 6 hyperparameter combinations, and the LR model averages are computed from 12 hyperparameter combinations.

Table 5: Average test set performance for random forest and logistic regression. Each value is presented as the mean performance plus or minus 1 standard deviation.

	Test set performance				
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
Random Forest	0.69 ± 0.08	0.28 ± 0.23	0.44 ± 0.17	0.78 ± 0.07	0.62 ± 0.10
Logistic Regression	0.67 ± 0.08	0.25 ± 0.19	0.45 ± 0.15	0.76 ± 0.07	0.61 ± 0.09

MCC includes the highest absolute and relative SD. “F1 negative” is the F1-score for the scenario where the target label has been flipped, such that a patient being alive after 5 years of treatment is the positive class. The algorithms are more accurate at predicting these patients correctly.

The combined average train set performance scores for all RF and LR models are presented in Table 6 below:

Table 6: Average train set performance for random forest and logistic regression. Each value is presented as the mean performance plus or minus 1 standard deviation.

	Train set performance				
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
Random Forest	0.93 ± 0.08	0.84 ± 0.19	0.88 ± 0.16	0.95 ± 0.06	0.91 ± 0.11
Logistic Regression	0.89 ± 0.08	0.74 ± 0.19	0.79 ± 0.16	0.92 ± 0.06	0.85 ± 0.11

The standard deviation reveals that these averages are a result of varied models. Further investigation found that the best performing RSKF model had test accuracy of 76.0% ± 7.0%.

Individual RF models without CV can have test accuracy of 0.9 or when zooming in on these RSFK models.

Averaging performance for each τ_1-value reveals that the difference between different τ_1 values are small, as shown in Table 7 below:

Table 7: Table with average performance for different τ_1 -values. There are between 4638 and 3420 models in each of these average scores. Each performance metric has extra digits so it is possible to notice any difference between the scores.

RF test performance for different τ_1 -values					
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
τ_1 = 0.1	0.680	0.265	0.443	0.771	0.614
τ_1 = 0.3	0.682	0.268	0.445	0.772	0.615
τ_1 = 0.5	0.681	0.267	0.447	0.771	0.616
τ_1 = 0.75	0.680	0.269	0.457	0.768	0.619
τ_1 = 1.0	0.674	0.263	0.463	0.759	0.617

A similar pattern is found for LR models with the same τ_1-values. This table can be found in the Appendix in chapter 7.1.

The table above can be recreated by averaging across different τ_2-values. Table 8 below presents this overview:

Table 8: Table with average performance for different τ_2 -values. There are between 4638 and 3420 models in each of these average scores.

RF test performance for different τ_2 -values					
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
τ_2 = 0.1	0.680	0.264	0.440	0.771	0.613
τ_2 = 0.3	0.681	0.266	0.444	0.772	0.615
τ_2 = 0.5	0.681	0.267	0.446	0.771	0.615
τ_2 = 0.75	0.682	0.275	0.463	0.770	0.621
τ_2 = 1.0	0.673	0.260	0.462	0.758	0.616

The table highlights that the average performance for different τ_2 values does not alone define the model performance. A similar table for LR performance instead of RF performance can be found in the Appendix in chapter 7.1.

4.2.2 High performing parameters

Averaging performance across fewer models yields more variation within the performance metrics. There is a subset of 284 RSKFs that have RF test accuracy higher than 73.0%. Among these models, certain C-values, L1-ratios, τ_1 - and τ_2 -values can result in a higher number of high performing models on the test set. Table 9 below shows this distribution.

Table 9: This table shows which parameters are most prevalent in models with test accuracy higher than 73%. There are 20 repeated stratified k-folds in each of these averages. The table is split into 4 tables, each detailing the distribution of models for each parameter.

C-value	# of models
0.1	170
0.5	6
1.0	4
5.0	62
10.0	42

L1-ratio	# of models
0.10	68
0.25	146
0.50	56
0.75	12
0.90	2

.

τ_1	# of models
0.1	36
0.3	34
0.5	28
0.75	52
1.0	133

τ_2	# of models
0.1	45
0.3	42
0.5	38
0.75	60
1.0	106

4.2.3 Best RF performance

The best performing RSKF has the following combination of parameters:

C:5, L1-ratio: 0.1, τ_1:1, τ_2 :0.5, Tau3:0.995, RF_Max_depth:2, RF_criterion:'gini'.

L1-ratio at 0.1 means that the model used almost entirely L2-regularization (ridge regression). However, τ_1 =1 means that the features selected for these models were chosen in 100% of the RENT models.

There are 20 RENT models that use this combination of parameters, so the average performance is a result of different test/train-splits and selected features. The best performance can be seen in Table 10 below:

Table 10: Best train and test performance for RF. Each of the computed averages contain 20 different models, all of which have unique train-test-splits. Each value is presented with the average score with the corresponding standard deviation of that value.

	Best performing RF model				
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
Test	0.76 ± 0.07	0.47 ± 0.16	0.57 ± 0.13	0.83 ± 0.05	0.69 ± 0.08
Train	0.88 ± 0.03	0.74 ± 0.06	0.80 ± 0.05	0.91 ± 0.02	0.84 ± 0.04

Interestingly, the train performance is lower for the best RF performance than the train performance is for the average RF performance. The results also show that the standard deviation is lower in the train performance than in the test performance.

4.2.3.1 Feature selection frequency for best models

The most frequently selected features for the best performing models are presented in Table 11 below. The leading prefix (either "FFE" or "DWI") indicates which image set the feature is derived from. Only 3 of the 31 features are from DWI images, 11 are from FFE images, while the remaining 17 are from the clinical dataset. No features from T1T2Sense images were selected by this subset of models.

Table 11: Overview over the 31 selected features from the 20 models that created the best performing RF RSFK model. Missing place of surgery and missing type of surgery indicate that the patients did not have surgery. This often happened due to their health rapidly deterioration and their subsequent passing. There are 80 more features that were chosen at least 1 time but fewer than 5 times.

Feature name	**Number of selections (max: 20)**
CEA (ug/L)	20
Suspected metastatic lesions at diagnosis: **Yes**	19
Adjuvant treatment: **Yes**	18
FFE: LBP_012	16
DWI: glcm_MCC_d_1	13
FFE: first_order_90Percentile	13
Type of surgery: **Hartmann**	13
FFE: LBP_102	11
Place Surgery: **MISSING**	11
Type of surgery: **MISSING**	11
p/ypT (TNM ed.7): **MISSING**	11
p/ypN (TNM ed. 7): **MISSING**	11
R classification: **MISSING**	11
Lymphocytes (10ˆ9/L)	9
p/ypN (TNM ed. 7): **0.0**	9
p/ypT (TNM ed.7): **4.0**	8
FFE: glrlm_ShortRunHighGrayLevelEmphasis	7
Blood type: **B+**	7
FFE: first_order_Median	6
FFE: glrlm_GrayLevelVariance	6
FFE: glcm_ClusterProminence_d_1	6
mrT (TNM ed.7)	6
GT (U/L)	6
DWI: first_order_InterquartileRange	5
DWI: first_order_RobustMeanAbsoluteDeviation	5
FFE: glrlm_HighGrayLevelRunEmphasis	5
FFE: gldm_HighGrayLevelEmphasis_d_1	5
FFE: glcm_Autocorrelation_d_1	5
FFE: glcm_ClusterShade_d_1	5
Type of surgery: **Lower anterior resection (LAR)**	5
p/ypT (TNM ed.7): **3.0**	5

4.3 Modelling on full dataset without RENT

ML models were also trained without RENT to benchmark any performance improvement or performance reduction from utilizing RENT. RF alone performed worse than RF with RENT, but not significantly worse. The performance is summarized in Table 12 below:

Table 12: Performance for a RF model without feature selection from RENT. The average performances are computed from 120 models.

