Optimizing patient care in rectal cancer: novel approach to predict treatment response

Raets, Camille

Publication date:
2025

VRIJE
UNIVERSITEIT
BRUSSEL

Thesis submitted in fulfilment of the requirements for the degree of Doctor in Medical Sciences

OPTIMIZING PATIENT CARE IN RECTAL CANCER

Novel Approach to Predict Treatment Response

Camille Raets

2024

Promotors: prof. dr. Kurt Barbé prof. dr. Mark De Ridder

Medicine and Pharmacy

Board of Examiners

Promotors
Prof. Dr. Kurt Barbé
Biostatistics and Medical Informatics (BISI) Research Group
Vrije Universiteit Brussel

Prof. Dr. Mark De Ridder
Department of Radiotherapy
Universitair Ziekenhuis Brussel

Jury Members
Prof. Dr. Eric Deutsch
Radiotherapy Department
Institut Gustave Roussy

Prof. Dr. Guy Nagels
Artificial Intelligence supported Modelling in clinical Sciences
Vrije Universiteit Brussel

Prof. Dr Francesco Lamonaca
Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica
Università Della Calabria

Prof. Dr. Gert Van Gompel
Medical Imaging
Universitair Ziekenhuis Brussel

Prof. Dr. Sonia Van Dooren
Molecular genetics labratory
Universitair Ziekenhuis Brussel
Chair
Dr. Sebastiaan Engelborghs
Neuroprotection & Neuromodulation
Vrije Universiteit Brussel

Acknowledgments

First and foremost, I would like to express my sincere gratitude for the opportunity I've been given. Medicine has always been a field I've been passionate about, and I am incredibly grateful for the chance to contribute to the Smart-Qi project.

I would also like to extend my deepest thanks to both of my promotors for their invaluable guidance throughout the years. Prof. Dr. Kurt Barbé, your mathematical expertise has been indispensable, and I am continually amazed by your ability to tackle even the most complex problems. Prof. Dr. Mark De Ridder, your medical guidance has been equally transformative. I now feel as though I've only just begun to scratch the surface of medical knowledge, thanks to your mentorship. I deeply value the time and patience you have shown in explaining everything to me, ensuring I understood even the most complex concepts.

A heartfelt thank you to the entire BISI team. From day one, I felt welcomed and supported by this diverse group, where different backgrounds and experiences came together to form a cohesive, warm, and collaborative team. I will fondly remember our after-work game nights and the Christmas parties we celebrated together. If I could, I would bring the entire team with me to my new job.

I am also deeply grateful to the Department of Radiology at UZB for their support, data provision, and access to their database. Although I may not have met everyone in the department, I am fully aware of the vital and often life-saving work that takes place there. Special thanks to Chaïmae for providing the Radiomics data, and to Kathleen Leemans and Ka Lun Law for setting up the initial data and handling all necessary paperwork. I am also grateful to Sven Van Laere for sharing so much information in the brief period after you joined the department, and for proactively helping to find additional patients for the study.

To the VUB and UZB, this is not goodbye, but rather a "see you later."

I would like to extend my thanks to the jury members for their time, effort, and insightful feedback, which helped enhance my thesis further.

A special thank you to my family for their unwavering support. To my mother, thank you for always trying to understand what I was working on, even when the details were far from easy to grasp. To my father, I am grateful not only for his support but for his encouragement. He always knew how to cheer me up by joking that no matter how difficult the mathematical problem, I could count on him to solve it. His humor and unwavering belief in me made even the toughest moments easier to handle. And to my sister, who, despite her dislike of mathematics, expressed her support in a way that only sisters can. Your encouragement and understanding meant the world to me.

Anyone who knows me even a little bit understands that the family I grew up in consists of more than just my parents and sister. Dogs have always been an important part of our lives, and I want to thank them for being the ones who, no matter the time, would always listen to me talk about mathematics and fully hear me out during all my mental breakdowns—something every PhD student experiences multiple times. Most especially, I want to honor the memory

of my late dog, Spikki, who was a loyal companion during late-night work sessions. While he often appeared bored and tired, I know he was always there for me, providing quiet, unwavering support.

Lastly, I want to express my deepest appreciation to the man who has become my own family, Kasper. Thank you for being my constant support throughout these years, for also standing by me during the breakdowns that come with a PhD, and for always encouraging me when I needed it most. I also want to express my gratitude for the time and effort you devoted to reading my thesis, meticulously checking for spelling and grammar mistakes.

Summary

Researching the optimization of patient care across all stages of the care continuum is crucial. Rectal cancer remains a very deadly disease, often causing major discomfort and decreasing the patient's quality of life (QOL) due to invasive surgery. Therefore, it is important to develop a prediction algorithm capable of predicting the tumor regression grade (TRG) before the start of any therapy. The TRG is typically determined by microscopic analysis of a biopsy obtained during surgery. The TRG used for rectal cancer patients at our hospital is the Dworak TRG. Since this system is a semi-quantitative grading system, there exists some subjectivity and variability in it. Consequently, we reclassified Dworak grades $0, 1$, and 2 as bad responders, while grades 3 and 4 were categorized as good responders, giving us a binary TRG classification problem.

We introduced a novel customized Random Forest (RF) algorithm to predict the binary TRG using radiomics extracted from the planning CT's. A total of 111 radiomic features were extracted using the open-source package PyRadiomics. Our proposed algorithm, called the Evolutionary Random Subspace Forest (ERSF), builds upon the algorithms developed by Ho and Breiman. We used subspaces for variable selection and pruned trees iteratively. Additionally, instead of utilizing traditional classification trees, we opted for linear discriminant analysis (LDA) trees. Our ERSF gave a proportional accuracy of 76.0% for the training and 65.4% for the validation.

Given the subjective nature of the Dworak, its impact on the prediction accuracy of ERSF cannot be overlooked. Moreover, we believe that understanding the relationship between the forest and expert opinions is extremely important. Thus, we revisited the surgical notes and identified patients whose grade (bad or good responder) was ambiguous, labeling them as "grey-zone patients". Our analysis revealed that the algorithm encountered greater difficulty in predicting outcomes for grey-zone patients, achieving only 63.5% proportional accuracy. In contrast, for non-grey zone patients, the proportional accuracy was notably higher at 92.5%. These findings underscore the influence of TRG subjectivity on the algorithm's misclassifications.

Additionally, we extracted Fourier features from the same CT images and fed them into the ERSF. The results were mediocre, giving a proportional accuracy of 82.4% for the training and 59.1% for the validation. Given the retrospective nature of our data, these results were somewhat expected. The data set contains variations in pixel spacing settings, which could potentially lead to prediction challenges. To address this issue, we experimented with several uniformizations of the pixel settings. We obtained encouraging results for some, with the most successful configuration giving a proportional accuracy of 90.6% for the training and 72.7% for the validation using Fourier data.

Finally, we extracted prior information from the Fourier data and constructed a new ERSF using both radiomics and the Fourier prior information in the LDA. The best results gave a proportional accuracy of 73.8% for the training and 72.7% for the validation, surpassing the initial radiomics results.

Samenvatting

Het onderzoeken van de optimalisatie van patiëntenzorg in alle fasen van het zorgcontinuüm is cruciaal. Rectum kanker blijft een zeer dodelijke ziekte, vaak met ernstige ongemakken en een vermindering van de kwaliteit van leven van de patiënt als gevolg van invasieve chirurgie. Daarom is het belangrijk om een predictie algorithme te ontwikkelen dat in staat is om de tumor regressiegraad (TRG) te voorspellen vóór de start van de therapie. De TRG wordt typisch bepaald door microscopische analyse van een biopsie verkregen tijdens de operatie. De TRG die wordt gebruikt voor patiënten met rectum kanker in ons ziekenhuis is de Dworak TRG. Aangezien dit systeem een semi-kwantitatief gradatiesysteem is, bestaat er enige subjectiviteit en variabiliteit in. Als gevolg hiervan hebben we Dworak graden $0, 1$ en 2 opnieuw geclassificeerd als slechte responders, terwijl graden 3 en 4 werden gecategoriseerd als goede responders, wat ons een binair TRG-classificatieprobleem opleverde.

We hebben een nieuw aangepast Random Forest algoritme geïntroduceerd om de binaire TRG te voorspellen met behulp van radiomics die zijn geëxtraheerd uit de planning-CT's. In totaal werden 111 radiomics geëxtraheerd met behulp van het open-source pakket PyRadiomics. Ons voorgestelde algoritme, genaamd de Evolutionary Random Subspace Forest (ERSF), bouwt voort op de algoritmen ontwikkeld door Ho en Breiman. We gebruikten subspaces voor variabele selectie en snoeiden bomen iteratief. Bovendien kozen we er in plaats van traditionele classificatiebomen voor om lineaire discriminantanalyse (LDA)-bomen te gebruiken. Onze ERSF gaf een proportionele nauwkeurigheid van 76.0% voor de training en 65.4% voor de validatie.

Gezien de subjectieve aard van de Dworak, kan de invloed ervan op de voorspellende nauwkeurigheid van ERSF niet over het hoofd worden gezien. Bovendien zijn we van mening dat het begrijpen van de relatie tussen het bos en expertmeningen uiterst belangrijk is. Daarom hebben we de chirurgische aantekeningen opnieuw bekeken en patiënten geïdentificeerd waarvan de graad (slechte of goede responder) ambigu was, ze labelend als "grijze-zone patiënten". Onze analyse onthulde dat het algoritme meer moeite had met het voorspellen van uitkomsten voor grijze-zone patiënten, waarbij slechts 63.5% proportionele nauwkeurigheid werd behaald. Daarentegen was de proportionele nauwkeurigheid voor niet-grijze zone patiënten aanzienlijk hoger met 92.5%. Deze bevindingen benadrukken de invloed van TRG-subjectiviteit op de misclassificaties van het algoritme.

Bovendien hebben we Fourier-features geëxtraheerd uit dezelfde CT-beelden en deze in de ERSF gestoken. De resultaten waren middelmatig, met een proportionele nauwkeurigheid van 82.4% voor de training en 59.1% voor de validatie. Gezien het retrospectieve karakter van onze gegevens waren deze resultaten enigszins te verwachten. De data set bevat variaties in de instellingen van de pixel groottes, wat mogelijk tot problemen in de voorspelling kan leiden. Om dit probleem aan te pakken, hebben we geëxperimenteerd met verschillende standardisaties van de pixel groottes. We behaalden bemoedigende resultaten voor sommige, waarbij de meest succesvolle configuratie een proportionele nauwkeurigheid van 90.6% voor de training en 72.7% voor de validatie met Fourier-features opleverde.

Tot slot hebben we voorafgaande informatie uit de Fourier-features geëxtraheerd en een nieuwe ERSF geconstrueerd met behulp van zowel radiomics als de Fourier-features informatie in de LDA. De beste resultaten gaven een proportionele nauwkeurigheid van 73.8% voor de training en 72.7% voor de validatie, waarbij de initiële radiomics resultaten werden overtroffen.

List of abbreviations

The list of abbreviations that have been used thoughout this thesis can be found here.

0-Z	
ACC	(Prediction) Accuracy
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
CNN	Convolutional Neural Network
CRC	Colorectal cancer
CRM	Circumferential Resection Margin
CRP	C-reactive protein
CRT	Chemoradiotherapy
CT	Computed tomography
cTNM	Clinical TNM
DFT	Discrete Fourier Transform
DL	Deep Learning
ECG	Electrocardiography
ERSF	Evolutionary Random Subspace Forest
FFT	Fast Fourier Transform
FIT	Fecal immunochemical test
FOB test	Fecal occult blood test
GLCM	Gray-Level Co-occurrence Matrix
GLDM	and Gray-Level Dependence Matrix features
GLRLM	Gray-Level Run Length Matrix
GLSZM	Gray-Level Size Zone Matrix
GTV	Gross Tumor Volume
IBD	Inflammatory bowel disease
IBSI	Image Biomarker Standardization Initiative
ICD	International Disease Classification
KNN	K-nearest neighbors
LAR	Low anterior resection
LARS	Low anterior resection syndrome
LDA	Linear Discriminant Analysis
LOOCV	Leave-One-Out Cross-Validation

ML	Machine Learning
MRI	Magnetic Resonance Image
NGDTM	Neighboring Gray-Tone Difference Matrix features
NIH	National Institute of Health
NIR spectroscopy	Near-infrared spectroscopy
pACC	Proportional (prediction) Accuracy
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PLT	Platelet Count Test
pTNM	pathological TNM
QOL	Quality of life
RF	Random Forest
ROI	Region-of-interest
RQS	Radiomics Quality Score
SEER	Surveillance, Epidemiology, and End Results Program
SVM	Support Vector Machines
TME	Total mesorectal excision
TNM	Tumor, Nodes, Metastasis staging
TRG	Tumor regression grade
UICC	Union for International Cancer Control
UZB	Universiteir Ziekenhuis Brussel (University Hospital Brussels)

Contents

Chapter 1

Introduction

Medicine is evolving rapidly, increasingly relying on mathematical tools and novel, innovative technologies. Cancer remains the leading cause of death worldwide, claiming nearly 10 million lives annually [1]. Among the various types of cancer, colorectal cancer (CRC) ranks as one of the top three in prevalence and mortality [2, 3]. With advancements in imaging techniques and computational methods, the potential to leverage artificial intelligence (AI) and machine learning (ML) for improved diagnostic accuracy and prognostic modeling has grown significantly. However, utilizing these techniques remains complex and challenging.

This thesis explores the intersection of computational and medical sciences in determining the tumor regression grade (TRG) of rectal cancer patients using planning Computed Tomography (CT) images. By extracting quantitative features from these images and applying ML methods, the research aims to contribute to a more precise and effective approach for rectal cancer prediction. The following sections outline the key concepts, motivation, research objectives, and structure of the thesis, providing a comprehensive roadmap for this interdisciplinary investigation.

1.1 Background and Context

We begin by providing the necessary background and context for this thesis. This section will briefly outline all the topics discussed in the subsequent chapters.

1.1.1 Cancer

Cancer, often called a malignant tumor, remains an omnipresent disease that claims millions of lives each year [1]. It can develop in any part of the body and is characterized by abnormal cells that grow aggressively. When cancer spreads to other parts of the body, it is referred to as metastasis, which is often the primary cause of cancer-related deaths.

CRC, which includes colon cancer, rectal cancer, and anal cancer, is the second most commonly diagnosed cancer and the third most lethal [3]. Patients with CRC often require aggressive therapies, such as invasive surgeries, which can significantly impact their quality-of-life (QOL).

Despite the widespread prevalence and high mortality of cancer, significant advancements have been made over the years across the entire cancer care continuum. These include improved screening methods, the adoption of more refined and personalized treatments, and advancements in precision medical equipment. In cancer care and detection, there is a growing reliance on machine learning algorithms, feature selection techniques, and other predictive modeling approaches

[5]. Some of these algorithms can predict cancer diagnoses even before symptoms manifest or become visible to the human eye.

1.1.2 CT Imaging and Feature extraction

CT images play a crucial role in various aspects of medicine. For rectal cancer specifically, they are used to stage patients after diagnosis and to plan radiotherapy.

CT images can also be utilized for predictive purposes, such as making predictions about diagnosis or forecasting the response to therapy. To use these predictive methods, quantitative features must be extracted from the images. One feature type that has gained significant popularity over the years is radiomics [14].

Radiomics

Radiomics is another method of extracting quantitative features from images. We chose to work with radiomics features for predicting the Dworak TRG of our rectal cancer patients. They are popular in the field of medical imaging. Radiomics can be used for the extraction of numerous quantitative features from a tumor segmentation in three-dimensional medical images. These features offer a more comprehensive characterization of the 3D segmentation extracted from the CT images, providing the researchers with potentially valuable information that may not be visible to the naked eye [14]. The features can be used for various aspects in cancer research, such as staging, tumor prognosis, prediction of treatment response, assessing survival, risk of recurrence and more [15].

Discrete Fourier Features

Discrete Fourier transformation (DFT) presents another method for feature extraction [16]. It is a valuable tool capable of converting signals into their frequency domain. In the context of medical images, DFT can be applied to translate voxels into their frequency domain, producing a complex-valued output. We will explore deep Fourier features extracted from three-dimensional images, as they can provide additional information that radiomic features may not capture.

1.1.3 Machine Learning

Nowadays, AI is a widely recognized term among the public. It consists creating all sorts of computer algorithms that perform tasks that typically require human input or at least human supervision [17][18]. Although often used interchangeably, AI and ML are distinct concepts. ML, a subset of AI, focuses on the development of statistical algorithms that learn from data to later apply learned patterns to new data [19]. In recent years, AI and ML have significantly advanced across various domains such as business, finance, education, and retail [17].

In medicine, AI and ML are extensively used for tasks including patient diagnosis, prognosis, and drug development. Early cancer diagnosis is among the beneficial applications of AI, with certain algorithms capable of diagnosing cancer even before observable symptoms emerge [5]. Early diagnosis improves treatment options and survival rates. Companies like PathAI utilize AI for precise diagnosis, allowing customized treatment plans tailored to individual patient needs [20]. However, diagnosis represents only the initial stage of the cancer care continuum where AI is applied. Moreover, AI's prominence in medicine is expected to increase in the coming years.

A frequently encountered problem in ML is the high dimensionality of data. Data is considered high-dimensional when it contains a large number of features. While more features can

provide additional useful information, they also increase the risk of overfitting and raise the computational cost of algorithms. To address high-dimensionality, feature selection techniques can be employed. These techniques identify and select features that are most relevant for prediction, helping to reduce the dimensionality while retaining the essential information.

There are various types of ML algorithms, including supervised and unsupervised learning methods. Supervised learning methods, such as logistic regression, Support Vector Machines (SVM), classification trees, and K-nearest neighbors (KNN), learn patterns from known data [21]. On the other hand, unsupervised learning methods like Principal Component Analysis (PCA) or K-means clustering are methods that learn patterns from data without knowing the output. In this thesis, we will only look at supervised learning methods.

we will briefly describe some of the supervised methods below:

Support Vector Machine

Support Vector Machine (SVM) is a classifier that constructs a linear function or hyperplane to separate data in such a way that the distance from the training data to the hyperplane is maximized [89].

Logistic Regression

Logistic regression is a statistical method that starts with a linear regression model and uses a transformation function to rescale the regression output to the $[0, 1]$ interval [91]. This interval can then be used with a threshold to make predictions.

Classification/Regression Trees

Classification/Regression trees are flowchart-like models that repeatedly split data into branches based on predictor variables [78].