RF model without RENT					
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
Test	0.66 ± 0.08	0.21 ± 0.21	0.36 ± 0.17	0.77 ± 0.06	0.58 ± 0.08
Train	0.97 ± 0.04	0.94 ± 0.09	0.95 ± 0.07	0.98 ± 0.03	0.96 ± 0.06

The models are less stable than the models created using feature selection from RENT. The standard deviations are also always equal or greater for the performance of the models without RENT. The models computed for this comparison all used the same RF hyperparameters as the models computed with RF and RENT.

4.4 Visual representation of features selected

The following frequency plot in Figure 12 shows a graphical representation of how often a certain feature was chosen. As seen in the figure, RENT most frequently choose the rightmost features, which are the clinical features.

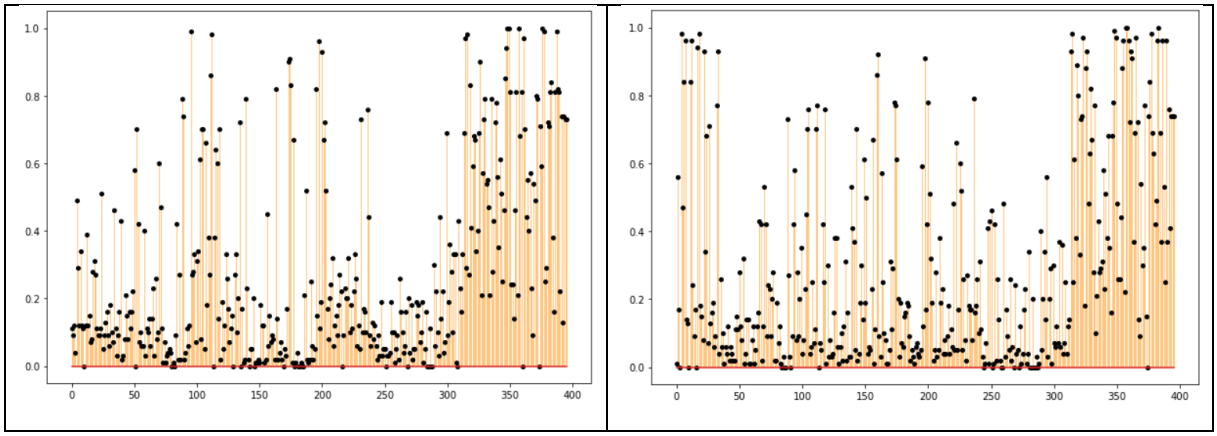


Figure 12: Frequency plots from the RENT analysis. The y-axis represents a percentage from 0% to 100%. Features that are always selected in every split will have a frequency of 100%. This RENT analysis used C=0.1

and L1-ratio=0.1, and the 2 plots have different train-test splits. Features are sorted left to right: DWI, FFE, T1T2Sense, Clinical.

Figure 12 above is created on the basis of the same C and L1-ratio, but the feature selection frequency differs due to different train-test splits. Thus, as $\boldsymbol{\tau_1}$ is changed for each of these selection frequencies, the features selected might differ for identical C and L1-ratio.

However, the feature selection frequency plot is altered by using different C and L1-ratios. By comparing the plots from Figure 12 and Figure 13, there are fewer features selected as the degree of regularization is changed, both in terms of C and L1-ratio.

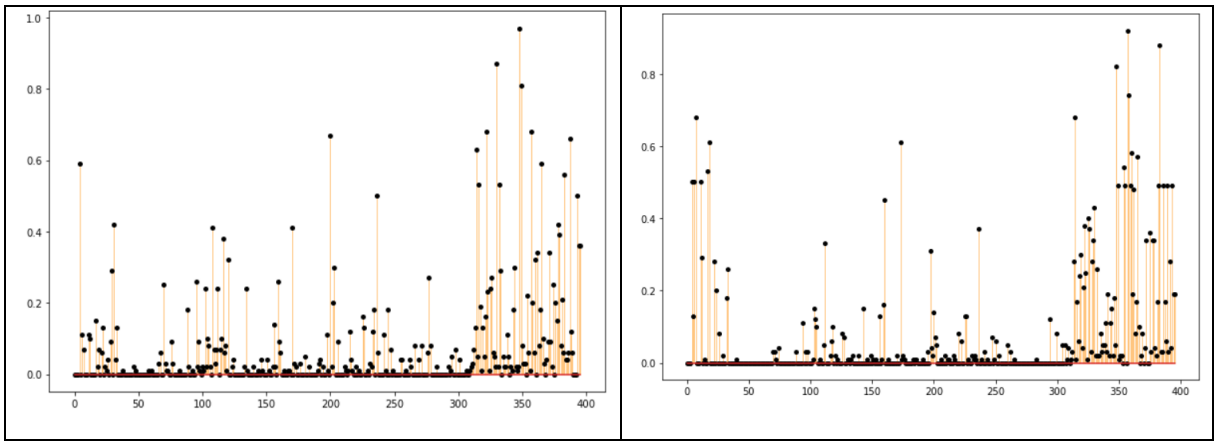


Figure 13: Frequency plots from the RENT analysis. The y-axis represents a percentage from 0% to 100%. Features that are always selected in every split will have a frequency of 100%. This RENT analysis used C=0.5 and L1-ratio=0.9, and the 4 plots have different train-test splits. Features are sorted left to right: DWI, FFE, T1T2Sense, Clinical.

4.5 Hard-to-predict patients

The outcome of certain patients was predicted incorrectly more often than other patients. The following table is sorted by the ratio of incorrect prediction. Patients who died (class 1) are in the majority in the group of patients that are frequently predicted incorrectly. Only 36% of the samples in the data set are class 1, while 73.3% of the 15 most frequently incorrectly predicted patients are class 1.

Table 13: Table highlighting which patients were most often incorrectly predicted after RENT analysis. Patients are randomly assigned to test set and train set, and there is therefore variance in the number of total incorrect predictions for any given patient.

Patient ID	Total incorrect predictions	Incorrect ratio (%)	Patient outcome
24	389 666	99.79	Dead
16	330 288	96.80	Dead
21	354 606	94.54	Dead
51	436 380	94.20	Dead
32	395 712	93.99	Dead
72	437 638	88.86	Dead
64	384 148	87.17	Dead
10	326 944	87.11	Dead
60	325 094	86.29	Alive
77	349 072	82.40	Alive
63	356 222	81.47	Dead
37	375 400	79.86	Dead
33	345 258	78.65	Dead
19	358 402	77.32	Alive
14	361 538	77.18	Alive

Chapter 5 Discussion

This section will provide a discussion surrounding the results and their implications. Firstly, the performance of the models will be benchmarked against similar work in literature. Secondly, there will be discussions about the potential overfitting by the average models. Thirdly, there will be a discussion regarding certain limitations of the results and the study. Finally, the features chosen by RENT will be compared to similar literature, to both validate the findings, as well as introducing radiomic features. Since the selected features have not been introduced in the theoretical background of this study, there will be provided several citations to literature during this discussion that the reader may use for further reading into the medical biomarkers.

5.1 Discussions on performance of models

The average RF and LR predictive performances are evaluated using accuracy, MCC, F1-score, and ROC-AUC. RF average test performance is slightly higher than LR average test performance (0.01 to 0.02 points). However, each of the performance metrics are presented with a corresponding standard deviation. Using only one standard deviation means that the RF and LR test performances are indistinguishable. The average F1-positive and F1-negative scores reveal that prediction of deceased patients is substantially harder for the algorithms than to predict the living patients.

The performance can be compared to similar studies to benchmark the performance as well as the methods. Recent studies have examined the relationship between CRC and the predictive ability of radiomic features, and they are in accordance with the findings in this thesis. The methods and data sets vary to some degree between these studies, but all the studies use feature selection before training their models so that the feature space is reduced. Badic et al. found that some positive correlations could be found between second order third order radiomic features and overall survival of CRC patients (Badic, Desseroit et al. 2019). Dai et al. computed the OS for CRC patients based on radiomic features with a test AUC of 0.768 (95% CI, 0.745-0.791, 10-fold CV) after using LASSO Cox to select the relevant features (Dai, Mo et al. 2020). Their results indicated that their selected radiomic features performed better than their clinicopathological features. Li et al. predicted the OS for patients with rectal cancer with clinical, radiomic, and clinicoradiologic features, and found that clinicoradiologic features had the highest test AUC of 0.802 (Li, Zhu et al. 2020). However, they applied only one split using

a 7:3 ratio between the train set and test set, instead of the more stable RSKF cross-validation performed in this thesis.

Nonetheless, the interest of exhaustive array modelling is not necessarily to find the average performance across all models, but to find the best performing model. The best performing model had test accuracy of 76.0 % ± 7 % and train accuracy of 88.0 % ± 3 %. Thus, the average model performance had test accuracy 7 percentage points lower than the best performing models. This result substantiates the importance of searching for a maximum in ML and AI. Since the best performing models is based on a 20 RSKF, the results were less likely to end up having high performance through a randomly advantageous split. Individual models have had test accuracy higher than 90%, but these numbers are not cross-validated, and cannot be reported as an individual result.

The benchmarking of a RF model without RENT provides further confidence in the results. The average RF model without feature selection performed worse than the RF model with RENT. Furthermore, its gap between test and train performance was larger, and multiple test-train splits had training accuracy of 100%. Overfitting like this is a well-known problem for high dimensional datasets, and this problem is confirmed in the analysis. On the other hand, performing the analysis without RENT could be computationally less demanding. The RF parameters would have 400 features to predict with, which could result in more complex computation if the number of decision trees was very large. An analysis without RENT would also not result in the feature analysis that can be done to examine feature importance.

5.2 Possible explanations for overfitting

The difference between the train and test performance of the models indicates that the models are on average overfit to the training samples. The difference between the RF models average test accuracy and train accuracy is 0.69 and 0.93, which is 24 percentage points (Table 5 and Table 6). While this was a surprising discovery, it has been reported in literature that RF can overfit (Segal 2003, Gashler, Giraud-Carrier et al. 2008), despite Breiman's claims when he introduced the RF technique (Breiman 1996). This difference in train accuracy and test accuracy is lower for the best RF models, 0.76 vs 0.88, which is 12 percentage points (Table 10). The models with the highest test accuracy are more generalizable, and these models also have lower train accuracy.

One explanation for this overfitting might be that the features provided by the RENT analysis are not optimal to explain the train set available to RF. If this assumption is true, it would mean that RF is limited to create a suboptimal model for the data samples. Naturally, RF can create a model designed to explain the train set, since the features provided are designed to explain the train set in particular. Thus, when the model is given new and unseen data, it will not be able to generalize well. One of RF's strength is its ability to decorrelate the features in the decisions trees by bagging. This ability is limited by limiting the available features. However, the results show that an RF model without RENT performs worse, so the feature selection might not be the issue.

Another explanation could be that the data set is difficult to predict, both due to its short-wide size and complexity. Therefore, each split in the CV is prone to overfitting due to the limited data available. Each train set consists of 60 samples, and by analysing the performance across different $\boldsymbol{\tau_1}$-values and $\boldsymbol{\tau_2}$-values, it seems that the strictness of feature selection is not alone an indicator of average model performance. Some evidence suggests that high $\boldsymbol{\tau_1}$ and $\boldsymbol{\tau_2}$ values (meaning stricter feature selection) lead to higher number of models with high test accuracy, but this effect may be due to a higher variance in performance within models with fewer features.

The analysis has also introduced some uncertainty in the modelling by using OS instead of PFS. This leads to some of the deceased patients being extremely hard to predict, since they may have died from something that is unrelated to their cancer. Results from the RENT analysis indicates that the class for the dead patients are predominant prevalent in list of the hardest-to-predict patients. This issue may introduce uncertainty in the models since each difficult sample substantially contributes to the data in each train set when the train set is only 60 samples long. There may be missing information explaining the outcome of these patients that is not available to the models.

5.3 Imputation and missing data

There are more complex and sophisticated ways (multiple imputation) to impute missing data, which may have yielded higher performance in the models (Raghunathan, Lepkowski et al. 2000, Rodgers, Jacobucci et al. 2021). Other methods, like MICE (van Buuren and Oudshoorn 2000, van Buuren 2007) and PPCA (Hegde, Shimpi et al. 2019) could also be considered. Recursive partitioning might also have yielded an improved performance (Doove, Van Buuren et al. 2014).

On the other hand, Morvan et al. argue that these imputation techniques have no theoretical grounding, and that their impute-then-regress methodology with powerful learners discover the true nature (Morvan, Josse et al. 2021). Perez-Lebel et al. argue that MICE is computationally intractable for large datasets (Perez-Lebel, Varoquaux et al. 2022). Furthermore, they argue that missingness is informative and that although the information is imputed, the imputation should be remembered in an "indicator column".