Random Forest

A Random Forest (RF) combines multiple Classification/Regression trees into a single model. By combining multiple trees, it reduces the risk of overfitting.

Hierarchical Clustering

Hierarchical clustering groups data into subgroups consisting of patients with similar predictor variables. It focuses on uncovering the structure of the data rather than making predictions [89].

K-Nearest Neighbors

In K-Nearest Neighbors (KNN), a patient is classified by identifying the K closest samples in the training data [92]. The class is assigned based on the majority vote of the K neighbors' classes.

1.1.4 Application of Machine Learning in Medicine

Numerous studies have demonstrated the performance of ML algorithms, as shown in Table 1.1. Recently, E. Avuçle published an article using various machine learning techniques for the classification of 64 breast cancer patients [22]. The RF algorithm achieved a training accuracy of 96.63% and a test accuracy of 70.37%. While these results are excellent, our primary objective

is to achieve a balance between training and testing data performance, while simultaneously striving to maximize their accuracy.

In 2015, a study reported by Chen *et al.* also used the RF algorithm on patients with colorectal cancer [23]. A total of 186 near infrared (NIR) Fourier spectra from 20 randomly-selected patients was used. The samples were grouped as cancerous or normal, and the feasability to predict the cancer stage was studied using NIR spectra. The RF respectively showed an error percentage of 5.1% on the training set and 10% on the test set. These are good results showing that the cancerous stage could be predicted using the NIR spectrum. Various articles also show promising results using radiomic features extracted from medical images. In 2015, Ypsilantis *et al.* published an article regarding the response of patients with esophageal cancer to neoadjuvant chemotherapy [24]. In total, 107 patients were included in the study and the radiomics were extracted from PET images. The Mandard grading system was used to obtain two groups of responders, namely good and bad responders. They used multiple classification techniques to predict the response including RF, SVM and logistic regression. They respectively obtained an accuracy of $57.3\% \pm 7.8\%$, $55.9\% \pm 8.1\%$ and $51.4\% \pm 3.0\%$ for the training set. A few years later, He *et al.* published results of their study concerning the grading of rectal cancer using the radiomics extracted from the MRI images [25]. They had a total set of 118 patients and a four-level tumor grade based on the tumor differentiation. The tumor grade was predicted from the 50 most frequent radiomic features used by the RF model. The AUC scores were reported for tumor grades I to IV, respectively giving AUC scores of 91.8%, 82.2%, 77.5%, and 100% for the training set and AUC scores of 71.7%, 68.3%, 69.0%, and 82.7% for the test set.

Citation	**Classification Method**	**Training**	**Validation**
Avuçle [22]	RF	96.6%	70.4%
Breast cancer	SVM	62.9%	59.3%
Chen *et al.* [23]	RF	94.9%	90.0%
NIR spectrum CRC			
Ypsilantis [24]	RF	57.3%	-
PET scan,	SVM	55.9%	-
esophageal cancer	logit	51.4%	-
He *et al.* [25]	RF grade I	91.8%	71.7%
Radiomics MRI,	RF grade II	82.2%	68.3%
rectal cancer	RF grade III	77.5%	69.0%
	RF grade IV	100.0%	82.7%

Table 1.1: Examples of studies showing the performance of some ML algorithms in medicine.

1.1.5 Incorporating Machine Learning in medicine

In recent years, there has been a tremendous surge in the use of AI in medicine, driven by the belief that AI holds immense potential to significantly benefit the medical field. However, when using AI algorithms specifically designed for medical applications, caution must be exercised. As a matter of fact, not all clinicians trust and accept to use AI in real medical settings as the algorithms are often too complex and lack transparency (black-box models), with no possibility of being interpreted or questioned when desired [6, 7]. Further, if clinicians do not understand the algorithm and/or their choices, communicating and explaining those choices to the patient will be difficult, and we can hardly expect patients to trust those medical decisions. This issue is e.g. raised by Visar Berisha and Julie Liss, two co-founders of Aural Analytics, who have

even referred to AI in medicine as overhyped [8]. They emphasize that mistakes made by AI algorithms in medicine can lead to life-or-death situations and that it is essential to comprehend how AI renders decisions before using it in hospitals and other medical settings: this is called AI explainability.

Even though doubts about AI in medical settings have been raised, they can still provide significant value. It is essential to thoroughly test an AI model to determine whether it reaches the same conclusions as doctors and whether doctors can accept the AI's findings. For example, Bum-Sup Jang *et al.* developed a deep learning (DL) model using post-chemotherapy MRI images to predict the response (complete response and good response) in rectal cancer patients. They not only created a DL model but also assessed whether doctors agreed with the pathological responses generated by the AI model [9]. This type of research is crucial for establishing a trustworthy model that doctors can have confidence in.

1.1.6 Incorporating additional data in Machine Learning

When dealing with high-dimensional data, the addition of new data to the model can pose challenges. However, new data may provide valuable information crucial for prediction accuracy. Adding newly obtained features would only worsen the issue of high dimensionality, increasing the risk of overfitting. Hence, it is preferable to explore alternative approaches. We have chosen to employ a Bayesian method, wherein we initially model the data using our ML algorithm and subsequently incorporate the acquired information into subsequent ML algorithms using new data. This approach allows us to mitigate the challenges associated with high dimensionality as much as possible.

1.2 Thesis Structure

Firstly, Chapter 2 will provide a more detailed overview of CRC. This chapter will outline the risk factors, screening methods, diagnostic processes, and staging of CRC. Additionally, it will cover current treatment options, the QOL for CRC patients, and their survival outcomes.

In Chapter 3, we will briefly cover the basic definitions of Fourier analysis and its applications in medicine. Establishing the framework of Fourier analysis is crucial for the subsequent chapters, where CT scans will be defined and a novel approach for extracting Fourier features will be introduced.

Chapter 4 provides an overview of the fundamental framework of CT scans and explores radiomics in greater detail. It also introduces a specific open-source Python package called PyRadiomics [64], which is commonly used for radiomics extraction. This chapter will include applications of both methods, showcasing how features are extracted from CT images and highlighting the successes of radiomics in medicine.

The next chapters will describe the main body of the thesis. Covering all the different aspects of our study. In Chapter 5, we describe our research problem and introduce our own ML model. We aim to predict the TRG of rectal cancer patients using radiomics extracted from CT images taken before the start of therapy.

We first introduce some state-of-the-art classifiers, including non-ensemble methods like logistic regression and ensemble methods like Random Forests (RF). However, we later show that the performance of these state-of-the-art classifiers is poor. As a result, we introduce our own methodology, called the Evolutionary Random Subspace Forest (ERSF), which incorporates aspects of RF theory and introduces additional steps to enhance the model's performance. The results demonstrate that, despite the challenges posed by CT images and radiomic features, our ERSF model performs well in predicting the TRG.

Chapter 6 focuses on establishing a stronger connection between medicine and the ERSF. In this chapter, we explore whether the challenges that doctors face when determining the TRG align with the difficulties encountered by the algorithm. Additionally, we investigate whether incorporating the valuable insights of a doctor into the algorithm can enhance the prediction results.

In the next chapter, Chapter 7, we introduce a novel method for extracting quantitative features from CT images using Fourier analysis. Additionally, we closely examine various CT scan parameters, such as slice thickness and pixel spacing, to assess whether these factors influence the performance of the prediction.

Lastly, in Chapter 8, we explain a Bayesian approach, which can be used to incorporate new data into an existing ERSF model that was previously built using another data source. This approach is valuable because it allows for the combination of different data sources without increasing the high-dimensionality of the data. Additionally, it enables the integration of new data at a later time point without the need to reprocess the entire modeling with the previous data.

The previously mentioned Chapters 5 - 8 introduce new methodologies that are based on our published papers:

1. "An Evolutionary Random Forest to Measure the Dworak Tumor Regression Grade Applied to Colorectal Cancer" - Chapter 5 [27]

2. "Bridging the Gap Between Machine Learning and Medicine: A Critical Evaluation of the Dworak Regression Grade in Rectal Cancer" - Chapter 6 [28]

3. "Optimizing Rectal Cancer Patient Care: Dworak TRG Prediction via Bayesian Evolutionary Fourier-Domain Random Subspace Forest" - Chapters 7 and 8 [29]

Since these chapters are only based on the articles, there are some differences between the chapters and the published article. Most introductions have been altered to avoid repetition of specific details that were already discussed in previous chapters. Additionally, some sections from the articles have been revised or replaced to improve the overall readability of the thesis. Specifically, Chapter 5 was slightly adapted to fit better within this thesis. In the original article, we used Monte-Carlo Cross-Validation to validate our method. However, in subsequent research, we found that Leave-One-Out Cross-Validation (LOOCV) improved the stability of the method. Given that Chapter 6 requires LOOCV due to its patient-specific optimization approach, and since we use LOOCV throughout the rest of the thesis, we decided to update the cross-validation technique in Chapter 5 to maintain consistency across the chapters.

Lastly, we chose to split our final article into two separate chapters, Chapter 7 and Chapter 8, as it covers both the introduction of extraction of our Fourier features and the Bayesian approach. We believed that discussing these topics separately would enhance the clarity and readability of the text.

1.3 Research Problem and Motivation

In this thesis, we aim to develop a classification algorithm to predict the TRG for rectal cancer patients using radiomics extracted from CT images taken before the start of the therapy. We chose to extract radiomic features because they are widely used in medicine and encompass a broad range of characteristics. Additionally, as we were also interested in exploring whether features from the frequency domain could be useful, we introduce a new technique for extracting

Fourier features from these same CT images and explore methods for combining both data sets within our ML algorithm.

If we can accurately predict the TRG beforehand, this would provide valuable information to help tailor treatments when feasible. In cases of a favorable prediction, unnecessary surgeries could be avoided, preserving patients' QOL. Moreover, this approach could lead to the usage of an algorithm capable of recommending more personalized treatment plans for all patients.

It is important to note that our goal is not to replace clinical expertise with AI algorithms but to use them as a supplementary tool that provides a second opinion to clinicians. Additionally, we aim to open up "black-box" models to create a transparent connection between biomedicine and the decision-making process.

We seek to establish a link between the radiomics used in our machine learning (ML) algorithm and the clinical data employed by clinicians. This would allow for collaboration and mutual learning between the two fields. To achieve this, we will focus on the TRG, a measure heavily influenced by the pathologist's judgment. This introduces a challenge, as the ML algorithm must predict a subjective and variable output. Since ML algorithms cannot inherently express subjectivity, predicting highly variable TRGs is difficult. By exploring connections between patients who are challenging to classify by the algorithm and those difficult to grade by the pathologist, we can bridge the gap between medicine and ML, making the algorithm more interpretable and explainable.

Chapter 2

Colorectal cancer

2.1 Introduction

Colorectal cancer (CRC) ranks as the third most prevalent cancer globally, accounting for 9.6% of all cancer incidences in 2022 [2]. Additionally, it ranks as the second most fatal cancer, accounting for 904, 019 deaths in 2022 [3]. Following the International Classification of Disease (ICD) codes, CRC is categorized under codes C18–C21. These can be further divided into specific codes: colon cancer (C18), rectal cancer (C19–C20), and anal cancer (C21). Both colon and rectal cancers are common, whereas anal cancers are rarer [3, 30, 31]. The incidence and mortality rates for all ICD codes associated with colorectal cancer are provided in Table 2.1. Additionally, Table 2.2 highlights that, except for anal cancer, both the incidence and mortality rates are higher in males.

	Incidence		**Mortality**	
Cancer Type	Absolute Number	Rank	Absolute Number	Rank
Colorectal Cancer	1,326,425	3th	904,019	2nd
Colon Cancer	729,833	8th	343,817	10th
Rectal Cancer	729,833	8th	843,817	10th
Anal Cancer	54,306	30th	22,035	30th

Table 2.1: Incidence and mortality for colorectal cancer: anal, rectal, and colon cancer in 2022.

	Incidence		**Mortality**	
Cancer Type	Male	Female	Male	Female
Colorectal Cancer	55.5%	44.5%	55.3%	44.7%
Colon Cancer	53.3%	46.7%	52.7%	47.3%
Rectal Cancer	59.8%	40.%2	59.7%	40.3%
Anal Cancer	44.3%	55.7%	49.3%	50.7%

Table 2.2: Incidence and mortality percentages per gender for colorectal cancer: anal, rectal, and colon cancer in 2022.

This number is expected to continue to rise in the coming years [32]. This increase can partly be attributed to improved screening and increased life expectancy, resulting in more cancer diagnoses [33]. However, lifestyle factors also play a role in the rise of incidences, with physical inactivity, rising obesity rates, and increased consumption of processed foods and alcohol all directly correlated with higher cancer rates [1], [33].

2.2 Risk factors

Although studies do not always agree, certain factors have been associated with either an increased or decreased risk of developing CRC. One such factor is the level of development of a country, with CRC being more prevalent in highly developed countries [34, 35]. Additionally, immigrants tend to adopt the incidence risk of their new environment, suggesting that environmental factors play a significant role in CRC risk.

Factors increasing the risk of CRC

The personal medical history of a patient plays a significant role in their risk of developing CRC. A family history of CRC increases a patient's probability of developing the disease compared to those without such a history. Additionally, patients who have previously suffered from, or continue to suffer from, inflammatory bowel disease (IBD) face a heightened risk of developing CRC.

As noted earlier in Section 2.1, Table 2.2, being male slightly increases the risk of developing CRC compared to being female. It remains unclear whether this increased risk is attributable to genetic factors or differences in lifestyle between men and women. However, the overall difference in risk between the sexes is minimal. For both men and women, age is a significant factor, with CRC being more frequently diagnosed after the age of 50.

In addition to personal and family history, environmental and lifestyle factors also play a critical role. A diet high in red meat and fat-rich foods but low in fiber is associated with a higher risk of CRC. Additionally, patients with obesity or diabetes are at greater risk. Finally, unhealthy habits such as smoking and excessive alcohol consumption further increase the likelihood of developing CRC.

Factors decreasing the risk of CRC

Several factors can help decrease the risk of developing CRC. Maintaining a healthy diet low in red meat and fat but rich in fiber is crucial for reducing this risk. Additionally, quitting smoking and reducing alcohol consumption can further lower the likelihood of developing CRC. Engaging in regular (moderate) physical activity is also beneficial in decreasing the risk. Moreover, some studies suggest that proper intake of vitamins and micronutrients may play a role in reducing the risk of CRC [34].

2.3 Screening, diagnosis and staging of CRC

2.3.1 Screening and diagnosis

Screening and diagnosing a patient are two similar concepts with a key difference. Screening involves looking for a disease in the absence of symptoms. This can still be valuable, as some patients may be asymptomatic and thus show no apparent signs of illness. In contrast, diagnosing

focuses on identifying a specific disease that may be causing symptoms or abnormalities detected during screening. Both processes involve the use of tests to detect disease in patients.

Diagnosing CRC patients can be challenging because symptoms are not always consistent and some patients may even be asymptomatic [35]. Therefore, regular screening is crucial. Improved and more frequent screening allows for earlier diagnoses, when the cancer is more likely to be in an early stage, thereby enhancing patient survival. Over the years, numerous screening and diagnostic tools have been developed for CRC, some of which are described below [34, 35, 36, 37, 38].

Stool tests

Even for asymptomatic patients, stool tests are often effective diagnostic tools, as CRC tends to cause bleeding that can be detected in the patient's feces. One such test is the fecal occult blood (FOB) test. A specific type of FOB test is the guaiac fecal occult blood test (gFOBT), which uses a chemical reaction to detect CRC in stool samples. However, a disadvantage of this test is that it requires certain dietary restrictions prior to testing. For instance, consuming red meat can lead to a false positive result.

Another commonly used stool test for detecting CRC is the fecal immunochemical test (FIT), which uses antibodies to identify cancer. Unlike the FOB test, the FIT does not require any dietary restrictions, making it a more convenient option for patients.

Lastly, DNA testing can be used to detect CRC. Cancerous cells in the colon or rectum shed DNA into the stool as they are naturally released from the intestinal lining. This approach, known as multitarget stool DNA testing, identifies CRC by detecting these DNA markers.

Colonoscopy

If the FOB test is positive or if CRC is suspected, a colonoscopy is performed. This allows for full visualization of the entire colon and rectum by inserting a very thin tube with a small camera attached to the end into the colon through the anus. During the colonoscopy, a better visualization of the cancer can be made as well as a biopsy can be extracted. This biopsy is then sent for pathological examination to determine whether the tumor is malignant and to determine the stage of the cancer.

Sigmoidoscopy

A sigmoidoscopy is similar to a colonoscopy but uses a shorter tube that only scans the lower part of the colon and rectum. The shorter tube makes the procedure less invasive than a colonoscopy. While colonoscopies are typically performed under sedation, this is not necessary for a sigmoidoscopy. Overall, sigmoidoscopy is a simpler and quicker procedure compared to a colonoscopy. However, its usefulness is limited to cases where the suspected tumor is located in the lower colon or rectum.

Imaging techniques

In recent years, many new and improved imaging techniques have been developed. These techniques are invaluable for staging cancers as they can be used to look for metastasis and often cause less discomfort for patients compared to other screening and diagnostic tools.

One such imaging technique is computed tomography colonography. This involves taking two-dimensional or three-dimensional computed tomography (CT) scans, which provide physicians

with a detailed visualization of the tumor. CT colonography is the primary imaging tool used to obtain a comprehensive view of the entire colon and rectum.

Magnetic resonance imaging (MRI) is another useful tool, particularly for detecting metastases in nearby regions or organs. It is commonly used in the assessment of rectal cancer. Additionally, positron emission tomography (PET), often used in combination with CT scans, can help detect metastases more effectively.

2.3.2 Cancer staging

There are several staging techniques for CRC, including the Dukes classification system and the Astler-Coller system [35, 34]. However, the most commonly used staging method is from the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC). This method utilizes the TNM classification system, which is documented and regularly updated by the UICC and AJCC [39, 36].

In the TNM classification system (see Table 2.3): T describes the invasion of the primary tumor; N represents the extent of regional lymph node metastasis, if present; M indicates the presence of distant metastasis.

Primary Tumor (T)	
TX:	Primary tumor cannot be assessed
T0:	No evidence of primary tumor
Tis:	Carcinoma in situ
T1-4:	Increasing size and/or local extent of the primary tumor
Regional Lymph Nodes (N)	
NX:	Regional lymph nodes cannot be assessed
N0:	No invasion of regional lymph nodes
N1-2:	Increasing involvement of regional lymph nodes
Distant Metastasis (M)	
M0:	No distant metastasis
M1:	Distant metastasis

Table 2.3: Clinical TNM staging as defined by the UICC [39].