5.4 Selected features

There are two tables of selected features presented in the results (Table 2 and Table 11). Interestingly, several of the features chosen frequently across all models are not present in the best models. This finding indicates that these frequently selected features may not be the ideal features to achieve high test set predictive ability. Since the best performing models contain 20 different test/train-splits as a part of RSKF, there is indication that the best performing features are generalizable to the entire dataset. This section will present published research as validation of the findings from Table 11, since these features predict the most accurately on test data. Certain details of these biomarkers and features are medical. Discussion surrounding the medical details should be performed by medical professionals. Citations in the text will lead the reader to sources for further reading.

5.4.1 Comparing selected clinical features to published literature

The feature named 'CEA' refers to *carcinoembryonic antigen*, a protein usually found in developing baby tissue, that is also a CRC biomarker (Hall, Clarke et al. 2019). While the systematic review from Hall et al. found that CEA is associated with advanced disease, they

also found that CEA was not sufficient alone to predict OS. The findings that CEA is an important feature remains in agreement with other published literature (Chu, Erickson et al. 1991, Thirunavukarasu, Talati et al. 2015). CEA levels has also been used as an indicator to predict patient response to ramucirumab treatment (Yoshino, Obermannová et al. 2017).

Metastasis is a well-known indicator of poor prognosis for CRC patients (Kennecke, Yu et al. 2014). Additionally, Wang et al. found that the location of metastasis significantly influenced patient OS (Wang, Li et al. 2020). The location of metastasis was recording for this study, but the feature was so sparse that it was excluded from the analysis.

Adjuvant treatment is given to patients after their operation to decrease the probability of cancer recurrence (Carrato 2008). Adjuvant treatment was given to 4 patients in the cohort from this analysis, and they were alive 5 years after treatment. Since there are only 4 patients that underwent this treatment, they all had the same outcome after treatment, this feature is an excellent predictor for this dataset. However, this feature may not be generalized to perform equally well on other data sets, since it may be a spurious correlation.

Several of the selected clinical features are related to the patients' treatment, e.g.: place of surgery, type of surgery, p/ypT staging, and p/ypN staging. Whenever these features were "missing", it usually meant that the patient died before surgery was performed (p/ypN and p/ypT staging is performed after surgery). Therefore, the features serve as an indicator for the algorithm to predict patient outcome. As such, the results do not mean that patients have a better prognosis if the doctors do not to operate. It could be argued that the selection of features related to treatment may lead to over-optimistic models, since the treatment is specific and individualized for each patient based on the clinical factors of the patient. Regardless, these findings build confidence in the results of the feature selection. The appropriate treatment will obviously decide how likely it is that a patient survives the treatment.

GT (Gamma Glutamyl transferase) was selected in 6 out of 20 models, and this biomarker has previously been reported as relevant to predict OS for CRC patients (He, Guo et al. 2013). Among the 5 patients with GT > 100, 4 died before their 5-year check-up, which could explain the predictive ability of the feature in this data set. GT has also been used as a predictor of patient outcome in CRC with liver metastasis (Xiao, Peng et al. 2020). GT is the feature selected in the analysis that has the least documentation in literature as a biomarker for CRC, although GT has been used as an indicator of liver disease and other cancer metastases.

Lymphocytes are a type of white blood cells that determine the immune response to foreign substances (Britannica 2020). In the best performing models, lymphocytes were selected 9 out of 20 times (Table 11). Lymphocyte count has shown some association with chemotherapy response (Cézé, Thibault et al. 2011, Liang, Zhu et al. 2016) and cancer recurrence and death (Chu-Yuan, Jing et al. 2013).

The RENT analysis indicates that B+ is an important factor to predict OS in CRC patients. This correlation could not be confirmed by published literature. Some studies have investigated the relationship between blood groups and risk of CRC (Huang, Wang et al. 2017, Kashfi, Bazrafshan et al. 2018). One plausible explanation for the finding in this thesis is that the feature is spuriously correlated to survival. There are only six B+ patients in the cohort, and they were all alive 5 years after their treatment. Thus, blood type B+ is an accurate predictor for this data set in particular, but the feature could underperform in other data sets.

5.4.2 Frequently selected radiomics features

Considering the clinical findings are supported by literature, there is more trust in the results regarding the radiomic features. Interestingly, T1T2Sense-features were not chosen frequently by the best performing model or the average performing model. The images chosen for T1T2Sense may be from a low-information echo, or maybe these images are less informative overall to predict OS in CRC patients. The ratio between DWI and FFE features reveal that the FFE features have a higher predictive ability for these patients. Again, since there are unused echo times for DWI and FFE, the results may be different if other echo times had been used as data input.

Interpretation of these features could be done by analysing the images in conjunction with the features extracted. Insight may be difficult to extract from this analysis, because the meaning of feature descriptions like “Cluster Prominence” may be challenging to understand as a human. One of the strengths of radiomics is that the algorithms find patterns hidden from the human perception. Therefore, it may be unprofitable to try to interpret and translate these features to human understanding. Descriptions and further reading regarding the different types of features extracted by PyRadiomics can be found in the documentation for PyRadiomics if the reader is interested [5] (van Griethuysen, Fedorov et al. 2017).

[5] https://pyradiomics.readthedocs.io/en/latest/

5.5 Importance of cross-validation and search techniques

The difference between the best models and the average models are substantial. This puts emphasis on the importance of using heuristics and search techniques to find the optimal models for any situation. The average model is barely better than a "dumb classifier", and several individual models have AUC < 0.5, which is worse than guesswork. The results also revealed that the best performing models had lower train performance, further indicating that these models are less overfit, in addition to having higher test performance. Overfitting and underfitting remains relevant, even after performing a features selection.

The overall spread of performance across different parameter combinations also supports the importance of using careful and thorough cross-validation to reduce the influence from randomness on the model performance. The local maximum test accuracy is higher than 90%, but it would not be realistic to expect this model to perform equally well on new data, since the high performance is likely due to an advantageous train-test-split. Other research should be equally careful to confirm that their findings are valid.

Likewise, the resulting average feature selection and the best performing feature selection differs. The feature selection should be evaluated with the models that has the smallest difference between train and test performance. In this thesis, that difference is the smallest for the models with the highest accuracy.

Chapter 6 Conclusion

6.1 Summary of thesis

The goal of this thesis was to test a new brute-force approach to feature selection with RENT to achieve more stable and generalizable machine learning models for cancer patients. The results indicate that RF models with RENT have higher predictive ability and lower degree of overfitting compared to RF models without RENT. The best models by test performance had average accuracy 0.76 ± 0.07, MCC 0.47 ± 0.16, F1 positive class 0.57 ± 0.13, F1 negative class 0.83 ± 0.05, and AUC 0.69 ± 0.08, which is comparable to similar studies reported in literature. The corresponding train performance was accuracy 0.88 ± 0.03, MCC 0.74 ± 0.06, F1 positive class 0.80 ± 0.05, F1 negative class 0.91 ± 0.02, and 0.84 ± 0.04, which indicates that there is some degree off overfitting and challenges to generalize the models, despite implementation of feature selection with RENT. This error is larger for a similar RF model in which no feature selection by RENT was performed.