A tumor is staged by obtaining a biopsy and examining it pathologically. The TNM classification determined from the pathological examination serves as the basis for the AJCC/UICC cancer staging system. The TNM classification is sometimes denoted as cTNM (clinical TNM classification) to indicate that the classification is conducted before treatment surgery. When the classification is performed after surgery, it is referred to as the pTNM (pathological TNM classification).

A summarized version of the TNM classification specific to CRC is provided in Table 2.4. The full, detailed TNM classification for CRC can be found in the appendix (see Table 2.7). It is important to note that regional lymph nodes are determined by the exact location of the cancer within the colorectum.

The staging of CRC using the TNM classification is summarized in Table 2.5. The fully detailed staging is available in the appendix (see Table 2.8).

It is important to note that with better screening techniques, a shift of more patients from early to later stages, known as stage migration, can be observed. Older techniques were less effective in providing a clear picture of the disease, which could lead to underestimating its

Primary Tumor (T)	
TX:	Primary tumor cannot be assessed
T0:	No evidence of primary tumor
Tis:	Carcinoma in situ - invasion of lamina propria
T1:	Tumor invades submucosa
T2:	Tumor invades muscularis propria
T3:	Tumor invades subserosa or into non-peritonealized pericolic or perirectal tissues
T4:	Tumor directly invades other organs or structures and/or perforates viscera peritoneum
Regional Lymph Nodes (N)	
NX:	Regional lymph nodes cannot be assessed
N0:	No invasion of regional lymph nodes
N1:	Metastasis in 1 to 3 regional lymph nodes
N2:	Metastasis in 4 or more regional lymph nodes
Distant Metastasis (M)	
M0:	No distant metastasis
M1:	Distant metastasis

Table 2.4: TNM classification for CRC as defined by the UICC [39].

Stage	**T**	**N**	**M**
Stage 0	Tis	N0	M0
Stage I	T1, T2	N0	M0
Stage II	T3, T4	N0	M0
Stage III	Any T	N1, N2	M0
Stage IV	Any T	Any N	M1

Table 2.5: CRC staging as defined by the UICC [39].

severity. In contrast, newer and more advanced techniques are better at accurately detecting the disease. As a result, historical comparisons between studies may not always be valid.

2.4 Treatment

When determining the most suitable treatment for a patient, it is crucial to consider various factors, including the CRC stage, the specific type of CRC, personal and family medical history, and other relevant details. After diagnosis, a team of medical specialists evaluates these factors to decide on the best possible treatment plan for the patient.

During treatment, it is essential to focus not only on controlling the primary tumor but also on monitoring and managing existing and potential future metastases. Control of the primary tumor is primarily achieved through radiotherapy and surgery. For metastases, on the other hand, treatments such as chemotherapy, immunotherapy, and other systemic therapies can complement the management of the primary tumor. The choice of treatment depends on a multidisciplinary approach, ensuring optimal outcomes tailored to the patient's unique needs. In this section, we will briefly discuss some of the treatment options available for managing CRC [35, 36].

Surgery

Surgery is commonly used to treat the primary tumor in colorectal cancer (CRC). If the cancer is in an early stage, minimally invasive techniques, such as removing the tumor during a colonoscopy, may be possible. However, for more advanced cancers, more aggressive surgical techniques are required, such as total mesorectal excision (TME), which is frequently used for advanced rectal cancers. During a TME procedure, the rectum and surrounding mesorectal tissue, which includes fat, lymph nodes, blood vessels, and potentially cancerous cells, are removed. This procedure has proven effective in controlling local recurrences.

Although surgery is one of the standard treatments for CRC, it is a highly invasive procedure that often leaves patients with significant discomfort, such as the need for a stoma. If it is expected that surgery is not immediately necessary, physicians may opt for a watch-and-wait approach, where the cancer is closely monitored and surgery is performed only when deemed necessary.

Radiotherapy

Radiotherapy is often categorized as either neoadjuvant or adjuvant. Neoadjuvant radiotherapy refers to treatment administered before surgery, while adjuvant radiotherapy is given after surgery. In both cases, radiotherapy uses high-energy radiation to kill or damage cancerous cells.

Neoadjuvant radiotherapy, which is used to shrink the cancer before surgery, has been shown to reduce the risk of local recurrence in rectal cancer. For this reason, it is frequently chosen as a treatment strategy [36, 40, 41].

Chemotherapy

Chemotherapy is a treatment that uses drugs to destroy fast-growing cells, such as cancer cells. Like radiotherapy, chemotherapy can be administered either preoperatively or postoperatively. When given in combination with radiotherapy, it is often referred to as chemoradiotherapy. If deemed in the patient's best interest, chemotherapy may be given both before and after surgery.

For rectal cancer, a study by R. Sauer *et al.* showed that local recurrence rates were 7% lower in patients receiving preoperative chemotherapy compared to those receiving postoperative chemotherapy, over a 30-month follow-up period [42].

Immunotherapy

Immunotherapy is a relatively new technique that aims to harness the patient's own immune system to fight cancer. It offers several advantages, such as its ability to target cancerous cells specifically while preserving healthy cells. Additionally, immunotherapy can "remember" the cancer, providing long-term protection by enabling the immune system to recognize and respond to future occurrences of the disease.

2.4.1 Tumor Regression Grade

As explained in the previous section, rectal cancer is typically treated initially with several fractions of neoadjuvant (chemo)radiotherapy (CRT), followed by surgery, unless a watchful waiting approach is preferred. During the surgery, a biopsy is taken and sent to the pathologists for microscopical examination. The pathologist can then determine the tumor response grade (TRG) for the patient. Various grading systems for CRC exist and are used since there is no standard grading system being used.

2.4.2 Dworak Regression Grade

One of such grading systems for rectal cancer is the Dworak regression grade, a semi-quantitative grading system consisting of five different grades [43]. The Dworak TRG was proposed by O. Dworak *et al.* in 1997 to determine the regression grade of rectal cancer after preoperative radiotherapy. All patients in the study had locally-advanced rectal cancer and all received (chemo)radiotherapy before surgery. Rectal resection specimens were sent for pathological examination under a microscope. The specimens were semi-quantitatively graded by looking at the tumor mass, fibrotic changes, irradiation vasculopathy and peri-tumorous inflammatory reaction resulting in a new grading system (Table 2.6), nowadays known as the Dworak regression grade or Dworak TRG.

Dworak Regression Grade:	
TRG 0	No regression;
TRG 1	Dominant tumor mass with obvious fibrosis and/or vasculopathy;
TRG 2	Dominantly fibrotic changes with few tumor cells or groups that are easy to find;
TRG 3	Very few and microscopically difficult to find tumor cells in fibrotic tissue with or without mucous substance;
TRG 4	No tumor cells, only fibrotic mass resulting in a total response.

Table 2.6: Dworak grade as proposed in [43].

2.5 Quality-of-life

However, better screening and treatment don't necessarily equal a better quality of life (QOL). Due to the intensive treatments, patients might experience a decrease in the QOL, with reports indicating that QOL worsened as CRC progressed into a more advanced stage [44]. The invasive surgeries often performed for CRC treatment can leave the patients with numerous inconveniences, such as a (permanent) stoma or other anorectal disorders like leakage around the stoma, significantly affecting the self-esteem and confidence of the patient. Other side effects may include persistent fatigue, fear of cancer recurrence, depression, and more. Furthermore, surgeries can lead to low anterior resection syndrome (LARS), which is a collection of symptoms that patients can experience after undergoing a low anterior resection (LAR) with total mesorectal excision (TME). The symptoms can include incontinence or leakage, frequency or urgency, incomplete bowel movement, tenesmus, dyspareunia, and so on. It is evident that all these symptoms contribute to a decreased QOL [45][46][47].

2.6 Survival

As advancements continue, survival rates have significantly improved, transforming what was once an extremely deadly illness into a more manageable one. According to the Surveillance, Epidemiology, and End Results Program (SEER) by the National Institute of Health (NIH) in the U.S., data collected between 2013 and 2019 reveals a 5-year survival rate of 89.8% for localized rectal cancer, 73.7% for regional cancer, 17.8% for distant cancer, and 60.7% for unstaged rectal cancer [48].

2.7 Appendix

2.7.1 Full TNM classification for CRC

Primary Tumor (T)	
TX:	Primary tumor cannot be assessed
T0:	No evidence of primary tumor
Tis:	Carcinoma in situ - invasion of lamina propria
T1:	Tumor invades submucosa
T2:	Tumor invades muscularis propria
T3:	Tumor invades subserosa or into non-peritonealized pericolic or perirectal tissues
T4:	Tumor directly invades other organs or structures and/or perforates viscera peritoneum
T4a:	Tumor perforates visceral peritoneum
T4b:	Tumor directly invades other organs or structures
Regional Lymph Nodes (N)	
NX:	Regional lymph nodes cannot be assessed
N0:	No invasion of regional lymph nodes
N1:	Metastasis in 1 to 3 regional lymph nodes
N1a:	Metastasis in 1 regional lymph node
N1b:	Metastasis in 2 to 3 regional lymph nodes
N1c:	Tumor deposits without invasion into regional lymph nodes
N2:	Metastasis in 4 or more regional lymph nodes
N2a:	Metastasis in 4–6 regional lymph nodes
N2b:	Metastasis in 7 or more regional lymph nodes
Distant Metastasis (M)	
M0:	No distant metastasis
M1:	Distant metastasis
M1a:	Metastasis confined to one organ without peritoneal metastases
M1b:	Metastasis in more than one organ
M1c:	Metastasis to the peritoneum with or without other organ involvement

Table 2.7: TNM classification for CRC as defined by the UICC [39].

2.7.2 Full CRC cancer staging diagram

Stage	T	N	M
Stage 0	Tis	N0	M0
Stage I	T1, T2	N0	M0
Stage II	T3, T4	N0	M0
Stage IIa	T3	N0	M0
Stage IIb	T4a	N0	M0
Stage IIc	T4b	N0	M0
Stage III	Any T	N1, N2	M0
Stage IIIa	T1, T2	N1	M0
	T1	N2a	M0
Stage IIIb	T1, T2	N2b	M0
	T2, T3	N2a	M0
	T3, T4a	N1	M0
Stage IIIc	T3, T4a	N2b	M0
	T4a	N2a	M0
	T4b	N1, N2	M0
Stage IV	Any T	Any N	M1
Stage IVa	Any T	Any N	M1a
Stage IVb	Any T	Any N	M1b
Stage IVc	Any T	Any N	M1c

Table 2.8: CRC staging as defined by the UICC [39].

Chapter 3

Fourier analysis

When considering a signal, let it be a wave or an image, we usually say we are looking at its time or spatial domain. When signals are measured over time, we can speak about the time domain.Audio signals, stock prices, and electrical signals are all signals that are measured in the time domain. It we, on the other hand look at signals or images measured using data points that are distributed in space we can speak about the spatial domain. Medical scanning techniques like MRI and CT create images on the spatial domain as they are represented in pixel values. For the remainder of this chapter, we will always refer to the time or spatial domain as just the time domain. In Fourier analysis one often speaks about the time domain rather than the spatial domain as Fourier was originally developed with the goal of looking at the frequencies of time domain data.

Fourier analysis is a foundational tool in signal processing that enables the decomposition of signals into their frequency components. This transformation allows for the analysis of patterns and characteristics that are not immediately apparent in the time domain. While Fourier methods were originally developed for analyzing time-series data, they have since been widely adopted across diverse fields, including image processing, telecommunications, audio engineering, and finance.

A signal or image is typically composed of multiple frequencies combined together. For example, a heartbeat can be visually represented as waves using electrocardiography (ECG) [49]. In this context, high frequencies correspond to a fast-beating heart, while low frequencies represent a slower heartbeat. The Fourier Transform is a mathematical method that analyzes a complex signal or image—measured over time or space—and maps it to its frequency domain. This process decomposes the signal into its individual frequency components, capturing both the amplitude and phase of each.

Returning to the ECG example, instead of focusing on how the signal changes over time, it may be more beneficial to analyze the frequencies, i.e. rhythms and cycles within the signal. For instance, this approach can help identify whether the heart is beating at a steady, normal pace or if irregularities, such as cardiac arrhythmias, are present. Fourier Transforms not only identify the frequencies present in a signal but are also versatile tools used for tasks like denoising, data compression, and beyond.

Fourier Transforms are used in various fields where they are used for many different reasons. They can be used for signal analysis in for example audio engineering or telecommunication. They are also useful to compress images or music, think about the JPEG and MP3 extensions. When changing an image with a PNG extension to a JPEG extension, this image will be compressed and can therefore be stored more efficiently. Fourier Transforms can also be used for noise

reduction in audio or images. They can enhance, for instance, audio fragments such that the conversation is more clear. Fourier Transforms are also applicable in finance, where they can be used to forecast future trends of for instance a stock. In medical imaging, Fourier Transforms are crucial for reconstructing images from raw scanner data, such as those obtained from CT or MRI machines. This capability makes them indispensable for creating clear and accurate diagnostic images.

In this chapter, we will look at some key concepts of Fourier analysis [16, 50, 51, 52]. We will explore the Continuous Fourier Transform, which transforms a continuous function into its frequency domain. This type of Fourier transform is essential in understanding how CT images work. Additionally, we will look at the Discrete Fourier Transform (DFT), as in a computer, the calculation is often performed on a discrete time domain rather than a continuous one.

3.1 One-dimensional the Fourier Transform

Continuous Fourier Transforms are often referred to as just Fourier Transforms. A Fourier transform can be used to transform a continuous signal on a time domain to its frequency domain. Unlike Fourier series we don't necessarily need periodic signals for a Fourier transform.

Definition 1. *The **Fourier Transform** of a function $f(t)$ is given by*

$$\hat{f}(\xi) = \int_{\infty}^{\infty} f(t)e^{-2\pi i\xi t}dt, \tag{3.1}$$

where ξ represents the frequency.

In reality however, most signals and images are represented in a discrete manner when uploading to a computer or program. The standard sample rate for recording music is 44.1 kHz or $44,100$ samples per second. For images we see that those are stored as pixels, where each pixel has a color value assigned to it. It is easy to see that those are not examples of continuous signals. We therefore need a discrete version of the Fourier Transform, which is called the Discrete Fourier Transform or DFT.

The DFT transforms a sequence of values that together for a discrete signal in the time domain into another sequence of values in the frequency domain. We will consider finite signals as in a computer the signal should also be finite.

Definition 2. *Consider a sequence $x[n]$ for $n = 0, \ldots, N-1$ with N samples. The **Discrete Fourier Transform** is given by*

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-2\pi ikn/N}. \tag{3.2}$$

Alternatively, the DFT is also denoted by $\mathcal{F}\{x[n]\}$.

3.2 Fourier Transform for N-dimensions

In the previous section, we gave the Continuous and Discrete Fourier Transforms for the one-dimensional case. This is used for signal processing in telecommunication, audio signals, financial stock prices, etc. However, Fourier transforms can also be used in more dimensional settings like for enhancing two-dimensional images or even three-dimensional images like MRI or CT images.

We can generalize the definitions for both the Continuous and Discrete Fourier Transforms to more dimensions. For the Continuous Fourier Transform, we get the following definition.

Definition 3. *Consider a continuous function d-dimensional function $f(\underline{x})$ with vector $\underline{x} = (x_1, \ldots, x_d)$. The **Fourier Transform in d dimensions** is then given by*

$$\hat{f}(\underline{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\underline{x}) e^{-i2\pi \underline{\xi} \underline{x}^T} d\underline{x}, \tag{3.3}$$

where $\underline{\xi} = (\xi_1, \ldots, \xi_d)$ denotes the frequency vector. Alternatively, the d-dimensional Fourier transform is denoted by $\mathcal{F}\{f(\underline{x})\}$.

Alternatively, we can define the d-dimensional Discrete Fourier Transform.

Definition 4. *Consider a d-dimensional sequence $x[n_1, \ldots n_d]$ with $n_j = 0, \ldots, N_{j-1}$ for all $j = 1, \ldots, d$. The **Discrete Fourier Transform in d dimensions** is then given by*

$$X[\underline{k}] = \sum_{n_1=0}^{N_1-1} \cdots \sum_{n_d=0}^{N_d-1} x[\underline{n}] \exp\left\{-2\pi i \left(\frac{k_1 n_1}{N_1} + \ldots + \frac{k_d n_d}{N_d}\right)\right\} \tag{3.4}$$

where $\underline{k} = (k_1, \ldots, k_d)$ denotes the frequency indices and N_{j-1} for $j = 1, \ldots d$ denotes the length of the signal for dimension j. Alternatively, the d-dimensional Discrete Fourier transform is denoted by $\mathcal{F}\{x[n_1, \ldots n_d]\}$.

3.3 Fourier applications in medicine

Fourier analysis is used in various aspects of medicine, with few studies focusing on the extraction of Fourier features. However, there is no universal definition of Fourier features, meaning that the only resemblance across studies is the use of Fourier and the extraction of features. The specific way in which Fourier is applied and the features extracted can vary between studies. In this section, we will summarize some of the applications of Fourier analysis and Fourier features in medicine. We will first discuss more general applications before delving into specific examples of Fourier's use in medical research.

Fourier Transforms play a fundamental role in several imaging techniques, such as CT and MRI scans [53]. In CT, for example, Fourier is used to reconstruct two-dimensional images from the raw data collected by the scanner's detectors. This process allows the generation of images that are interpretable by the human eye. Similarly, Fourier Transforms are also used to reconstruct images in MRI scans.

Another application of Fourier Transforms is in electrocardiography or ECG, which is used to monitor a patient's heart activity. However, the ECG signal often contains noise introduced by the machine or external sources. Fourier Transforms can be applied to filter out this noise, leaving a clean signal that accurately represents the patient's heartbeat [49].

We will now delve into some more specific applications of Fourier analysis in medicine. An overview of different studies is provided in Table 3.1. In 2024, an article by A. Nokhostin and S. Rashidi detailed the use of Fourier analysis for diagnosing COVID-19 from lung CT images. They applied the Fractional Fourier Transform (FrFT) to the CT images, extracting features such as mean and median values [54]. The Fractional Fourier Transform is a generalization of the standard Fourier Transform. While the standard Fourier Transform maps a signal from the time domain to the frequency domain, the FrFT transforms the signal to intermediate domains that are mixtures of both time and frequency. For further details on the FrFT, refer to the work of A. Nokhostin and S. Rashidi. Using the extracted features, the researchers employed two different statistical modeling methods. Both approaches yielded high accuracy rates, achieving 99.90% and 99.84%, respectively.

Another application of Fourier Transforms is in the segmentation of colon cancer glands in histopathological images, as demonstrated by Y. B. Luo *et al.* [55]. They introduced a Fourier-based segmentation network, called FFS-Net, which utilizes Fourier transforms to enhance the segmentation of histopathological images. In their study, they tested the method on three different data sets and reported F1 scores—a metric used to evaluate classification model performance—of 0.930, 0.845, and 0.853, respectively. These results demonstrate the method's effectiveness in accurately segmenting glandular structures in colon cancer tissue using Fourier based features.

One final example of Fourier application in medicine is the study by T. Yoshimasu *et al.* [56]. In their research, they employed the Fast Fourier Transform (FFT), an efficient algorithm for computing the Fourier Transform, to analyze the complexity of tumor outlines on CT scans and differentiate between primary lung tumors and metastatic lung tumors. Their results demonstrated an accuracy of 89.7% in correctly classifying the tumors as either primary lung tumors or metastatic lung tumors.

Citation	**Application**	**Description**
A. Nokhostin & S. Rashidi [54]	Covid-19	Fractional Fourier Transform (FrFT) features for diagnosing Covid-19 on lung CT images
Y. B. Lou *et al.* [55]	Colon cancer	Fourier-based segmentation network for colon cancer glands in histopathological images
T. Yoshimasu *et al.* [56]	Lung cancer	Fast Fourier Transform to differentiate between primary lung tumors and metastatic lung tumors.

Table 3.1: Applications of Fourier in medicine.

Chapter 4

Computed Tomography (CT) and Feature Extraction

CT scanning is a well-known technological tool that has become immensely important in medical imaging offering us the capability to look inside a body and gather valuable information for clinicians. Over the years CT scanning has undergone extraordinary improvements, from the invention of X-ray to high-speed modern CT scanners as we know them today. They are inter alia used for detecting bone fractures, checking for head injuries, planning cancer treatments. This chapter delves into the fundamentals of CT imaging.

4.1 Introduction to CT

Although CT scans are frequently used in medical settings the introduction of the first CT scanner was not that long ago. As CT scans are a more advanced imaging technique that uses X-rays, we have to go back to the invention of the X-ray. In 1895, Wilhelm Conrad Röntgen, a German mechanical engineer and physicist, was doing some experiments with a cathode-ray tube when he, by hazard, discovered X-rays [53, 57]. A cathode-ray tube is a vacuum tube where an image is created by emitting electron beams from one or more electron guns on a phosphorescent screen [58].

A theoretical explanation of X-rays as high-energy electromagnetic radiation was provided in 1912 by Max von Laue, and Johann Radon developed a mathematical framework for image reconstruction in 1917 [53, 59]. However, it was not until 1967 that Sir Godfrey Hounsfield invented the first practical CT scanner [60]. His design used a rotating X-ray tube and image processing to create cross-sectional images of the human body. The first CT scanner, developed in 1971, took 20 minutes to process and revealed a brain lesion in 1972 [61]. Hounsfield and Allan Cormack received the Nobel Prize in 1979 for their contributions.

Hounsfield's first CT scanner was specifically created to scan the head. It was Robert Ledley however, who in 1974 firstly created a CT scanner that could scan the entire human body, and by the 1980s, CT scanners were in widespread use. Modern scanners use multidetector technology, allowing high-speed scans to be completed in minutes.

4.1.1 The basics of CT scanning technology

The process of making a CT image is very complex and consists of many complicated mathematical formulas. A CT image is created using one or more rotating X-ray tubes, forming a cross-sectional image.

X-rays are electromagnetic radiation that transmit energy in the form of short-waves and photons, which are particles transmitting light. The wavelength or frequency of electromagnetic radiation is used to classify it into different groups. X-rays, for instance, have wavelengths that are closer to those of ultraviolet (UV) light and even visible light, whereas the radiation used in MRI lies at a much longer wavelength, far removed from the X-ray spectrum.

When X-rays pass through different types of materials, they lose a different amount of energy. It is because of this difference in energy loss that we are able to capture differences in light on the image. For example, soft tissue will give a low intensity and will appear much darker than high intensity bone structures. This is because bone structures absorb more of the X-ray radiation and therefore will appear very bright, white even, on the image.

The absorption of X-rays varies over different materials. There will be a reduction of radiation intensity when there is an interaction between X-rays and the (medical) matter. This reduction of intensity is the reduction of the number of (positively charged) photons that are going through the matter and arrive at the detector. This process is called the (linear) attenuation. The (linear) attenuation coefficient of a certain matter is in medicine usually denoted by μ and measured in cm^{-1}. For an object with thickness x, the (linear) attenuation coefficient $\mu(x)$ is calculated as:

$$\mu(x) = -\frac{1}{x} \ln\left(\frac{I(x)}{I_0}\right), \tag{4.1}$$

with radiation intensity $I(x)$ of passing through the object x. As the attenuation coefficient denotes the reduction of the number of X-ray photons, it is clear that the attenuation coefficient will depend on the energy of the X-ray photons, where higher energy photons will penetrate the material more easily than lower energy photons.

X-rays give two-dimensional images that are often useful like when finding fractures in the human body after a fall. However, sometimes we could benefit from a more detailed three-dimensional image rather than only a two-dimensional one. CT images use X-rays but consist of multiple projections from multiple angles to create a cross-sectional view and 3D reconstruction. There are many different techniques for acquiring CT images, with helical and axial CT scanning two of the most well known techniques. In axial CT scanning the patient remains stationary while the X-ray tube rotates around him/her, giving one slice after which the patient is moved and a new slice is created. Helical is a faster scanning technique where both the table on which the patient lays and the X-ray tube move at the same time creating a spiral like pattern that creates the scan in one continuous motion. We will not delve into the mathematical framework of reconstructing CT images, as this involves complex computations. However, it is worth noting that reconstructing two-dimensional cross-sectional images from multiple projections relies heavily on Fourier Transformations (see Definition 3).

4.1.2 CT image representation: Hounsfield units

In the computer, CT images are represented as a three-dimensional grid consisting of pixels or cuboids, known as voxels. Each voxel represents a gray value that corresponds to the density of the tissue in that specific area of the body. The units in which the CT gray values are represented are the Hounsfield Units (HU). These units are a standardized scale and are calculated using the

linear attentuation coefficient of water and the linear attenuation coefficient of x:

$$HU_X := \frac{\mu_x - \mu_{\text{water}}}{\mu_{\text{water}}} \tag{4.2}$$

with μ_x the attenuation coefficient of a substance x and μ_{water} the attenuation coefficient of water [62].

4.1.3 Challenges in CT imaging

CT images provide valuable information across many medical fields. However, they are not without challenges [63]. The quality of a CT scan significantly impacts its utility. For instance, if the slice thickness is set too wide, the resulting scan may appear blurry and lack detail. Fortunately, advancements in CT technology have continually improved image quality over time.

Another challenge involves artifacts, which can distort the image. These artifacts come in various forms. For example, patient movement can cause motion artifacts. If a patient moves or shakes excessively during the scan, the resulting image quality will be compromised. A specific instance is chest CT scans, where the patient must hold their breath for several seconds. Failure to do so may lead to a blurry image.

Metal artifacts are another common issue. If a patient has a metal plate near the area being scanned, it can create an overly bright spot on the image. This brightness may obscure surrounding tissue, making it appear more prominent than it should.

4.2 Radiomics Features

Processing CT image data is not easy as it has a three-dimensional structure. It would be much more beneficial to work with quantitative features extracted from the CT image that capture the complexity of the image. Radiomics is such a method used in medicine for extracting a large number of quantitative features from medical images such as CT images using mathematical formulas. They can be used to extract quantitative information from complex scans that may not be visible with the naked eye [14]. Radiomics can be used in all stages of for instance cancer care, from screening and staging to recurrence risk assessment[15].

4.2.1 PyRadiomics

PyRadiomics is a specific open-source Python package used to extract a wide range of radiomic features [64]. They are well documented by Griethuysen *et al.*, where they explain all the different settings that can be used in their package for the feature extraction [64]. It is important to note that radiomics extraction tools do not all extract the same set of features. Some tools include unique features or provide users with a tailored subset of features designed for specific applications. In this context, we will examine the standard settings of PyRadiomics, which consist of seven distinct feature classes: First-order statistics, shape-based features, Gray Level Co-occurrence Matrix, Gray Level Run Length Matrix, Gray Level Size Zone Matrix, Neighboring Gray Zone Difference Matrix, and Gray Level Dependence Matrix. The last five feature classes are often referred to as the texture features, as they give more information about the spatial distribution and patterns of the pixel or voxel intensities within the ROI. We will describe all the different feature classes in the PyRadiomics package into more depth.

First-order statistics

The First-order statistics describe the distribution of the voxel intensities within the ROI. They do not take into consideration the spatial distribution of the voxels. The features in this class are derived from the first-order histogram of the voxel intensities. A total of 18 different first-order features are calculated, among which the mean and median value of the voxel intensity as well as the energy, interquartile range and many more.

Shape-based features

The shape-based features describe the three-dimensional shape and size of the ROI. They give features such as volume, surface area, maximal diameter, compactness and more. The shape-based features are independent of the voxel intensities, as they only look at the shape of the ROI.

The three-dimensional surface is created around the ROI using the Marching Cubes algorithm, developed by William E. Lorensen and Harley E. Cline in 1987 [65]. In essence, what the algorithm does is create a surface by interpolating between two slices. Each image slice consists of pixels with four corner points, such that between two slices a cube can be formed consisting of eight corner points. Each of the eight corners either lies within the ROI or not, giving a total of $2^8 = 256$ possibilities that the ROI can pass through the cube. The points that do not lie within the ROI are used to create vertices that will form the surface mesh. However, since the cube is symmetric, the number of possibilities can be reduced to 14 distinct options. We will not provide a detailed explanation of the Marching Cubes algorithm, as it is beyond the scope of this thesis.

The closer the pixels are spaced to each other, the finer the mesh grid obtained from the marching cubes algorithm will be. Hence, the surface will be drawn more accurately when the number of pixels in the image is high. The features will therefore be impacted by the pixel spacings and slice thickness used by the scanner.

Gray Level Co-occurrence Matrix (GLCM)

The Gray Level Co-occurrence Matrix (GLCM) is the first subgroup of the texture features [66, 67]. These texture features are less intuitive and more challenging to explain, so examples will be provided to clarify their meaning. The matrix calculates how often a certain pair of pixel values separated by a certain distance δ and along an angle θ appears. In the PyRadiomics package, the GLCM is always symmetrical, meaning that the order in which a pair occurs doesn't matter.

Example 1. *Assume we have a two-dimensional square ROI containing gray values given in Figure 4.1.*

1	1	3	7	4
2	3	5	4	3
6	7	2	6	1
2	6	3	2	1
3	3	4	5	2

Figure 4.1: Two-dimensional square ROI with gray values ranging from 1 to 7.

The GLCM with a distance $\delta = 1$ and angles $\theta = 0$ and $\theta = \pi/2$ is a 7×7 matrix with 7 different gray levels. For the angle $\theta = 0$ horizontal co-occurrences are considered (See Figure 4.2a), whereas for the angle $\theta = \pi/2$ vertical co-occurrences are considered (see Figure 4.2b).

	1	2	3	4	5	6	7
1	2	1	1	0	0	1	0
2	1	0	2	0	1	2	1
3	1	2	2	2	1	1	1
4	0	0	2	0	2	0	1
5	0	1	1	2	0	0	0
6	1	2	1	0	0	0	1
7	0	1	1	1	0	1	0

(a) Caption

	1	2	3	4	5	6	7
1	2	2	2	0	0	0	0
2	2	0	2	0	2	3	0
3	2	2	0	2	1	1	1
4	0	0	2	0	0	1	1
5	0	2	1	0	0	0	0
6	0	3	1	1	0	0	1
7	0	0	1	1	0	1	0

(b) Caption

Figure 4.2: Gray Level Co-occurrence matrix from the gray values in the ROI given in Figure 4.1 for a distance $\delta = 1$ and angles $\theta = 0$ (a) and $\theta = \pi/2$ (b).

To illustrate how the values in the GLCM matrix are calculated, we focus only on the angle $\theta = 0$, where we look at horizontal co-occurrences of 1 vs 1 and 2 vs 3. For the co-occurrence of 1 vs 1, we see in Figure 4.1, that only in the left upper corner the case of 1 vs 1 is present and the first element of the GLCM matrix will therefore be 2. Remark that here (and since we are considering a symmetrical GLCM) we have 1 vs 1 from both left to right and right to left viewpoint (see Figure 4.3). In the case of 2 vs 3, we can identify two cases, once from left to right and once from right to left (see Figure 4.4), such that the GLCM element at the second row and third column and third row and second column is two.

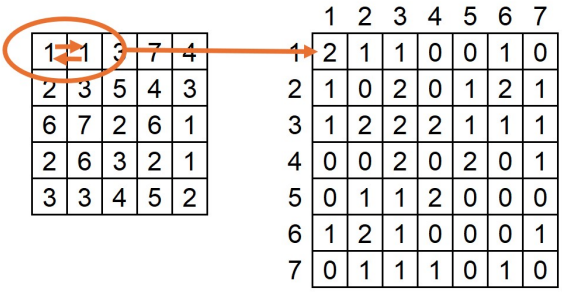


Figure 4.3: In the ROI we see 1 vs 1 from left to right and right to left.

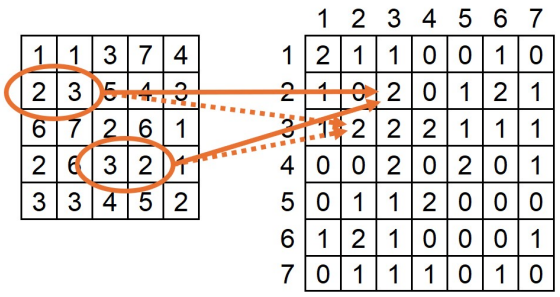


Figure 4.4: In the ROI we see the pair 2 vs 3 occur two times.

From this GLCM matrix several features, like the joint average, correlation, and autocorrelation, can be extracted. In the PyRadiommics package, by default, the features are calculated for each angle separately and then averaged.

Gray Level Run Length Matrix (GLRLM)

The Gray Level Run Length Matrix (GLRLM) is a second subgroup of the texture feature. This matrix calculates how often certain pixels will appear next to each other along a certain angle. If e.g. pixel number one appears only once four times next to each other, in the GLRLM, it will be visible as the element on the first row and the fourth column will be one. From this matrix there is again a number of features that can be extracted like the short and long run emphasis and the non-uniformity of the gray levels.

Example 2. *Consider again the two-dimensional ROI defined in example 1 (Figure 4.5).*

1	1	3	7	4
2	3	5	4	3
6	7	2	6	1
2	6	3	2	1
3	3	4	5	2

Figure 4.5: Two-dimensional square ROI with gray values ranging from 1 to 7.

The GLRLM matrix consists of the number of times a certain levels intensity appears consecutive next to each other. For this ROI, we obtain a GLRLM matrix for an angle $\theta = 0$ (horizontal) given in Figure 4.6.

Intensity Level \ Consecutive obs.	1	2	3	4	5
1	2	1	0	0	0
2	5	0	0	0	0
3	4	1	0	0	0
4	3	0	0	0	0
5	2	0	0	0	0
6	3	0	0	0	0
7	2	0	0	0	0

Figure 4.6: GLRLM matrix for the ROI defined in Figure 4.5.

We observe that the GLRLM matrix for ROI contains many zero values, indicating that not many intensity levels with the same value appear next to each other. We will explain the process of creating the GLRLM for intensity level one. In the ROI matrix, we can see that the value one appears two times in the matrix without the same level of one appearing next to it horizontally. Furthermore, we see that the only other observation of one in the matrix is given in the first row of the matrix where we can see that one appears two times next to each other (see Figure 4.7).

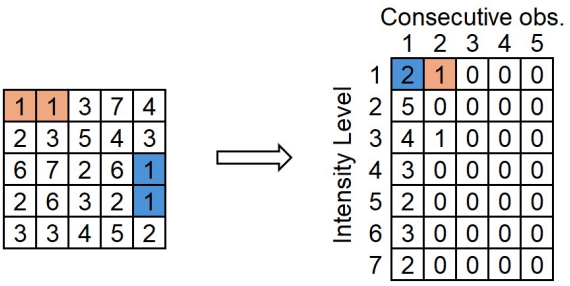


Figure 4.7: The number of times that the intensity level one is observed without another intensity level one left or right to it is given in orange. The number of times that the level was observed with two times next to each other is given in blue.

Gray Level Size Zone Matrix (GLSZM)

A third subgroup of the texture features are features extracted from the Gray Level Size Zone Matrix (GLSZM). The GLSZM determines the gray level zones that are present in a ROI of an image. The zones are defined by the number of voxels with the same gray level intensity that are connected to each other. In the three-dimensional case, there are 26 possible connected regions, whereas in the two-dimensional case, only 8 connected regions are present. From the GLSZM matrix, we can extract features like the short and long area emphasis and the non-uniformity of the gray levels.

Example 3. *Consider again the two-dimensional ROI defined in example 1 (Figure 4.8).*

1	1	3	7	4
2	3	5	4	3
6	7	2	6	1
2	6	3	2	1
3	3	4	5	2

Figure 4.8: Two-dimensional square ROI with gray values ranging from 1 to 7.

The GLSZM is similar to the GLRLM with the difference that connections can go in any direction. For our example, most of the intensity levels are connected with maximally eight surrounding intensity levels. The size of the GLSZM matrix depends on the maximal connections over all intensity levels. For our ROI we observe a GLSZM matrix given in Figure4.9.

Connections

Intensity Level	1	2	3
1	0	2	0
2	2	0	1
3	1	1	1
4	1	2	0
5	2	0	0
6	1	1	0
7	2	0	0

Figure 4.9: The GLSZM matrix for the ROI given in Figure 4.8.

Looking at the intensity level three, we can observe, in our ROI, one case where no other pixel with level three is connected to a pixel with level three, one case where two pixels with the same intensity level of three are connected, and one case where three pixels with intensity level three are connected (See Figure 4.10).