The analysis in this thesis used a brute-force approach to search through 450.000 model combinations. The average models have considerably worse performance than the performance of the best models. This highlights the importance of using search techniques to improve model performance. Additionally, this thesis used extensive validation of the results to build stable models with high confidence in the results. The variance in the model performances has proved to be so high that it suggests that other research must be equally vigilant and careful to not report fortunate local maxima. Furthermore, this brute-force framework has been tested in small-scale with low processing power, which means that the framework has proven it's usefulness to scale up. This framework could also be applicable for machine learning in other similar medical situations, and potentially in other research topics.

All the frequently selected clinical features can be supported by published literature. The radiomic features with the highest predictive ability have also been reported in the results, but these features have not been confirmed in other literature yet. These features provide a basis to perform further research and may aid prognosis for CRC patients in the future. Considering the sample size is small in this analysis (n=81), one should be careful to generalize the findings in regarding important features without conducting further research on a larger sample size.

6.2 Suggestions for future work

6.2.1 PCA and unsupervised learning on difficult patients

There may be clusters of different patients that could be revealed from using unsupervised learning methods like clustering or PCA. This analysis might reveal that certain dead patients are different from other dead patients. Could this confirm our suspicion that “dead” is too coarse/rough? The results indicated that patients that are the most frequently incorrectly predicted also have an unusual high rate of dead patients. Future work should investigate if there is a pattern to these patients, for example by using PCA with a union of the features that RENT have selected. Thus, the analysis could reveal if there is a pattern to these patients that is not detected through the analysis in this thesis. The PCA could also reveal that there is not a pattern to these patients, and that the relevant information is not available. Determining this information is crucial for the medical community, so that they may more accurately take the relevant history of cancer patients.

6.2.2 Expansion of the framework introduced in breadth/depth

This framework can be expanded upon in both breadth and depth, and it should be! It serves as a proof of concept that has a potential to dig thoroughly through a small dataset to find an optimal combination of parameters, given a powerful enough processor. This research compressed the result to 450.000 model entries. This resulting dictionary was built so that accessing the results and slicing them was convenient with *for-loops.* Thus, expanding the analysis in breadth or depth is a matter of expanding the search space and devoting more computational resources to the analysis.

The three image modalities used in this thesis holds six, five, and 180 echo times. The other echo times may include other predictive information. Since only one echo time was used for each of the image modalities, other features may be selected by RENT if other echo times are analysed. Alternatively, such an analysis could strengthen our results by reselecting the same image features. Potentially, these finding may be compared to other radiomic research to investigate if any features are generalizable to all CRC patients. Especially T1T2Sense performed poorly overall, and the image modality might perform better by changing the echo time.

6.2.3 Improve interpretability by removing features describing treatment

Both Table 2 and Table 11 describes the rate of which features were selected. Some of these features describe the treatment (or the lack of treatment) a patient received. These features are useful to predict the outcome of the patient because they are derived from the state of the patient. However, knowing that these features have high predictive ability is not information that doctors can act upon in their treatment and prognosis for patients, since the patients have already received treatment at this point. These features are also correlated since they explain the same treatment information. If a study aimed to give doctors further confidence in their or find new biomarkers, it could be more useful to exclude the features derived from the given treatment. An analysis as such could potentially result in higher interpretability of the resulting features.

Chapter 7 Appendix

7.1 LR test performance for different Tau-values

LR test performance for different τ_1 -values					
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
τ_1 = 0.1	0.669	0.252	0.452	0.755	0.613
τ_1 = 0.3	0.670	0.253	0.452	0.755	0.613
τ_1 = 0.5	0.671	0.255	0.453	0.757	0.614
τ_1 = 0.75	0.673	0.260	0.454	0.759	0.616
T τ_1 au1 = 1.0	0.68	0.256	0.417	0.773	0.610

LR test performance for different τ_2 -values					
	Accuracy	**MCC**	**F1 positive**	**F1 negative**	**ROC AUC**
τ_2 = 0.1	0.669	0.252	0.452	0.755	0.613
τ_2 = 0.3	0.669	0.252	0.452	0.755	0.613
τ_2 = 0.5	0.671	0.255	0.452	0.757	0.614
τ_2 = 0.75	0.673	0.259	0.454	0.759	0.615
τ_2 = 1.0	0.682	0.257	0.418	0.774	0.610

References

Aerts, H. J., E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies and P. Lambin (2014). "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." Nat Commun **5**: 4006.

Aerts, H. J. W. L., E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies and P. Lambin (2014). "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." Nature Communications **5**(1): 4006.

American Cancer Society. (2020). "What Is Colorectal Cancer?" Retrieved 24.04.2022, from https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html.

American Cancer Society. (2022). "Survival Rates for Colorectal Cancer." Retrieved 24.04.2022, from https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html.

Badic, B., M. C. Desseroit, M. Hatt and D. Visvikis (2019). "Potential Complementary Value of Noncontrast and Contrast Enhanced CT Radiomics in Colorectal Cancers." Academic Radiology **26**(4): 469-479.

Balagurunathan, Y., V. Kumar, Y. Gu, J. Kim, H. Wang, Y. Liu, D. B. Goldgof, L. O. Hall, R. Korn, B. Zhao, L. H. Schwartz, S. Basu, S. Eschrich, R. A. Gatenby and R. J. Gillies (2014). "Test–Retest Reproducibility Analysis of Lung CT Image Features." Journal of Digital Imaging **27**(6): 805-823.

Barentsz, J. O., J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager and J. J. Fütterer (2012). "ESUR prostate MR guidelines 2012." Eur Radiol **22**(4): 746-757.

Bellman, R., R. E. Bellman and R. Corporation (1957). Dynamic Programming, Princeton University Press.

Bousquet, O. and A. Elisseeff (2002). "Stability and Generalization." Journal of Machine Learning Research **2**: 499-526.

Breiman, L. (1994). Bagging Predictors, Department of Statistics, University of California Berkeley.

Breiman, L. (1996). "Bagging predictors." Machine Learning **24**(2): 123-140.

Britannica, T. E. o. E. (2020). "Lymphocyte." Retrieved 07.06.2022, from https://www.britannica.com/science/lymphocyte.

Brown, G., A. Pocock, M.-J. Zhao and M. Luján (2012). "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection." The Journal of Machine Learning Research **13**: 27-66.

Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller and O. Grisel (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research **12**: 2825-2830.

Buuren, S. v. (2021). Flexible Imputation of Missing Data, Chapman and Hall/CRC.

Bøvelstad, H., S. Nygård, H. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi and O. Lingjærde (2007). "Predicting Survival from Microarray Data – a Comparative Study." Bioinformatics (Oxford, England) **23**: 2080-2087.

Carpenter, J., J. Bartlett and M. Kenward. (s.a.). "Missing At Random (MAR)." Retrieved 24.01.2022, from https://web.archive.org/web/20150910180057/http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=76:missing-at-random-mar&catid=40:missingness-mechanisms&Itemid=96.