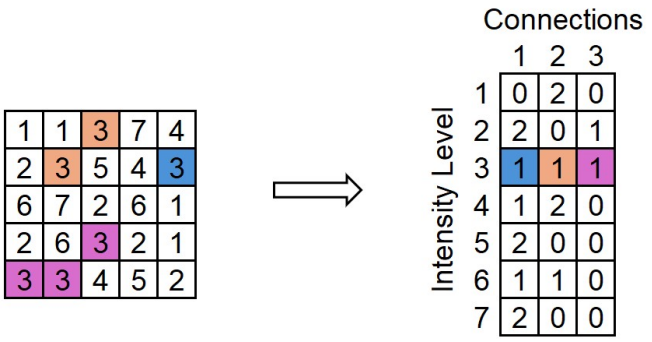


Figure 4.10: The number of connections for an intensity level of three.

Neighboring Gray Zone Difference Matrix (NGZDM)

The fourth and most complicated subset of the texture features are those extracted from the Neighboring Gray Zone Difference Matrix (NGZDM). Despite what the name suggests, this is more a table than a matrix. Consider a segmented image ROI $\boldsymbol{x} = x(i,j,k)_{i,j,k}$ consisting of gray levels. It first calculates how frequently a certain gray value i appears in the image $\boldsymbol{x}$, n_i, and then it calculates the frequency that it occurs, p_i. Lastly, we calculate the average gray level in the neighborhood of voxel $x(i,j,k)$. This means that the average gray value for the surrounding voxels of $x(i,j,k)$ is calculated. Mathematically, this can be denoted by,

$$A_(i,j,k) = \frac{1}{W} \sum_{m_i=-\delta}^{\delta} \sum_{m_j=-\delta}^{\delta} \sum_{m_k=-\delta}^{\delta} x(i+m_i, j+m_j, k+m_k), \tag{4.3}$$

where W is the number of voxels in the neighborhood of the (i,j,k)-th voxel. This average is then used to calculate the sum absolute difference between the gray levels of pixels with the same gray level and the average. For the gray levels with value m we get then

$$s_m = \sum_{\{(i,j,k)|x(i,j,k)=m\}} |m - A(i,j,k)|. \tag{4.4}$$

Using the values of n_i, p_i and s_i for gray level value i, several features can be calculated like the complexity, contrast, and strength.

Example 4. *Consider again the two-dimensional ROI defined in example 1 (Figure 4.11).*

1	1	3	7	4
2	3	5	4	3
6	7	2	6	1
2	6	3	2	1
3	3	4	5	2

Figure 4.11: Two-dimensional square ROI with gray values ranging from 1 to 7.

The NGTDM matrix is actually a table instead of a matrix. The NGTDM for the first three intensity levels is given in Table 4.1.

i	n_i	p_i	s_i
1	4	0.16	9.900
2	5	0.20	8.767
3	6	0.24	5.417

Table 4.1: NGTDM 'matrix' for the ROI given in Figure 4.11

The value for s_1 is calculated using the formula:

$$s_1 = \left|1 - \frac{6}{3}\right| + \left|1 - \frac{14}{5}\right| + 2\left|1 - \frac{16}{5}\right| = 9.9. \tag{4.5}$$

Looking at the first row and columns of our ROI matrix. We see that this element of intensity level one is surrounded by 3 pixels, hence the denominator in the first fraction of the absolute value. Furthermore, if we sum all the intensity values of the surrounding pixels we get $1+3+2 = 5$ *which is the numerator. We calculate the fraction in the same manner for all pixel values.*

Gray Level Dependence Matrix (GLDM)

The fifth and last subgroup of the texture features are the features extracted from the Gray Level Dependence Matrix (GLDM). This matrix gives the gray level dependencies in the ROI of the image. The dependency is the number of voxels connected within a distance δ dependent on the center voxel. the $(i,j)-$th element of the GLDM matrix is given by the number of times a voxel with gray value i appeared with j dependencies, i.e. $|i-j| \leq \alpha$, for a chosen α.

Example 5. *Consider once more the two-dimensional ROI defined in example 1 (Figure 4.12).*

1	1	3	7	4
2	3	5	4	3
6	7	2	6	1
2	6	3	2	1
3	3	4	5	2

Figure 4.12: Two-dimensional square ROI with gray values ranging from 1 to 7.

The GLDM for $\alpha = 0$ and $\delta = 1$ is given in figure 4.13.

	0	1	2	3	4
1	0	4	0	0	0
2	2	2	1	0	0
3	1	4	1	0	0
4	1	2	0	0	0
5	2	0	0	0	0
6	1	2	0	0	0
7	2	0	0	0	0

Figure 4.13: The GLDM for the ROI given in Figure 4.12.

We will consider the intensity level one in this example. We can identify four pixels with intensity level one in the ROI (see figure 4.14).

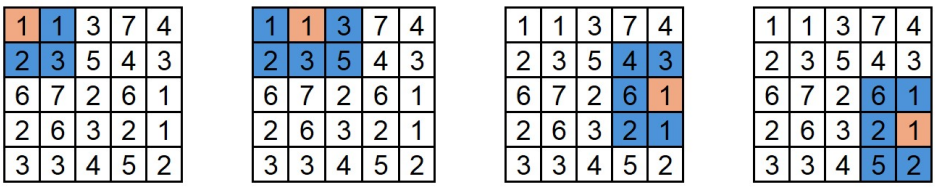


Figure 4.14: The four pixels highlighted in orange with intensity level one and the surrounding pixels highlighted in blue.

We see that for all four cases, there is each time one pixel with intensity level one around the pixel we are considering. We therefore denote on the first row and second column of the matrix the number four. The same is done for all other intensity levels.

4.2.2 Successes with radiomics

Over the years, many studies demonstrating promising prediction results have been published (see Table 4.2). In 2016, C. Liang *et al.* published an article investigating the predictive ability of radiomics for staging primary CRC [26]. Their study included a total of 494 patients with stage I-IV CRC. Using LASSO logistic regression, they identified 16 predictive radiomics. They achieved a training AUC of 79.2% and a validation AUC of 70.8%, indicating that radiomics can be effectively be used for staging CRC. Another example of successful radiomics can be found in an article by H. Yu published in 2021, which looked at 121 patients with breast cancer [68]. The authors extracted a total of 612 radiomics from the patient's mammography using PyRadiomics. Their aim was to predict the tumor-infiltrating lymphocyte levels categorized as low and high levels. Their classification results showed a training accuracy of 70.6% and a validation accuracy of 63.9%.

Lastly, Brunese *et al.* reported exceptionally good results in their study. They extracted radiomic features from MRI scans of patients with brain cancer with the purpose of predicting the tumor grade using ensemble learning with radiomics [69]. They achieved an average accuracy of 99.0% across the different brain cancer grades. However, despite the impressive results, they did not validate their model, which is crucial for generalizing results. We will focus on both training and validation of our method as we aim to work towards a method that is applicable in real-world settings.

Citation	**Description**	**Results**
C. Liang *et al.* [26]	Staging CRC using radiomics	Training AUC = 79.2%; Validation AUC = 70.8%
H. Yu [68]	Prediction of tumor-infiltrating lymphocyte levels for patients with breast cancer	Training accuracy = 70.6%; Validation accuracy = 63.9%
L. Brunese *et al.* [69]	Predicting the tumor grade for brain cancer patients using radiomics	Average accuracy = 99.0%

Table 4.2: Successful applications of radiomics in medicine.

4.2.3 Challenges and problem with radiomics

Even though radiomics are frequently used in the medical world and are yielding promising results, it is essential to acknowledge that they still have some limitations that need to be addressed.

One of the challenges reported in radiomics data is the issue of repeatability and reliability. Repeatability refers to the consistency of features when scans are repeated under the same conditions, while reliability refers to testing the consistency of the features when different scanners and/or settings are used. Varghese *et al.* conducted a phantom study to evaluate the repeatability and reliability of the radiomic features [70]. For repeatability, where scans were retaken with a 15-minute interval, they found that repeatability is dependent on the scanner model. They measured the repeatability and reliability score, which is the percentage of similar variables. The Philips Brilliance 64 CT scanner had a better repeatability score of 97.08% compared to the Toshiba Aquillion Prime 160 CT scanner, which had a repeatability score of 74.01%. Texture features, in particular, showed poor repeatability for both scanners. Regarding reliability, they tested 21 different settings for the Philips CT scanner and 16 for the Toshiba CT scanner. They found that changing the settings had varying effects on the consistency of radiomics. For example, changing the field of view resulted in a reliability of 97.02% whereas for changing the

slice thickness yielded a much lower reliability of only 8.93%. In conclusion, they showed that the repeatability of radiomics depends on the scanner type, and that the settings are important for the consistency. However, they did not test whether the different scanners between themselves had a good reliability or whether different slice thicknesses posed a problem.

Another problem that cannot always be avoided is the problem of contrasts and artifacts. Both will alter the gray-values of the data and will therefore generate incorrect radiomics [71]. The researcher should always check if and when contrasts were used previous to their scan and whether there would be still some contrast left in the patient's body if contrasts were used. For artifacts, researchers are facing some more problems as not all artifacts can be removed. One could try to counteract the artifacts and retrieve/calculate the correct gray-value score.

High correlation between features also causes a challenge in radiomics studies [72]. With numerous radiomics extracted from CT scans, models may have more or equal features than samples, leading to multicollinearity issues. We must check for multicollinearity and either select algorithms immune to it or reduce the number of highly correlated variables through feature selection, which also helps mitigate overfitting risks.

The manner in which the region-of-interest (ROI) is delineated on the image also significantly impacts the stability of radiomics [73]. Images are typically preprocessed using either manual, semi-automated or automated segmentation techniques. Manual segmentation, although time consuming, is prone to human errors as physicists manually delineate the ROI slice by slice. Inter-observer manual segmentations have been shown to result in significant differences in the drawn ROIs, consequently leading to variations in the extracted radiomics. However, manual segmentation remains the standard method for segmenting medical images. In the future it would be beneficial to eliminate some of the noise that originates from human errors and subjectivity when drawing in the ROI manually. Currently, some research is being done, by the research group BISI at the Vrije Universiteit Brussel (VUB), looking at (semi-)automatic segmentation algorithms on CT images for rectal cancers.

To reduce the variability in the radiomics, researchers can opt for (semi-)automated segmentation methods. While many (semi-)automated segmentation algorithms demonstrate good reproducibility, surpassing manual segmentation in some cases, it is important to note that the repeatability of radiomics using different automated segmentation algorithms may not always be as robust. For certain segmentation methods, radiomics can vary by up to 50% [73].

4.3 Texture analysis in CT

Texture features, like radiomics, are frequently utilized to extract quantitative data from medical images such as CT scans. These features have a wide range of applications, including optimizing disease diagnosis, predicting tumor grades, and more. In this section, we will provide an overview of several studies that have used texture features extracted from CT scans (see Table 4.3).

First, let's look at a study from S. G. Mougiakakou *et al.*, where they aimed to extract different texture features—first-order statistics, spatial gray level dependence matrix, gray level difference method, Laws' texture energy measures, and fractal dimension measurements—from different regions of interest (ROIs) drawn on non-enhanced liver CT images to create an optimal computer-aided diagnosis (CAD) architecture [74]. The goal was to correctly diagnose liver tissues as normal liver tissue, hepatic cyst, hemangioma, and hepatocellular carcinoma. They used several ensembles of classifiers and obtained the best mean results with a training accuracy of 100%, a validation accuracy of 88.43%, and a test accuracy of 84.96%, proving that they were successful in using texture features for diagnosing liver tissues.

Secondly, texture features have also been applied in predicting the grade of pancreatic neu-

roendocrine tumors (PNETs). R. Canellas *et al.* conducted a study to assess the predictive value of both texture features and CT features—such as tumor location, size, and pattern—for diagnosing PNETs [75]. The researchers aimed to determine which features were most predictive and to evaluate the combined performance of these features. By extracting 36 texture features from the images, they found that combining them with CT features resulted in an accuracy of 79.3%.

Lastly, Y. Zeng *et al.* used radiomics extracted from ankle CT images to classify gout versus non-gout patients [76]. They identified five radiomic features that were significant for distinguishing between gout and non-gout cases. Using these five features in a logistic regression model, they achieved an accuracy of 74.0%. When applying machine learning techniques—Random Forest, XGBoost, and SVM—they found that the accuracies improved to 90.1%, 83.3%, and 87.5%, respectively, demonstrating that the Random Forest algorithm produced the best results for their prediction problem.

Citation	**Description**	**Results**
S. G. Mougiakakou *et al.* [74]	Computer-aided diagnosis architecture for the classification of liver tissue	Mean training accuracy = 100% Mean validation accuracy = 88.43% Mean test accuracy = 84.96%
R. Canellas *et al.* [75]	Grade prediction of pancreatic neuroendocrine tumors (PNETs) using texture and CT features	accuracy = 79.3%
Y. Zeng *et al.* [76]	Statistical and Machine Learning methods for the classification of gout vs non-gout using CT radiomics	Accuracy: Logistic regression = 74.0% Random Forest = 90.1% XGBoost = 83.3% SVM = 79.3%

Table 4.3: Successful applications of radiomics in medicine.

Chapter 5

The Evolutionary Random Subspace Forest

This chapter is based on the publication:
C. Raets, C. El Aisati, M. De Ridder, A. Sermeus, and K. Barbé, "An Evolutionary Random Forest to measure the Dworak tumor regression grade applied to colorectal cancer," *Measurement*, vol. 205, pp. 112–131, Nov. 2022.

Abstract

In order to optimize a patient's cancer treatment, the prediction of the tumor's response to a planned radiotherapy is vital. For patients with rectal cancer, the quality of life depends heavily on the surgery leaving some patients with a stoma or other problems. If we would be able to predict the response upfront, we would be able to personalize the treatment and therefore preserve the QOL as much as possible. The Dworak tumor regression grade is a typical diagnostic tool to assess the tumor response of colorectal cancer patients. However, the Dworak grade is determined by a pathologist by inspection of the tumor biopsy without a dedicated measurement instrument. The measurement of the Dworak grade in an automated way by using the pre-operative CT scans, is a challenge. In this paper, we propose a novel methodology to measure the Dworak grade based on a customized Random Forest.

We created a new Random Forest based on the methods from Breiman and Ho. These methods are further enhanced by evolutionary computation. We give the results of both our new method and other classical classification methods and highlight that the choice of classifier and knowledge is crucial.

We extracted 111 radiomic features from 141 patients with colorectal cancer, of which 97 bad and 44 good responders. The evaluation gave a 77.507% accuracy and the cross-validation used for validation gave a 67.081% accuracy. Our results were the best compared with some other classification methods.

5.1 Introduction

Rectal cancer patients are typically initially treated with radiation therapy, sometimes in combination with chemotherapy, followed by a surgical procedure. During surgery, a biopsy is taken and sent to the lab for pathological examination. From the biopsy, pathologists determine the regression grade of the patient, which measures how well the tumor has responded to the treatment. Unfortunately, this TRG can only be calculated when a biopsy from the surgery is taken, and since these surgeries are highly invasive, minimizing them is a priority.

Consequently, there is a pressing need to explore methods capable of predicting the Dworak TRG before the start of the therapy. By extracting quantitative data from planning CT images, which are obtained before the start of the therapy, we aim to develop a prediction algorithm for estimating the Dworak TRG. Predicting the Dworak TRG prior to start of the therapy and surgery enables us to identify patients who are likely to respond well to treatment, allowing for a wait-and-watch approach. Opting for this approach helps us avoid possible side effects and diminished QOL associated with the surgical invasiveness. Furthermore, avoiding unnecessary surgeries also leads to a more efficient allocation of healthcare resources.

Furthermore, by integrating additional clinical and biomedical data such as radiation dose, fraction plan, chemotherapy type, etc., a step can be made towards more personalized treatment, improved patient outcomes, and optimized healthcare resource utilization.

In this thesis, we introduce our custom-created ML algorithm for predicting the TRG of rectal cancer after neoadjuvant (chemo)radiation therapy. We will categorize the Dworak TRG into a binary classification, distinguishing between bad and good responders. To enhance the robustness of our ML algorithm and mitigate overfitting, we will employ a variable selection technique.

Our approach used a custom-built version of a Random Forest (RF), originally introduced by Ho in 1995 under the name of *Random Decision Forest* [77]. A RF is an ensemble method used for classification or regression tasks. It builds multiple decision trees—flowchart-like models designed to solve classification or regression problems—and combines their results to produce a single, more robust outcome. Both classification and regression trees are carefully explained in the eponymous book of Leo Breiman [78]. In 2001, Breiman proved that a RF is less prone to overfitting the data when the number of trees in the forest increases [79]. We will use the RF classification idea as described by Ho to perform variable selection and to enhance the prediction of the Dworak regression grade of patients after (chemo)radiotherapy [77].

Our RF algorithm is evolutionary in nature: with each iteration, the forest improves by obtaining more knowledge about predictor variables and possible linear combinations thereof. Additionally, it eliminates trees performing less effectively than newly discovered ones.

Moreover, we incorporate Linear Discriminant Analysis (LDA) as a classifier in our algorithm. Initially introduced by Fisher in 1936, LDA is an intuitive method for straightforward classification [80].

It is important to note that the algorithm presented in this thesis is tailored specifically to our data set. Unfortunately, ML algorithms are not universally applicable magic tools. Instead, their parameters must be specifically tuned to the characteristics of the data.

5.2 Problem statement, data and methods

5.2.1 Problem statement

The goal of this chapter is to investigate whether the Dworak regression grade, as defined in Section 2.4.2, can be predicted for patients with rectal cancer following (chemo)radiotherapy

using radiomics derived from their planning CT scans. If we were able to predict the regression grade, we would be able to assess ahead of time whether surgery may be avoided. It is important to note that it is absolutely not the point of the study to replace the discernment of the doctor, and that prediction models should only be seen as a tool to help doctors in the decision-making process.

5.2.2 Redefining the Dworak Regression Grade

The definition of the Dworak TRG, provided in Section 2.4.2 and Table 2.6, is clearly subjective, leading to variations in the grade assigned by different pathologists. In fact, a given pathologist may even change opinion over time, resulting in surgical error. Again, these errors arise because the grades depend solely on the pathologist's visual observations [81]. The bottom line of this is that, the Dworak TRG has some significant limitations due to its variability. Several articles have suggested reducing the number of grades to specifically reduce this variability [82, 83, 84]. We adopted a dichotomous grading system (see Figure 5.1), dividing patients into bad responders (Dworak grades 0-1-2) and good responders (Dworak grades 3-4). We refer to this classification system as the regrouped Dworak TRG. Nonetheless, residual variability in the regrouped TRG remains, presenting challenges in constructing a classification algorithm.

	Dworak Regression Grade:	
Bad responder	TRG 0	No regression;
	TRG 1	Dominant tumor mass with obvious fibrosis and/or vasculopathy;
	TRG 2	Dominantly fibrotic changes with few tumor cells or groups that are easy to find;
Good responder	TRG 3	Very few and microscopically difficult to find tumor cells in fibrotic tissue with or without mucous substance;
	TRG 4	No tumor cells, only fibrotic mass resulting in a total response.

Figure 5.1: Regrouped Dworak Regression Grade.

5.2.3 Radiomics Features

Today, CT imaging is an integral part of your typical radiotherapy workflow, specifically for the accurate segmentation of internal volumes and precise dose calculations. In this chapter, we analyze Gross Tumor Volume (GTV) segmentations performed on the CT images of colorectal cancer patients—the very same images used to plan their radiation therapy—by extracting quantitative features from these segmentations. The features, called radiomics, can then be used in numerous statistical tools for deeper analysis. In our case, this will consist of building prediction models of the Dworak TRG. Radiomics are becoming more and more popular as a feature extraction technique as they are easy to extract and have been proven to be very useful in previous research (see Chapter 4).

We extracted the following radiomic features with the Python package PyRadiomics [64]. This open-source package divides the features in seven classes: shape-based features, first-order features, Gray-Level Co-occurrence Matrix features (GLCM), Gray-Level Run Length Matrix features (GLRLM), Gray-Level Size Zone Matrix features (GLSZM), Neighboring Gray-Tone Difference Matrix features (NGTDM), and Gray-Level Dependence Matrix features (GLDM).

The shape-based features extract quantitative data from the 3D shape, such as the surface area, diameters, volume, etc. These features are calculated using the marching cube algorithm, introduced in 1987 by Lorensen and Cline [65]. Using linear interpolation, the marching cube algorithm generates triangle vertices to form a surface around the region-of-interest (ROI), enabling calculations such as diameter determination.

The first-order features do not rely on 3D information. Instead, they utilize voxel intensities given in Hounsfield Units to derive more straightforward and intuitive features such as maximum and minimum voxel intensity, mean voxel intensity, and so forth. A right-skewed histogram would indicate that voxel intensities tend to lean toward lower Hounsfield Units, resembling air, while a left-skewed histogram would suggest intensities leaning toward higher Hounsfield Units, resembling bone structures. The frequency of intensity occurrences could thus provide insights into the structure of the substance present in the image, which could be valuable in predicting the Dworak TRG.

The third feature class, GLCM, provides features derived from the square Grey Level Co-occurrence matrix with a size equal to the number of discrete intensity levels present in the image. Features of this square matrix, which is symmetric by default, are calculated for all angles separately, returning the average over all angles as the feature value. The $(i, j$-th element in the GLCM matrix $P(i, j|\theta)$ represents the number of times the i-th discrete gray value, with a distance of δ pixels to the j-th gray level value, occurs along an angle θ [64]. By default, the distance δ is set to one, considering only adjacent gray level intensities.

GLSZM features analyze gray level zones in the image. These zones comprise connected voxels with the same gray level intensity, where the $(i, j$-th matrix element equals the number of zones with gray level i that are connected j times. Another feature class is GLRLM, which calculates features from the Gray Level Run Length Matrix (GLRLM). This matrix resembles GLSZM, with the $(i, j$-th element representing the number of occurrences of the i-th gray level appearing consecutively for j times in the image ROI. The key difference is that GLRLM considers an angle θ along which connected gray values must be found, unlike GLSZM, where no angle is specified, allowing for angle variations within zones of equal gray values.

Next, we have the NGTDM features, which are derived from the Neighboring Gray Tone Difference matrix. This matrix quantifies the difference between a gray value and the average values of its neighbors within a distance δ. The NGTDM's first column comprises all discrete gray level values in the image, while the second column contains the frequency of each gray level's occurrence. The third column holds the probabilities of encountering a certain gray level, calculated by dividing the second column by the total number of voxels in the image. In the last column, the element s_i for the i-th discrete gray level is zero if the i-th gray level is absent in the image; otherwise, it is calculated as the sum over all gray levels i of the absolute difference between i and the average of all gray levels around that gray level i.

Lastly, there are the GLDM features, calculated on the Gray Level Dependence Matrix. As per the PyRadiomics Documentation, the GLDM matrix captures gray level dependencies in the image, where dependency is defined as the number of connected voxels within a distance δ that depend on the center voxel. For a gray level j to be dependent on the center voxel i, $|j - i| \leq \alpha$ for a certain α.

Altogether, these features give a rather complete characterization of the 3D segmentations extracted from the CT scan images— in our case, these will be rectal tumor segmentations. The mathematical definition of all features is standardized in a dictionary from the Image Biomarker Standardization Initiative (IBSI), so as to standardize definitions across studies [85].[1]

[1] nog figuren toevoegen

5.2.4 Medical data

A retrospective data set, with ethical committee approval number EC1010135, was constructed at the University hospital of Brussels (UZ Brussel). It consists of $n = 141$ patients with colorectal cancer who were admitted to the hospital to undergo chemo- and radiotherapy. All patients had a planning CT scan taken prior to their radiotherapy (see Fig. 5.2 for an illustrative example). All the CT images in the data set were taken between 2005 and 2017. From each CT scan a

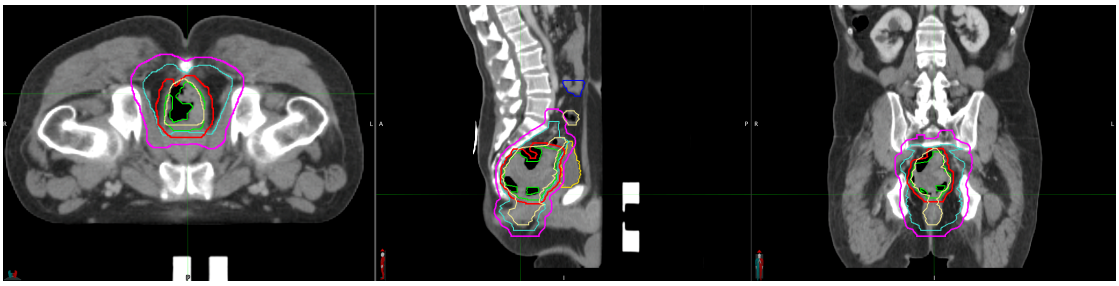


Figure 5.2: Example of a planning CT from a patient included in the study. From left to right, we respectively show an axial, sagittal and coronal view of the patient. The different colored lines in the images correspond to the delineations performed by the doctor (e.g. gross tumor volume, expansions of it, organs at risk). The rectal cancer is clearly visible in these images (presence of a mass of tissue in the rectum, highlighted by the green contour). The radiomics are solely calculated on the gross tumor volume.

total of 109 radiomics features were extracted using the default settings of the PyRadiomics Python package. We added two more variables to the features set, one where we divided the energy level by the number of voxels and one where we divided the total energy by the number of voxels. All together, this resulted in $p = 111$ radiomic features. Since some parameters were not normally distributed, we tried to find a good transformation of the parameters to obtain normality. The normality of the prediction parameters is required for the Linear Discriminant Analysis (see Sec. 5.3.1). Figure 5.3 gives an example of such a transformation, where the variable *GLDM Small Dependence Low Gray Level Emphasis* was transformed using the log-function. In the Appendix, we have listed all the radiomic features and the transformations needed for normalization in Table 5.4. Some variables required no transformations, while for others that did, no suitable transformation was found.

Of the 141 patients, 90 are male and 51 are female. The average age is 65 years old, with the youngest patient being 33 years old and the oldest patient being 86 years old. In Table 5.1 some descriptions are given regarding the monitored blood values obtained from the patients. The clinical reference blood values used in the hospital are also displayed for comparison. We can see that several patients have blood values that do not lie within the reference ranges (e.g. C-reactive protein (CRP) and the Platelet Count Test (PLT)).

The Dworak TRG varies among patients, with 46 patients exhibiting TRG 1, 51 patients exhibiting regression grade 2, 33 patients exhibiting TRG 3, and only 11 patients exhibiting TRG 4. Notably, there are no patients with a regression grade of 0 in the data set.

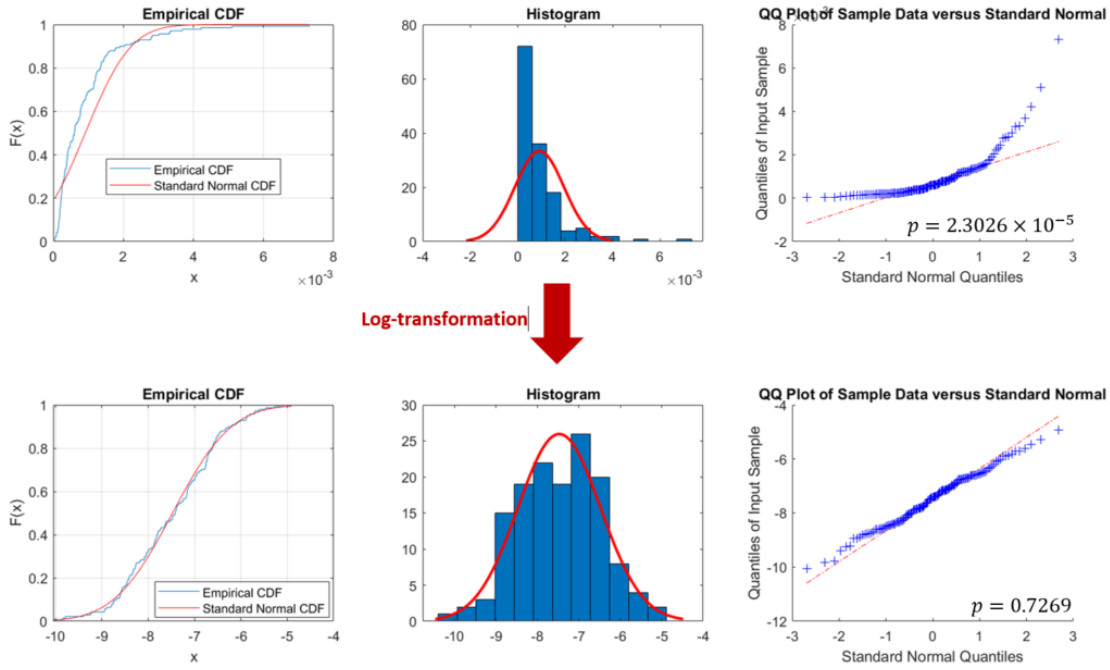


Figure 5.3: (Top) Distribution of the variable *GLDM Small Dependence Low Gray Level Emphasis*. The empirical distribution function and the normal cumulative distribution function (CDF) are shown on the left graph, respectively in blue and red. The histogram with normality curve (right) are displayed in the middle and the qq-plot is displayed on the right. All three graphs clearly differ from the normal distribution. The Kolmogorov-Smirnov test has a p-value of $p = 2.3026 \times 10^{-3}$ such that the null-hypothesis is. (Bottom) After a log-transformation of the variable, the empirical distribution function, the histogram and the qq-plot now resembles more a normality distribution and the p-value is now $p = 0.7269$ such that the null-hypothesis is not rejected.

Parameter	N	Mean	Std	Minimum	Maximum	Reference
HB (g/dl)	105	13.450	1.982	8.6	18	$13 - 16.5$
neutro $(\times 10^3/mm^3)$	96	5.101	2.254	1.047	14.3	$1.4 - 6.7$
lympho $(\times 10^3/mm^3)$	96	2.035	4.420	0.637	21.28	$1.2 - 3.5$
WBC $(\times 10^3/mm^3)$	103	8.103	3.312	3.6	28	$3.6 - 9.6$
PLT $(\times 10^3/mm^3)$	105	287.18	110.245	37	786	$158 - 341$
CEA $(\mu g/l)$	94	7.194	11.957	0.39	79.5	$0 - 3$
CRP (mg/l)	81	10.668	92.8	0	92.8	< 5
Albumine (g/l)	67	40.528	5.812	24	57.3	$36 - 50$
Age	134	65.437	11.087	33	86	/

Table 5.1: Descriptive statistics of the patient's age and blood values in the rectum study.

The number of observations per TNM staging, chemotherapy and surgery are given in Tab. 5.2.

Parameter	Category		Observations
T	T2		11
	T3		117
	T4		13
N	N0		17
	N1		51
	N2		72
	N3		1
M	M0		128
	M1		10
Preoperative chemo	No		89
	Yes	5-FU	10
		Xeloda	33
		Capecitabine	2
Type of surgery	Abdominoperineal res.		30
	Anterior resection		8
	TME		75
	PME		6

Table 5.2: Number of observations (last column) per parameter (columns 1 and 2) for the patients in the study.

5.2.5 Mathematical problem

As our data is high dimensional and contains several radiomic features that are highly correlated with one another, performing a reduction of the variables at hand is a good idea, if not necessary. Feature reduction, also called feature selection, was originally only used on small features sizes as the technology was not evolved enough to handle high-dimensional data. Nowadays, we are able to handle bigger data with dedicated algorithms that are able to reduce their dimensionality [86]. Furthermore, feature selection is used to limit over-fitting and control multicollinearity [87]. This is of special importance since the application of the prediction model to new data is where lots of studies fail. In other words, we must find the right classification and feature selection method to build a good prediction model without overfitting the training data, so that the model can later successfully be exported.

Finally, it's important to acknowledge that besides the residual variability in the binary TRG, another potential source of systematic uncertainty may exist in the data due to the lack of information on the CT scanners used. The patients were selected over a span of approximately 20 years, during which different CT scanners with varying settings and operating systems were used. Moreover, we did not test the reproducibility of the analysis for any of the CT scanners as this is a retrospective study, and therefore could not test the stability of the results in these regards. Without reproducible data, it is nearly impossible to get a high prediction rate even if the TRG would be stable. It is therefore not possible to obtain perfect or even great results for the training without causing the validation to be awful and therefore over-fitting the model. This obviously has to be kept in mind, as the primary goal of the machine learning algorithm is to offer the doctor a second opinion which is evidence-based and supported by the data.

5.3 State-of-the-art classifiers

There are many classification techniques that have been developed and that can be found in the literature. To understand the choice of the approach that we will be using for our work, and for the sake of completeness, we give in this section an overview of the state-of-the-art classification

techniques. We respectively discuss non-ensemble and ensemble methods. The reader can jump without loss of generality to the next section if desired.

5.3.1 Non-ensemble methods

Support Vector Machine

Support vector machines (SVM) have been around for over two decades. They have been introduced by Cortes and Vapnik in 1995 as *support vector networks*, nowadays better known as SVMs [88]. SVMs are a generalization of the *maximal margin classifier*. This classifier constructs a linear function or hyperplane that perfectly separates the data, such that the distance of the training data to the separating hyperplane is maximal [89]. The smallest distance to the training data from the hyperplane is called the *margin*. However, the assumption that the training data can perfectly be separated is rarely satisfied. Therefore, new classifiers have been formulated to allow some training points to be on the wrong side of the hyperplane, i.e. to be wrongly classified. Lastly, the possibility to modify the linear hyperplane using kernels has been added, resulting in the modern SVM technique [90]. SVMs are usually successful in high-dimensional data, which is quite common in medical data.

Logistic Regression

Another very popular method for the classification of binary data is the *logistic regression* or *logit*. The idea of the *logistic regression* is to start from the linear regression model and use a so-called link function to transform the range of the linear function to the interval $[0, 1]$ [91]. Several classifiers make use of this approach. For the *logistic regression,* the link function is given by $g(x) = \log(x/(1-x))$. Since the data is transformed to the interval $[0, 1]$, it can be seen as a probability measure such that a prediction can be made using the probability values. Logistic regression provides a classification model with straightforward calculations, making it ideal when interpretability of the model is wanted.

Classification trees

Classification trees rely on splitting data into branches based on predictor variables. This process begins with a root node and progresses by repeatedly dividing branches until reaching terminal nodes or meeting a stopping criterion. Efficient splits at each node are crucial, and various metrics exist in the literature to evaluate the quality of these splits [92], [89], [91]. Note that not all predictor variables need to be used in this process, meaning that it performs some kind of variable selection on its own [78]. An advantage of classification trees is that they are very easy to understand and interpret.

A pioneer in this field, Leo Breiman, significantly influenced the development of regression and classification trees. His eponymous book elaborates on the methodology for constructing and pruning trees, solidifying their theoretical foundations [78].

Hierarchical Clustering

Cluster methods consist of creating subgroups—or clusters—of the data in order to find a particular structure instead of focusing on prediction [89]. Hierarchical clustering starts with all patients separated and then clusters two patients with the most similar features together. The algorithm continues clustering patients or clusters of patients based on specific similarity tests, until there is nothing left to be clustered.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) method classifies the data by looking at the K closest datapoints [89]. A prediction is made by looking at the different classes in the K-nearest neighbors and taking the class which is the most present [92]. The performance of the KNN method heavily depends on the choice of K. When the number of observations per class is small, the algorithm typically fails to correctly converge for large values of K. On the other hand, a really small value of K might not be representative enough. It is therefore crucial to wisely choose the value of K.

Fisher's Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is similar to logistic regression, but has the upside of not requiring any iterative process to predict classes. Although having the advantage of using simple and straightforward calculations, the LDA requires predictor variables to be normally distributed. Therefore, we will assume that the p-dimensional random vector $\underline{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ consists of $p \in \mathbb{N}$ normally distributed random variables. The LDA technique was first introduced by Fisher in 1936, but we will follow the description made by Friedman in 1989 [80, 93].