Carrato, A. (2008). "Adjuvant treatment of colorectal cancer." Gastrointestinal cancer research : GCR **2**(4 Suppl): S42-S46.

Cézé, N., G. Thibault, G. Goujon, J. Viguier, H. Watier, E. Dorval and T. Lecomte (2011). "Pre-treatment lymphopenia as a prognostic biomarker in colorectal cancer patients receiving chemotherapy." Cancer Chemother Pharmacol **68**(5): 1305-1313.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth. (2000). "CRISP-DM 1.0." Step-by-step data mining guide Retrieved 24.05.2022, from https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72.

Chu-Yuan, H., P. Jing, W. Yi-Sheng, P. He-Ping, Y. Hui, Z. Chu-Xiong, L. Guo-Jian and W. Guo-Qiang (2013). "The impact of chemotherapy-associated neutrophil/ lymphocyte counts on prognosis of adjuvant chemotherapy in colorectal cancer." BMC Cancer **13**: 177.

Chu, D. Z., C. A. Erickson, M. P. Russell, C. Thompson, N. P. Lang, R. J. Broadwater and K. C. Westbrook (1991). "Prognostic significance of carcinoembryonic antigen in colorectal carcinoma. Serum levels before and after resection and before recurrence." Arch Surg **126**(3): 314-316.

COCIR. (s.a.). "Medical Imaging." Retrieved 12.05.2022, from https://www.cocir.org/our-industry/medical-imaging.html.

Cramer, J. S. (2002). "The Origins of Logistic Regression." Tinbergen Institute, Tinbergen Institute Discussion Papers.

Dai, W., S. Mo, L. Han, W. Xiang, M. Li, R. Wang, T. Tong and G. Cai (2020). "Prognostic and predictive value of radiomics signatures in stage I-III colon cancer." Clinical and translational medicine **10**(1): 288-293.

Decourselle, T., J.-C. Simon, L. Journaux, F. Cointault and J. Miteran (2012). Noise Robustness of a Texture Classification Protocol for Natural Leaf Roughness Characterisation.

Demirel, H. C. and J. W. Davis (2018). "Multiparametric magnetic resonance imaging: Overview of the technique, clinical applications in prostate biopsy and future directions." Turk J Urol **44**(2): 93-102.

Doove, L. L., S. Van Buuren and E. Dusseldorp (2014). "Recursive partitioning for missing data imputation in the presence of interaction effects." Computational Statistics & Data Analysis **72**: 92-104.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern Recognition Letters **27**(8): 861-874.

Garavaglia, S. and S. Asha (1998). "A smart guide to dummy variables: Four applications and a macro." Proceedings of the northeast SAS users group conference **43**.

Gashler, M., C. Giraud-Carrier and T. Martinez (2008). Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. 2008 Seventh International Conference on Machine Learning and Applications.

Gillies, R. J., P. E. Kinahan and H. Hricak (2016). "Radiomics: Images Are More than Pictures, They Are Data." Radiology **278**(2): 563-577.

Gillies, R. J., P. E. Kinahan and H. Hricak (2016). "Radiomics: Images Are More than Pictures, They Are Data." Radiology **278**(2): 563-577.

Grossmann, P., O. Stringfield, N. El-Hachem, M. M. Bui, E. Rios Velazquez, C. Parmar, R. T. Leijenaar, B. Haibe-Kains, P. Lambin, R. J. Gillies and H. J. Aerts (2017). "Defining the biological basis of radiomic phenotypes in lung cancer." Elife **6**.

Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." Journal of machine learning research **3**(Mar): 1157-1182.

Guyon, I. and A. Elisseeff (2003). "An Introduction to Variable and Feature Selection." Journal of Machine Learning Research **3**: 1157-1182.

Hall, C., L. Clarke, A. Pal, P. Buchwald, T. Eglinton, C. Wakeman and F. Frizelle (2019). "A Review of the Role of Carcinoembryonic Antigen in Clinical Practice." Annals of coloproctology **35**(6): 294-305.

Hanley, J. A. and B. J. McNeil (1983). "A method of comparing the areas under receiver operating characteristic curves derived from the same cases." Radiology **148**(3): 839-843.

Harrell, F. E., K. L. Lee and D. B. Mark (1996). "MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS." STATISTICS IN MEDICINE **15**: 361-387.

Harris, C. R., K. J. Millman and S. J. van der Walt (2020). "Array programming with NumPy." Nature **585**: 357–362.

He, W. Z., G. F. Guo, C. X. Yin, C. Jiang, F. Wang, H. J. Qiu, X. X. Chen, R. M. Rong, B. Zhang and L. P. Xia (2013). "Gamma-glutamyl transpeptidase level is a novel adverse prognostic indicator in human metastatic colorectal cancer." Colorectal Dis **15**(8): e443-452.

Hegde, H., N. Shimpi, A. Panny, I. Glurich, P. Christie and A. Acharya (2019). "MICE vs PPCA: Missing data imputation in healthcare." Informatics in Medicine Unlocked **17**: 100275.

Hilt, D. E. and D. W. Seegrist (1977). "Ridge, a computer program for calculating ridge regression estimates." Research Note NE-236.

Ho, T. K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition.

Ho, T. K. (1998). "The random subspace method for constructing decision forests." IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(8): 832-844.

Hosmer Jr., D. W., S. Lemeshow and R. X. Sturdivant (2013). Applied Logistic Regression, Third Edition.

Hua, J., Z. Xiong, J. Lowey, E. Suh and E. R. Dougherty (2004). "Optimal number of features as a function of sample size for various classification rules." Bioinformatics **21**(8): 1509-1515.

Huang, J. Y., R. Wang, Y. T. Gao and J. M. Yuan (2017). "ABO blood type and the risk of cancer - Findings from the Shanghai Cohort Study." PLoS One **12**(9): e0184295.

Hughes, G. (1968). "On the mean accuracy of statistical pattern recognizers." IEEE Transactions on Information Theory **14**(1): 55-63.

Huilgol, P. (2020). "Bias and Variance in Machine Learning – A Fantastic Guide for Beginners!" Retrieved 13.06.2022, from https://www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning/.

Hush, D. R. and B. G. Horne (1993). "Progress in supervised neural networks." IEEE Signal Processing Magazine **10**(1): 8-39.

Höhne, K. H., H. Fuchs and S. M. Pizer (1990). 3D Imaging in Medicine, Algorithms, Systems, Applications, Springer.

James, G., D. Witten, T. Hastie and R. Tibshirani (2013). An Introduction to Statistical Learning, Springer.

Jenul, A., S. Schrunner, B. N. Huynh and O. Tomic (2021). "RENT: A Python package for repeated elastic net feature selection." Journal of Open Source Software **6**(63): 3323.

Jenul, A., S. Schrunner, K. H. Liland, U. G. Indahl, C. M. Futsæther and O. Tomic (2021). "RENT—Repeated Elastic Net Technique for Feature Selection." IEEE Access **vol. 9**: 152333-152346.