The goal is to assign the object to a class $m \in \{1, \ldots, M\}$ using the information of the random vector $\underline{X}$. Bayes' rule and the rule of conditional probability state that

$$\mathbb{P}(m|\underline{X}) = \frac{\mathbb{P}(m)\mathbb{P}(\underline{X}|m)}{\mathbb{P}(\underline{X})}, \tag{5.1}$$

where $\mathbb{P}(\underline{X})$ is the prior of $\underline{X}$ and $\mathbb{P}(m|\underline{X})$ the sampling distribution, or data distribution [94]. The risk of wrongly classifying a sample of level m by any other level is simply given by:

$$1 - \frac{\mathbb{P}(m)\mathbb{P}(\underline{X}|m)}{\mathbb{P}(\underline{X})}. \tag{5.2}$$

Because $\mathbb{P}(\underline{X})$ is a constant, the class $\hat{m}$ that minimizes the risk in Eq. (5.2) is given by

$$\hat{m} = \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \mathbb{P}(m)\mathbb{P}(\underline{X}|m). \tag{5.3}$$

Since it is assumed that $\underline{X}$ follows a multivariate Gaussian distribution (see above), the conditional distribution function of $\underline{X}$ is given by:

$$\mathbb{P}(\underline{X}|m) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_m|}} \exp\left\{ -\frac{1}{2} \left(\underline{X} - \underline{\mu}_m\right) \Sigma_m^{-1} \left(\underline{X} - \underline{\mu}_m\right)' \right\},$$

where $\mu_m := \mathbb{E}[\underline{X}|m]$ is the population mean vector for the m-th class and $\Sigma_m := Cov(\underline{X}|m)$ is the population covariance matrix for that m-th class. Using the CDF of $\underline{X}$, Eq. (5.3) may then be rewritten as

$$\hat{m} = \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \frac{\mathbb{P}(m)}{\sqrt{(2\pi)^p |\Sigma_m|}} \exp\left\{ -\frac{1}{2} \left(\underline{X} - \underline{\mu}_m\right) \Sigma_m^{-1} \left(\underline{X} - \underline{\mu}_m\right)' \right\}. \tag{5.4}$$

Maximizing the argument in Eq. (5.4) is the same as maximizing its natural logarithm, with the advantage of simplifying formulas. It is assumed in the LDA that the covariance matrices do not vary over the different classes, i.e. that $\Sigma := \Sigma_m$ for all $m \in \{1, \ldots, M\}$. This, together with the assumption that $\mathbb{P}(m)$ is independent of m, leads to

$$\hat{m} = \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \left(\underline{X} - \frac{1}{2}\underline{\mu}_m\right) \Sigma^{-1} \underline{\mu}_m' \tag{5.5}$$

Unfortunately, the population mean and covariance matrix are rarely known. In practice, we use the sample mean $\hat{\mu}_m$ for each level m and sample covariance matrix $\hat{\Sigma}$ as estimators. These are known to converge to the population mean and covariance with increasing sample size. For each i-th patient, with $i \in \{1, \ldots, n\}$, the predicted group $\hat{y}_i$ is then calculated as,

$$\hat{y}_i = \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \left(\underline{x}_i - \frac{1}{2} \underline{\hat{\mu}}_m \right) \hat{\Sigma}^{-1} \underline{\hat{\mu}}'_m. \tag{5.6}$$

5.3.2 Ensemble methods

The field of ensemble methods has gained a lot in popularity, in part with the revolution of computer technology. The main idea behind ensemble methods is to combine different results such that the new pooled result will perform better than any of the individual ones, improving classification and prediction power.

Leo Breiman's Random Forest

The RF algorithm of Leo Breiman is a non-parametric algorithm as it does not use any distribution of the prediction variables. Breiman described a RF as a collection of trees, which we denote as $\{h(\underline{x}, \Theta_k), k = 1, 2, \ldots\}$, with Θ_k the k-th random vector independently drawn and identically distributed from the previous ones [79]. Furthermore, he pointed out that a RF is less likely to over-fit than one tree, particularly when the number of trees in the forest increases.

For each random vector Θ_k, we choose different bootstrapped learning samples $\mathcal{L}_k$, $k = 1, 2, \ldots$, consisting of $n_k \leq n$ observations independently drawn from the learning sample $\mathcal{L} = \{(\underline{x}_i, y_i) | i = 1, 2, \ldots, n\}$. From these different bootstrapped samples we can create multiple tree classifiers $h(\boldsymbol{x}, \mathcal{L}_k), \{k = 1, 2, \ldots, n_F\}$ with $n_F \in \mathbb{N}$, which are then combined into one forest. A prediction of the Random Forest can be made by aggregating the votes of each tree such that, for the sample $(\underline{x}_i, y_i)$, the most popular prediction among all trees, i.e.

$$\underset{m=1,\ldots,M}{\operatorname{argmax}} \#\{k | h(\underline{x}_i, \mathcal{L}_k) = m\} \tag{5.7}$$

is selected as the final prediction. Breiman coined this method *bagging*, an abbreviation for *bootstrap aggregation* [95].

Random Subspace Decision Forest

The Random Subspace Decision Forest was introduced by Tin Kam Ho in two different papers [77], [96]. She pointed out that trees that are grown too complex might over-fit, and that it is therefore necessary to find a method that overcomes this problem. The solution proposed by Ho independently draws subsets of the predictor set instead of bootstrapping from the learning sample. By bootstrapping from the predictor set, the number of predictors used for each tree classifier is different. Ho constructs many different tree classifiers, where each time a subset of the predictors is bootstrapped, and bundles them together into one forest. For the evaluation of the forest, Ho looks at the probability that an observation belongs to a class m when it reaches a terminal node in the tree. She calls the average probability that a patient belongs to a certain class over the different trees the *discriminant function* of that level. The level with the highest discriminant function is then chosen as the prediction level.

5.3.3 Evaluation of classifiers

When predicting a categorical outcome variable, the percentage of patients that are correctly predicted, defined as the *accuracy* (ACC), can be calculated. A visual representation of the prediction with respect to the observed values can be given with the help of a confusion matrix. The rows and columns of this matrix respectively represent the classes observed and predicted by the classifier, as described in [97]. The accuracy is calculated as the number of correctly predicted patients divided by the total number of patients. In this way, the percentage of misclassification, known as the *error*, is given by $1 - ACC$.

In the case of unbalanced data, the class with more observations in the data might overshadow, in a positive or negative way, the other classes. Since we are interested in the performance of all classes simultaneously, it is a good practice to normalize the confusion matrix row-wise, i.e. to divide each entry by the sum of its corresponding row. As our TRG is unbalanced, we will always use the normalized confusion matrix to calculate the accuracy, which we will call the *proportional accuracy* ($pACC$).

5.4 Novel approach: the Evolutionary Random Forest

We propose a new RF classifier, which we call the Evolutionary Random Subspace Forest (ERSF), that combines the best of Breiman's and Ho's methods: to construct a tree, we bootstrap our data and randomly select subspaces of the predictors, exclusively used for the construction of each given tree. We do not use the standard tree classifier of Sec. 5.3.1, but rather the LDA, making our approach a parametric one. Merging parts of the RF construction methods of Breiman and Ho and additionally improving it with an iterative pruning process results in a new method that is evolutionary as it keeps on learning from the data in each iterative step.

As mentioned, there are different ways to construct a tree. N. Deo describes in his book a more formal definition of a tree is given: a tree is defined as an acyclic connected graph, meaning that it is a graph with no cycles and where there is at least one path between all pairs of vertices [98]. Furthermore, any *rooted* tree, i.e., a tree that has a vertex without an incoming path, is called a *decision* tree or a *classification* tree.

5.4.1 Building a tree

Consider again a data set with design matrix $\boldsymbol{x} = (x_{ij})_{ij} = (\underline{x}_1, \ldots, \underline{x}_n)' \in \mathbb{R}^{(n \times p)}$ consisting of $n \in \mathbb{N}$ patients and $p \in \mathbb{N}$ possible predictor variables and an output vector $\underline{y}$ consisting of the output variable for each patient. Let $\mathcal{L} = \{(y_i, \underline{x}_i) | i \in \{1, \ldots, n\}\}$ be the learning data sample.

Each tree will be constructed by selecting a random subset of predictor variables and using only these variables to construct the tree. For the j-th tree, let Θ_j be a random vector independently drawn from $\Theta_1, \ldots, \Theta_{j-1}$ where Θ_j is a possible subset of the set $\Theta = \{1, \ldots, p\}$ which is a set containing all predictor indices. This subset Θ_j contains the indices of predictor variables that are used for the construction of the j-th tree, such that only those predictor variables with indices in Θ_j are used in the construction of the tree. Note that the amount of predictor variables that are randomly selected may differ from tree to tree.

When the random vector Θ_j is drawn, a subset of the data is created using only the predictor variables in Θ_j. This subset is used to construct a tree. However, instead of splitting the subset into training and validation we choose to do K-fold cross-validation (CV). This CV is used to check whether the results are uniform over all the K different splits, as we obviously want to create trees that are stable through different splits. It might happen that one of the K splits

produces supreme results while the other do not perform as well. By using the K splits, we avoid that such a superior "accidental" split overpowers the results.

As the population mean vectors and covariance matrix are unknown, the training data in each K-th fold is used to estimate the sample means and covariance matrix. The validation data of the K-th fold is then used to validate the LDA using the mean vector and covariance matrix from the training data. The performance of the training and validation data can be measured for all folds and is then brought together to give one performance metric for the tree.

5.4.2 Construction of the LDA tree

To construct the LDA tree, T_j, we first select a random subset Θ_j from the set of predictor indices Θ as in Section 5.4.1 and create a subset of the data using only the predictors of which the indices are in the random vector Θ_j. We then use K-fold CV to create K splits of training and validation data. The number of training and validation data is dependent of the value chosen for K. The number of splits should be chosen with respect to the sample size. Using the sample mean vectors for each possible level and the sample covariance matrix of the training data, we can predict the outcome for both training and validation data with LDA. The accuracy of each tree can be calculated for both training and validation, where the normalized confusion matrix is used such that we can correct for a possible imbalance of the output variable. As mentioned above, the tree is constructed using a K-fold cross validation, implying that there are K accuracies calculated for both training and validation. In order to get a measure of the goodness-of-fit (GoF) for training and validation, we use the mean of the K accuracies,

$$\begin{cases} \overline{ACC}_{TR,j} & := \frac{1}{K}\sum_{k=1}^{K} ACC_{TR,j,k} \\ \overline{ACC}_{VAL,j} & := \frac{1}{K}\sum_{k=1}^{K} ACC_{VAL,j,k} \end{cases} \tag{5.8}$$

with $ACC_{TR,j,k}$ ($ACC_{VAL,j,k}$) the accuracy of the j-th tree and the k-th fold of the training (validation) data. We want to express the GoF of the j-th tree by one value which we can compare with other trees to see which one is better. Obviously, we want to maximize the training accuracy to get as many corretly classified patients as possible. However, what many researchers do not include in their published work is a test of the generalizability of their model. This is essential to prove that the statistical model may further be implemented in the clinical workflow. In other words, we want the most optimal trade-off between the accuracies on the training and validation samples. To do so, we define the GoF measure as the following weighted average accuracies:

$$GoF_i = \frac{1}{4}\overline{ACC}_{TR,i} + \frac{3}{4}\overline{ACC}_{VAL,i}. \tag{5.9}$$

Algorithm 1 (Construction of a LDA tree)**.** *For the j-th tree, T_j, as described above, we use only the predictor variables from the set Θ_j. We denote $\tilde{\boldsymbol{x}}$ the new design matrix consisting of the data of the n patients and predictor variables in Θ_j. Split the data in K folds, $F_1, \ldots, F_K$, where each fold has a n/K sample size and $\cap_{j=1}^{K} F_j = \emptyset$. Make sure that all output levels are equally represented in each fold. Then, for all $k \in \{1, \ldots, K\}$, go through the following steps:*

1. *Set the validation set equal to fold F_k and denote it by $\mathcal{V}_k := F_k$. Define the training data as the remaining data, $\mathcal{T}_k := \mathcal{L} \backslash F_k = \bigcup_{j=1, j\neq k}^{K} F_j$.*

2. *Calculate the sample mean vectors $\hat{\mu}_{m,\mathcal{T}_k}$ and the covariance matrix $\hat{\Sigma}_{\mathcal{T}_k}$ of the training data.*

3. *Use the LDA and Eq. (5.6) to extract the training and validation predictions. Eventually, compute the accuracies $ACC_{TR,j,k}$ and $ACC_{VAL,j,k}$, which may be obtained from the normalized confusion matrix.*

Lastly, the formulas in Eqs. (5.8) and (5.9) can be calculated. Any tree j can therefore be constructed as the classifier $T_j := T(\Theta_j, F_1, \ldots, F_K)$, where $T(\cdot)$ represents the LDA tree building algorithm explained here above.

5.4.3 Bundling and pruning trees in the forest

The procedure to bundle and prune the trees in order to create the forest is thoroughly explained in the following algorithm.

Algorithm 2 (Evolutionary Random Forest)**.** *Start with an empty forest $\mathcal{F}^{(0)}$ and a zero performance metric $PM^{(0)} = 0$ and set iter $= 1$.*

1. *Create n_T trees, $T_1, \ldots, T_{n_T}$, as described in Algorithm 1, and bring the previous forest $\mathcal{F}^{(iter-1)}$ and those newly created n_T trees together as one new Random Forest $\mathcal{F}^{(iter)}$.*
2. *Let $|\mathcal{F}^{(iter)}|$ denote the number of trees in the forest $\mathcal{F}^{(iter)}$ and calculate the performance metric of the forest $\mathcal{F}^{(iter)}$*

$$PM^{(iter)} := \max\left(PM^{(iter-1)}, \text{med}\{GoF_i, i \in \{1, \ldots, |\mathcal{F}^{(iter)}|\}\}\right). \tag{5.10}$$

   All trees satisfying at least one of the following pruning criteria can be removed from the Random Forest $\mathcal{F}^{(iter)}$:

   (a) *the tree, denoted by T_j with $j \in \{1, \ldots, \mathcal{F}^{(iter)}\}$, performs less than the new performance metric, i.e. $GoF_j <= PM^{(iter)}$;*

   (b) *the GoF value of tree T_k is smaller than 0.5;*

   (c) *the accuracy of the validation set is larger than that of the training set.*

3. *Return to step 1 and set iter $=$ iter $+1$ unless one of the following stopping criteria is met:*

   (a) *the number of remaining trees after pruning is equal to n_T;*

   (b) *the number of maximal rounds n_R to build a forest is reached, with $n_R \in \mathbb{N}$ a predetermined number.*

The resulting RF and number of trees in that forest are respectively denoted by $\mathcal{F}$ and n_F.

A schematic representation of the construction of our Evolutionary Random Forest without pruning is given in Figure 5.4.

5.4.4 Evaluation of the Evolutionary Random Forest

Evaluating trees is simple and intuitive in contrast to evaluating a RF. When evaluating a single tree, only one result is obtained, making it straightforward to interpret. However, with a Random Forest (RF), results from multiple trees must be interpreted collectively, and there is no intuitive method for doing so. We use an aggregating method that differs from Breiman's popular voting technique, described in [95]. The evaluation however resembles the aggregating method for regression that Breiman described in the same article and it also resembles Ho's method of discriminant functions.

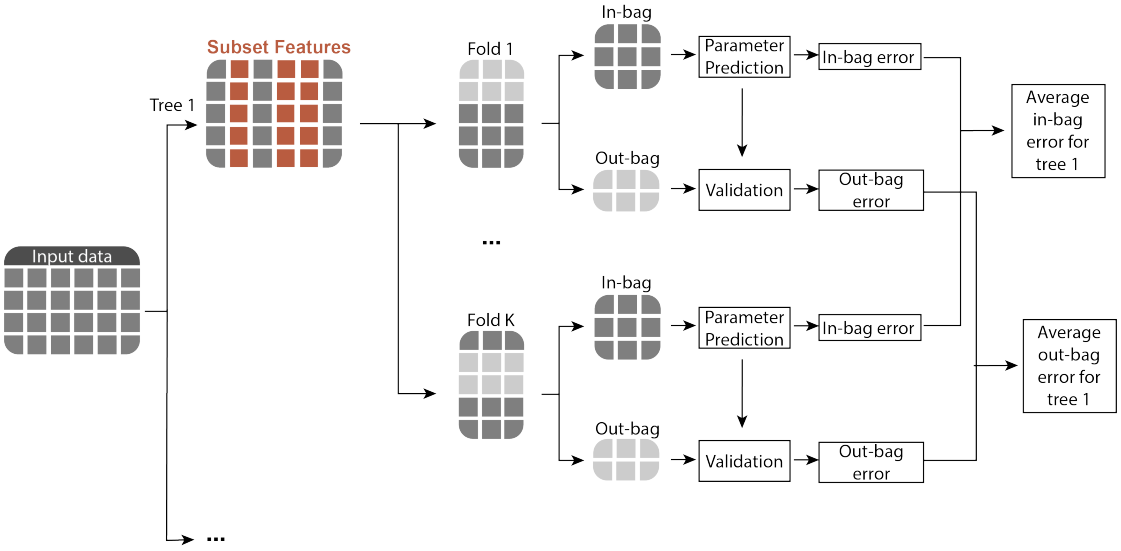


Figure 5.4: A schematic representation of the construction of our Evolutionary Random Forest without the pruning process. In a next step, the trees that perform less than the performance metric are pruned and the process is started again.

In our RF, the response of a sample $(\underline{x}_i, y_i)$ is predicted using Eq. (5.6). For each tree T_j, the following discrimination score is computed,

$$B_{i,j,m} = \frac{\left(\underline{x}_i - \frac{1}{2}\hat{\underline{\mu}}_{j,m}\right)\hat{\Sigma}_j\hat{\underline{\mu}}'_{j,m}}{\sum_{k=1}^{M}\left(\underline{x}_i - \frac{1}{2}\hat{\underline{\mu}}_{j,k}\right)\hat{\Sigma}_j\hat{\underline{\mu}}'_{j,k}}, \tag{5.11}$$

where $\hat{\underline{\mu}}_{j,m}$ is the mean vector calculated for tree T_j and the m-th class. We need the normalization in equation (5.11) in order to compare different trees. This is due to the fact that different trees are built with different predictor variables and hence results in scores of different order of magnitude.

Instead of predicting the class for each patient and each tree such that we can perform popular voting, we now use the discriminant scores $B_{i,j,m}$ and calculate for each patient and each possible class the mean over all trees,

$$\overline{B}_{i,m} = \frac{1}{n_F}\sum_{j=1}^{n_F} B_{i,j,m}. \tag{5.12}$$

With Eq. (5.12), we can now select the class for which the average is the highest, i.e.

$$\hat{y}_i = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}}\ \overline{B}_{i,m}. \tag{5.13}$$

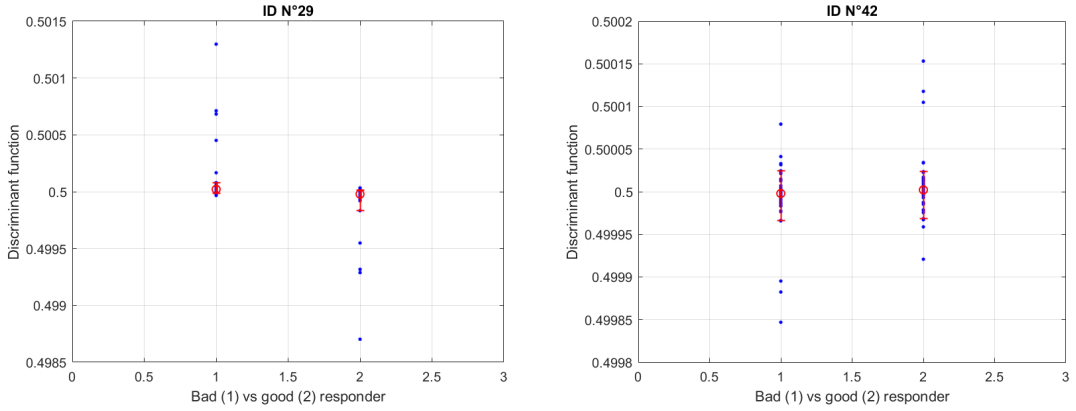
Using this technique, every single patient can get a prediction out of our Evolutionary Random Forest.