Jenul, A., O. Tomic and B. N. Huynh. (2021). "RENT applied to a binary classification problem." Retrieved 21.04.2022, from https://github.com/NMBU-Data-Science/RENT/blob/master/examples/Classification_example.ipynb.

Juliani, C. and S. Ellefmo (2019). "Prospectivity Mapping of Mineral Deposits in Northern Norway Using Radial Basis Function Neural Networks." Minerals **9**.

Kadir, T. and F. Gleeson (2018). "Lung cancer prediction using machine learning and advanced imaging techniques." Transl Lung Cancer Res **7**(3): 304-312.

Kashfi, S. M., M. Bazrafshan, S. H. Kashfi and A. Khani Jeihooni (2018). "The Relationship Between Blood Group and Colon Cancer in Shiraz Namazi Hospital During 2002 - 2011." **7**(1): e59474.

Kennecke, H., J. Yu, S. Gill, W. Y. Cheung, C. D. Blanke, C. Speers and R. Woods (2014). "Effect of M1a and M1b category in metastatic colorectal cancer." Oncologist **19**(7): 720-726.

Kleppe, A., O.-J. Skrede, S. De Raedt, K. Liestøl, D. J. Kerr and H. E. Danielsen (2021). "Designing deep learning studies in cancer diagnostics." Nature Reviews Cancer **21**(3): 199-211.

Kluyver, T. e. a. (2016). "Jupyter Notebooks – a publishing format for reproducible computational workflows." Positioning and Power in Academic Publishing: Players, Agents and Agendas.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc.**:** 1137–1143.

Larobina, M. and L. Murino (2014). "Medical Image File Formats." Journal of Digital Imaging **27**(2): 200-206.

Le Bihan, D., E. Breton, D. Lallemand, P. Grenier, E. Cabanis and M. Laval-Jeantet (1986). "MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders." Radiology **161**(2): 401-407.

LeCun, Y., Y. Bengio and G. Hinton (2015). "Deep learning." Nature **521**(7553): 436-444.

Lee, S. Y., Y. R. Shin, H. J. Park, M. H. Rho and E. C. Chung (2017). "Usefulness of multiecho fast field echo MRI in the evaluation of ossification of the posterior longitudinal ligament and dural ossification of the cervical spine." PLoS One **12**(8): e0183744.

Li, M., Y.-Z. Zhu, Y.-C. Zhang, Y.-F. Yue, H.-P. Yu and B. Song (2020). "Radiomics of rectal cancer for predicting distant metastasis and overall survival." World journal of gastroenterology **26**(33): 5008-5021.

Liang, L., J. Zhu, H. Jia, L. Huang, D. Li, Q. Li and X. Li (2016). "Predictive value of pretreatment lymphocyte count in stage II colorectal cancer and in high-risk patients treated with adjuvant chemotherapy." Oncotarget **7**(1): 1014-1028.

Loh, W.-Y. (2015). A Brief History of Classification and Regression Trees, Department of Statistics University of Wisconsin–Madison.

Lv, L., B. Xin, Y. Hao, Z. Yang, J. Xu, L. Wang, X. Wang, S. Song and X. Guo (2022). "Radiomic analysis for predicting prognosis of colorectal cancer from preoperative 18F-FDG PET/CT." Journal of Translational Medicine **20**(1): 66.

Mariscal, G., Ó. Marbán and C. Fernández (2010). "A survey of data mining and knowledge discovery process models and methodologies." The Knowledge Engineering Review **25**(2): 137-166.

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochimica et Biophysica Acta **405**(2): 442-451.

Merboldt, K.-D., W. Hanicke and J. Frahm (1985). "Self-diffusion NMR imaging using stimulated echoes." Journal of Magnetic Resonance (1969) **64**(3): 479-486.

Mohammadi, N. (2021). "Radiomics using MR brain scans and RENT for identifying patients receiving ADHD treatment." Retrieved 26.05.2022.

Morvan, M. L., J. Josse, E. Scornet and G. Varoquaux (2021). "What's a good imputation to predict with missing

values?" NeurIPS 2021 - 35th Conference on Neural Information Processing Systems.

National Cancer Institute. (2021). "Colorectal Cancer Prevention (PDQ®)–Patient Version." Retrieved 04.01.2022, from https://www.cancer.gov/types/colorectal/patient/colorectal-prevention-pdq.

National Cancer Institute. (s.a.). "Overall Survival Rate." Retrieved 07.01.2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/overall-survival-rate.

Nogueira, S., K. Sechidis and G. Brown (2018). "On the stability of feature selection algorithms." Journal of Machine Learning Research **18**: 1-54.

Olofsson, C. K. (2021). "Using machine learning and Repeated Elastic Net Technique for identification of biomarkers of early Alzheimer's disease." Retrieved 26.05.2022.

Olofsson, C. K. (2021). Using machine learning and Repeated Elastic Net Technique for identification of biomarkers of early Alzheimer's disease, Norwegian University of Life Sciences, Ås.

Our World in Data. (2018). "Number of deaths by cause." Retrieved 04.01.2022, from https://ourworldindata.org/grapher/annual-number-of-deaths-by-cause?time=latest&country=~OWID_WRL.

Parekh, V. S. and M. A. Jacobs (2017). "Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI." npj Breast Cancer **3**(1): 43.

Parekh, V. S. and M. A. Jacobs (2019). "Deep learning and radiomics in precision medicine." Expert Review of Precision Medicine and Drug Development **4**(2): 59-72.

Parekh, V. S. and M. A. Jacobs (2020). "Multiparametric radiomics methods for breast cancer tissue characterization using radiological imaging." Breast Cancer Res Treat **180**(2): 407-421.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford and A. R. Feinstein (1996). "A simulation study of the number of events per variable in logistic regression analysis." J Clin Epidemiol **49**(12): 1373-1379.

Perez-Lebel, A., G. Varoquaux, M. Le Morvan, J. Josse and J. B. Poline (2022). "Benchmarking missing-values approaches for predictive models on health databases." Gigascience **11**.

Posse, S., C. A. Cuenod and D. Le Bihan (1993). "Human brain: proton diffusion MR spectroscopy." Radiology **188**: 719-725.

Preston, D. C. (2006). "Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics." Retrieved 11.06.2022, from https://case.edu/med/neurology/NR/MRI%20Basics.htm.

Preziosi, P., A. Orlacchio, G. Di Giambattista, P. Di Renzi, L. Bortolotti, A. Fabiano, E. Cruciani and P. Pasqualetti (2003). "Enhancement patterns of prostate cancer in dynamic MRI." European Radiology **13**(5): 925-930.

Python Software Foundation. (2022). "datetime — Basic date and time types." Retrieved 18.01.2022, from https://docs.python.org/3/library/datetime.html.

Raghunathan, T., J. Lepkowski, J. Hoewyk and P. Solenberger (2000). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." Survey Methodology **27**.

Raschka, S. and V. Mirjalili (2019). Python Machine Learning. Machine Learning and Deep Learning in Python, scikit-learn, and TensorFlow 2. Birmingham - Mumbai, Packt.

Raschka, S. and V. Mirjalili (2019). Python Machine Learning. Machine Learning and Deep Learning in Python, scikit-learn, and TensorFlow 2. Birmingham - Mumbai, Packt**:** 115-120.

Raschka, S. and V. Mirjalili (2019). Python Machine Learning. Machine Learning and Deep Learning in Python, scikit-learn, and TensorFlow 2. Birmingham - Mumbai, Packt**:** 211-215.

Redalen, K. R. (2018). "Functional MRI of Hypoxia-mediated Rectal Cancer Aggressiveness (OxyTarget)." Retrieved 21.04.2022, from https://clinicaltrials.gov/ct2/show/study/NCT01816607.

Rinck, P. A. (2021). Magnetic Resonance in Medicine: A Critical Introduction, BoD.

Rodgers, D. M., R. Jacobucci and K. J. Grimm (2021). "A Multiple Imputation Approach for Handling Missing Data in Classification and Regression Trees." Journal of Behavioral Data Science **1**: 127–153.

Sagi, O. and L. Rokach (2020). "Explainable decision forest: Transforming a decision forest into an interpretable tree." Information Fusion **61**: 124-138.

Sana, H. B. (2021). Sequential and orthogonalized partial least squares regression applied to healthcare data acquired from patients diagnosed with gastrointestinal carcinoma, Norwegian University of Life Sciences, Ås.

Segal, M. (2003). "Machine Learning Benchmarks and Random Forest Regression." Technical Report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco.

Shearer, C. (2000). "The CRISP-DM Model: The New Blueprint for Data Mining." JOURNAL OF DATA WAREHOUSING **5**(4): 13-22.

Sima, C. and E. R. Dougherty (2008). "The peaking phenomenon in the presence of feature-selection." Pattern Recognition Letters **29**: 1667–1674.

Smithsonian Institution (1946). Annual Report of the Board of Regents of the Smithsonian Institution. Washington.

Smola, A. and S. Vishwanathan (2008). "Introduction to machine learning." Cambridge University, UK **32**(34): 2008.

Sollini, M., L. Antunovic, A. Chiti and M. Kirienko (2019). "Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics." European Journal of Nuclear Medicine and Molecular Imaging **46**.

Stefanowski, J. (2008). Discovering Decision Trees, Institute of Computing Science Poznań University of Technology.

Stone, M. (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions." Journal of the Royal Statistical Society. Series B (Statistical Methodology) **36**: 111-147.

Särndal, C.-E., B. Swensson and J. Wretman (1992). Model assisted survey sampling, Springer Science & Business Media.

Søvdsnes, M. S. (2021). Predicting treatment outcome of colorectal cancer from MRI images using machine learning, Norwegian University of Life Sciences, Ås.

Teruel, J. R., M. G. Heldahl, P. E. Goa, M. Pickles, S. Lundgren, T. F. Bathen and P. Gibbs (2014). "Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer." NMR Biomed **27**(8): 887-896.

The pandas development team. (2022). "Pandas 1.4.0." Retrieved 18.01.2022.

Theodoridis, S. and K. Koutroumbas (2008). Pattern Recognition, 4th edition. Florida, US, Academic Press.

Thirunavukarasu, P., C. Talati, S. Munjal, K. Attwood, S. B. Edge and V. Francescutti (2015). "Effect of Incorporation of Pretreatment Serum Carcinoembryonic Antigen Levels Into AJCC Staging for Colon Cancer on 5-Year Survival." JAMA Surg **150**(8): 747-755.

Tubiana, M. (1996). "[Wilhelm Conrad Röntgen and the discovery of X-rays]." Bull Acad Natl Med **180**(1): 97-108.

TURING, A. M. (1950). "I.—COMPUTING MACHINERY AND INTELLIGENCE." Mind **LIX**(236): 433-460.

Udupa, J. K. and G. T. Herman (1999). 3D Imaging in Medicine, Second Edition, CRC Press.

van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification." Statistical Methods in Medical Research **16**(3): 219-242.

van Buuren , S. and C. G. M. Oudshoorn (2000). Multivariate Imputation by Chained Equations

van der Meulen, P., J. P. Groen, A. M. Tinus and G. Bruntink (1988). "Fast Field Echo imaging: an overview and contrast calculations." Magn Reson Imaging **6**(4): 355-368.

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

van Smeden, M., J. A. H. de Groot, K. G. M. Moons, G. S. Collins, D. G. Altman, M. J. C. Eijkemans and J. B. Reitsma (2016). "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis." BMC medical research methodology **16**(1): 163-163.

Vidal, T. and M. Schiffer (2020). Born-Again Tree Ensembles. Proceedings of the 37th International Conference on Machine Learning. D. Hal, III and S. Aarti. Proceedings of Machine Learning Research, PMLR. **119:** 9743--9753.

Vilagut, G. (2014). Test-Retest Reliability. Encyclopedia of Quality of Life and Well-Being Research. A. C. Michalos. Dordrecht, Springer Netherlands**:** 6622-6625.

Wang, J., S. Li, Y. Liu, C. Zhang, H. Li and B. Lai (2020). "Metastatic patterns and survival outcomes in patients with stage IV colon cancer: A population-based analysis." Cancer Med **9**(1): 361-373.

WHO. (2020). "All cancers fact sheet." Retrieved 04.01.2022, from https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf.

WHO. (2021). "Cancer - Key Facts." Retrieved 04.01.2022, from https://www.who.int/en/news-room/fact-sheets/detail/cancer.

Wibmer, A., H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, V. E. Reuter, J. Eastham, E. Sala and H. A. Vargas (2015). "Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores." Eur Radiol **25**(10): 2840-2850.

Wiemken, T. L. and R. R. Kelley (2020). "Machine Learning in Epidemiology and Health Outcomes Research." Annual Review of Public Health **41**(1): 21-36.

Xiao, B., J. Peng, J. Tang, Y. Deng, Y. Zhao, X. Wu, P. Ding, J. Lin and Z. Pan (2020). "Serum Gamma Glutamyl transferase is a predictor of recurrence after R0 hepatectomy for patients with colorectal cancer liver metastases." Therapeutic Advances in Medical Oncology **12**: 175883592094797.

Yoo, S., J. K. Kim and I. G. Jeong (2015). "Multiparametric magnetic resonance imaging for prostate cancer: A review and update for urologists." Korean J Urol **56**(7): 487-497.

Yoshino, T., R. Obermannová, G. Bodoky, R. Garcia-Carbonero, T. Ciuleanu, D. C. Portnoy, T. W. Kim, Y. Hsu, D. Ferry, F. Nasroulah and J. Tabernero (2017). "Baseline carcinoembryonic antigen as a predictive factor of ramucirumab efficacy in RAISE, a second-line metastatic colorectal carcinoma phase III trial." European Journal of Cancer **78**: 61-69.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society. Series B (Statistical Methodology) **67**(2): 301-320.

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

Postboks 5003
NO-1432 Ås
Norway