5.4.5 The Evolutionary Random Forest

We choose to perform our RF with the LDA instead of any other classifier. Since the logistic regression is a popular classifier, we created a similar algorithm where the LDA was replaced by

the logistic analysis while keeping the same RF mechanism. The total amount of trees created per round was $n_T = 200$ and the total number of rounds was $n_R = 1000$. After pruning the last forest, there were in total 196 trees left, with a normalized accuracy of 58.532%. The LOOCV was again poor with a normalized accuracy of only 22.225%. The results of our RF with the logistic regression does not perform better than the state-of-the-art classifiers or even the ensemble methods. Hence, we continued our research by looking at the LDA with the non-penalized RF. With again $n_R = 1000$ rounds where each time $n_T = 200$ trees were created per round, we achieved an evaluation accuracy of 77.507% and a LOOCV accuracy of 67.081%. This accuracy is the best one we have obtained whilst keeping a decent evaluation accuracy.

We finally choose in our algorithm (see Eq. (5.12)) to take the mean of our discriminant scores rather than the median. Taking the average of the discriminant scores brings the final accuracy of training and validation closer to one another as opposed to choosing the median of the scores: these were respectively 91.846% and 60.262%. Not only was the validation accuracy lower, but the difference between training and validation accuracies was larger than 30%. Remember that our goal is to create an optimal RF algorithm, in the sense that it should properly fit the training data, but validation should not suffer from this.



(a) patient with ID N°29

(b) patient with ID N°42

Figure 5.5: Scatter plot (blue) of the discriminant functions per response type and a error plot (red) displaying the 95% empirical confidence interval with mean. A discriminant of 0.5 means that the algorithm cannot differentiate between the two level, larger and lower values respectively mean that a bad and good response are predicted.

5.4.6 Examples

Even though we are only interested in the prediction given by the Evolutionary RF for each patient, it is interesting to look at the discriminant scores $B_{i,j,m}$ (see Eq. (5.11)) for all j trees and m levels of the i-th patient. As we will illustrate, these discriminant scores give some useful insights into the data. Although the forest makes a decision based on the mean value over the different scores, we analyze and discuss the variation of these discriminant scores over the forest.

As an illustration, we will look at the discriminant scores generated by the Evolutionary RF for two different patients. To analyze these scores, Figure 5.5 shows them for each tree as a function of the predicted outcome associated with that specific tree. Additionally, we add the 95% confidence intervals of the obtained discriminant scores for each outcome category together with

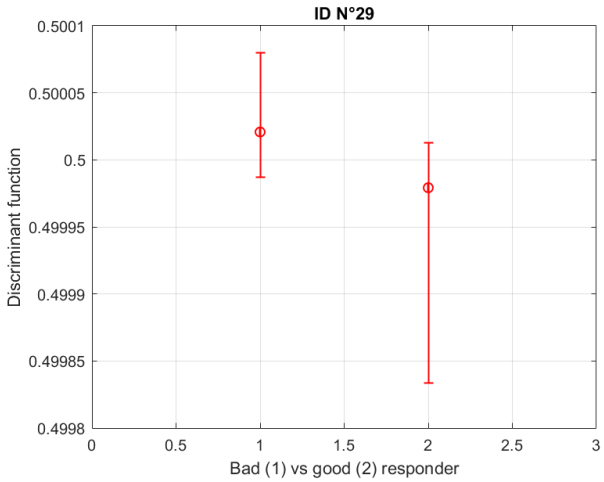


Figure 5.6: The error plot displaying the 95% empirical confidence interval with mean for patient number 29.

its group mean. Note that the discriminant scores are probabilities representing the likelihood according to a specific tree in the forest for a certain outcome. The more "certain" the forest is about its decision, the higher the mean value of that group and the smaller the confidence interval should be. Thus, we can assess the "certainty" of the ensemble for a specific outcome by assessing the width of the confidence intervals.

First, we investigate in more depth the scores generated by the forest of patient N°29. To have a better view of the confidence intervals, we included Figure 5.6 where solely the confidence intervals for both response groups are given. In this graph it is clear that the mean is significantly higher for the first response group and the discriminant scores are also concentrated in a more narrow bandwidth compared to the second group. Therefore, we can conclude that our Evolutionary RF predicted the patient as a bad responder. This is in line with the response given by the pathologist, who also classified the patient as a bad responder.

Secondly, we take a closer look at patient N°42 whose discriminant scores can be found in Figure 5.5b. At first glance, we see no significant difference between the two response groups. Although the average probability of the ensemble is in favor of classifying this patient as a "good responder", the uncertainty for both outcomes is high, which is an indication that our RF algorithm is not able to confidently make a prediction for the patient. Since the mean of the discriminant scores of the good responders is slightly higher, the algorithm predicted that the patient must be a good responder. If we look at the response value for this patient given by the pathologist, we find that he/she was classified as a bad responder. We conclude that this type of visual assessment of the ensemble's uncertainty is of importance for clinicians to understand the "certainty" of any decision reached by the machine learning method. In the presence of large uncertainties, a clinician should carefully review the prediction.

5.5 Discussion

Our data is a challenging data set for which it is not straightforward to make a good prediction. We are dealing with a regression grade that is not quantitatively determined: the output has some level of systematic error since the pathologist cannot be objective when grading the Dworak

regression grade of a patient. Therefore, there is a variability in the data set due to different pathologists that have graded different specimens. As the algorithm creates its own opinion based on the data given, it is crucial to have a stable output variable ("garbage in, garbage out"). As this was not the case in our data, we could not expect great results for classification.

On top of that, it is shown that some CT scanners do not produce stable radiomics when the scan is retaken. As we have not tested this in our study, we cannot confirm whether the radiomics are reproducible. Besides, although literature shows the great potential of the radiomics in the field of oncology, it is still at an early stage. The radiomic features that can be extracted from medical images go beyond what the eye can see. It is important to find a connection between the mathematically computed features and the findings of the physician before handling our RF to them as a prediction tool.

These problems show that the classical prediction models are not always suited for real (medical) data as there are too many assumptions in the statistical/mathematical theory that can be violated. Instead the researchers should find out what the characteristics of the data are and construct a classification algorithm that works well with the problems/challenges the data poses.

(Non-)ensemble	Classification Method	Evaluation	LOOCV
Non-Ensemble	SVM - linear kernel	77.074%	42.596%
	SVM - radial kernel	55.166%	48.969%
	Logit	100.000%	56.654%
	LDA	93.908%	52.700%
	Classification tree	92.573%	46.837%
Ensemble	RF Leo Breiman	46.708%	45.677%
	RF Tin Kam Ho	100.000%	46.720%
	Evolutionary RF	77.507%	67.081%

Table 5.3: Performance of both non-ensemble and ensemble methods. The accuracy was calculated from the normalized confusion matrix.

As state-of-the-art classifiers and machine learning algorithms (MLA's) are sometimes considered to be promising tools that solve all your problems, we thought it to be important to show that the performance of the classical classifiers and MLA's depend on the considered data set. Moreover, many classification methods require the researcher to set some parameters, like the number of trees created in Breiman's forest. Most built-in functions for such classification methods provide the researcher with standard values for these parameters, which is not always the best option if you want to fine-tune the model according to your data.

We have shown that the state-of-the-art classifiers do not perform well on our data (see Tab. 5.3). Most of them showed to be useful for the training of the data. However, we obtained the validation of the data through CV which showed that most of the viewed classifiers did not even reach 50%, proving that overfitting was observed.

The ensemble methods should be more robust when it comes to overfitting. Breiman has proved that a forest is less difficult to overfit when the number of trees in the forest increases. However, the results of the standard forest of Breiman were disappointing as both evaluation and CV were below fifty percent accuracy. The forest of Tin Kam Ho did show perfect training results, but seemed to be over-fitted as the CV was below fifty percent.

Logit, LDA and SVM all make use of the inverse of the covariance matrix. Since the radiomics data contains multicollinearity, the covariance matrix will be close to a singular matrix, such that

the inverse cannot be calculated properly. We cannot expect a method to be stable and work well for validation data when the needed parameters are wrongly predicted as the inverse of the covariance matrix is not correctly calculated. This multicollinearity shows that we need to perform some feature selection.

We looked at using a classification tree as prediction tool but ended up with poor results. Classification trees are grown using binary recursive partitioning where each time a split is chosen that maximizes the reduction in impurity. The number of splits are limited and chosen from a limited set of options as it is not feasible to try out all options. We are faced with the curse of high dimensionality in the construction of our classification tree leading to possibly a badly chosen tree that overfits.

Breiman tackled the idea of overfitting by bundling multiple trees together. However, the evaluation of Breiman's forest using our data was extremely poor. Ho's method of subspacing the predictors is a good idea as the subsets might contain variables that aren't as highly correlated. However, some trees might be constructed using a subset that only contains highly correlated variables resulting in poor prediction results. All these trees bundled together result in a bad performance as badly constructed trees are kept in the ensemble. Our RF resolved this problem by removing badly performing trees, making new trees to add to the pruned forest, and repeating the process. In this way we can keep on learning from the data.

Furthermore, we changed the classification tree as introduced by Breiman by the LDA. We opted for the LDA over decision trees as we can use all of our subspaced prediction variables in the LDA. If we would accidentally select a collection of variables with highly correlated variables, then the tree will be pruned later on if it produces a bad prediction result. We also chose the LDA over the logit as the latter is a method where prediction coefficients are estimated by MLE which is done by an iterative process leaving room for rounding errors to pile up. We therefore ended up with a new method of constructing a RF that suits the characteristics of our data.

We showed in this paper that the choice of classification method is crucial for our data. By constructing our Evolutionary RF algorithm, we were able to create a prediction model that performs fairly well on both training and validation. The validation results might not be superb, but if we keep in mind all the challenges we are faced with like the instability of our response variable and the radiomics data that is extracted over a long period of time using many different scanners and many different settings, the results are better than we expected them to be. We initially thought that the Evolutionary RF would still experience problems due to the challenging data characteristics but our results showed that the RF is capable of predicting the data to a decent level in spite of the challenges.

We showed that our ERSF can perform quite well for our challenging data set. However, the ERSF is a black box system meaning that it is not easily interpretable or transparent. Specifically, although the ERSF algorithm can make accurate predictions, the internal decision-making process—i.e., how it uses the data to make the prediction—is not easily understood by humans. We want to break down the black box system with more research using animal trials [99]. This research is conducted within our research group, Biostatistics and Medical Informatics research group (BISI), at the Vrije Universiteit Brussel (VUB). These trials will be used to understand the decision-making process even better and will search for the biological meaning of the features. This biological interpretability will be necessary for the clinical application of the ERSF. A. Rifi *et al.*

5.6 Appendix

5.7 Transformation of the Radiomics data

	Variable (X)	p-value	Transformation of X	new p-value	mean	standard deviation
First Order	Energy	0.000	$\ln(X)$	0.894	18.311	1.946
	Total Energy	0.000	$\ln(X)$	0.569	20.194	1.610
	Entropy	0.355	/	/	2.953	0.602
	Minimum	0.000	$\ln(X - \min(X) + 1)$	0.158	4.712	1.557
	10th percentile	0.000	$\ln(-X - \min(-X) + 1)$	0.064	4.445	1.045
	90th percentile	0.305	/	/	75.212	20.511
	Maximum	0.000	No transformation found	/	225.010	349.420
	Mean	0.000	$\ln(-X - \min(-X) + 1)$	0.243	4.083	0.739
	Median	0.074	/	/	38.433	24.226
	Interquartile Range	0.000	$\ln(X)$	0.122	4.042	0.559
	Range	0.000	No transformation found	/	994.794	476.864
	Mean Absolute Deviation	0.000	$\ln(X)$	0.176	3.974	0.699
	Robust Mean Absolute Deviation	0.000	$\ln(\frac{-X}{1-X})$	0.758	0.046	0.022
	Root Mean Squared	0.000	$\ln(X)$	0.154	4.648	0.590
	Skewness	0.904	/	/	-3.660	2.012
	Kurtosis	0.000	$\ln(X)$	0.234	3.020	0.957
	Variance	0.000	$\ln(X)$	0.556	8.985	1.415
	Uniformity	0.570	/	/	0.192	0.065
Shape 3D	Mesh Volume	0.000	$\ln(X)$	0.745	10.890	0.790
	Voxel Volume	0.000	$\ln(X)$	0.7412	10.897	0.787
	Surface Area	0.040	$\ln(X)$	0.843	9.287	0.560
	Surface Area to Volume ratio	0.013	$\ln(X)$	0.719	-1.603	0.304
	Sphericity	0.419	/	/	0.644	0.085
	Maximum 3D diameter	0.650	/	/	78.007	21.858
	Maximum 2D diameter (slice)	0.230	/	/	59.661	16.397
	Maximum 2D diameter (column)	0.674	/	/	63.873	18.068
	Maximum 2D diameter (row)	0.410	/	/	72.967	22.147
	Major axis length	0.375	/	/	67.129	20.534
	Minor Axis length	0.096	/	/	43.432	11.857
	Least Axis length	0.334	/	/	35.578	10.866

	Elongation	0.416	/	/	0.669	0.137
	Flatness	0.892	/	/	0.544	0.120
GLCM	Autocorrelation	0.000	$\ln(-X - \min(-X) + 1)$	0.081	6.306	0.971
	Joint Average	0.000	$\ln(-X - \min(-X) + 1)$	0.193	2.233	0.802
	Cluster Prominence	0.000	$\ln(X)$	0.205	10.232	2.991
	Cluster Shade	0.000	No transformation found	/	-4580.506	7715.485
	Cluster Tendency	0.000	$\ln(X)$	0.400	3.651	1.526
	Contrast	0.000	$\ln(X)$	0.298	2.190	1.145
	Correlation	0.674	/	/	0.590	0.165
	Difference Average	0.006	$\ln(X)$	0.398	0.464	0.483
	Difference Entropy	0.297	/	/	2.279	0.502
	Difference Variance	0.000	$\ln(X)$	0.239	1.807	1.257
	Joint Energy	0.026	$\ln(X)$	0.926	-2.969	0.598
	Joint Entropy	0.629	/	/	5.325	1.067
	Informational Measure of Correlation 1	0.263	/	/	-0.142	0.050
	Informational Measure of Correlation 2	0.944	/	/	0.703	0.122
	Inverse Difference Moment	0.677	/	/	0.545	0.080
	Maximal Correlation Coefficient	0.813	/	/	0.673	0.128
	Inverse Difference Moment Normalizes	0.018	$\ln(\frac{X}{1-X})$	0.237	5.126	0.862
	Inverse Difference	0.574	/	/	0.583	0.068
	Inverse Difference Normalized	0.057	/	/	0.959	0.017
	Inverse Variance	0.289	/	/	0.437	0.045
	Maximum Probability	0.065	/	/	0.139	0.067
	Sum Average	0.000	$\ln(-X - \min(-X) + 1)$	0.311	2.848	0.866
	Sum Entropy	0.520	/	/	3.723	0.702
	Sum Squares	0.000	$\ln(X)$	0.0546	2.497	1.453
GLSZM	Small Area Emphasis	0.276	/	/	0.657	0.057
	Large Area Emphasis	0.000	$\ln(X)$	0.869	3.588	1.598
	Gray Level Non-Uniformity	0.000	$\ln(X)$	0.893	4.012	0.888
	Gray Level Non-Uniformity Normalized	0.106	/	/	0.087	0.044
	Size-Zone Non-Uniformity	0.000	$\ln(X)$	0.730	5.666	1.238
	Size-Zone Non-Uniformity Normalized	0.148	/	/	0.404	0.069

	Zone Percentage	0.034	$\ln(X)$	0.529	-2.428	0.539
	Gray Level Variance	0.022	$\sqrt{(X)}$	0.158	7.557	3.453
	Zone Variance	0.000	$\ln(X)$	0.829	9.491	1.617
	Zone Entropy	0.128	/	/	5.955	0.713
	Low Gray Level Zone Emphasis	0.000	$\ln(\ln(X)) - \min(\ln(X)) + 1)$	0.563	0.978	0.315
	High Gray Level Aone Emphasis	0.001	$\ln(-\sqrt{(X)} - \min(-\sqrt{X}) + 1))$	0.052	2.539	0.616
	Small Area Low Gray Level Emphasis	0.000	No transformation found	/	0.008	0.011
	Small Area High Gray Level Emphasis	0.000	No transformation found	/	576.38	290.66
	Large Area Low Gray Level Emphasis	0.000	$\ln(X)$	0.999	2.881	1.897
	Large Area High Gray Level Emphasis	0.000	$\ln(X)$	0.841	16.341	1.917
GLRLM	Short Run Emphasis	0.912	/	/	0.805	0.048
	Long Run Emphasis	0.217	/	/	2.614	0.677
	Gray Level Non-Uniformity	0.000	$\ln(X)$	0.721	6.845	1.098
	Gray Level Non-Uniformity Normalized	0.763	/	/	0.165	0.053
	Run Length Non-Uniformity	0.000	$\ln(X)$	0.619	8.206	1.196
	Run Length Non-Uniformity Normalized	0.984	/	/	0.610	0.076
	Run Percentage	0.814	/	/	0.735	0.064
	Gray Level Variance	0.000	$\ln(X)$	0.384	2.741	1.355
	Run Variance	0.169	/	/	0.684	0.323
	Run Entropy	0.221	/	/	4.183	0.470
	Low Gray Level Run Emphasis	0.000	No transformation found	/	0.006	0.008
	High Gray Level Run Emphasis	0.000	$\ln(-\sqrt{X} - \min(-\sqrt{X}) + 1)$	0.138	2.208	0.813
	Short Run Low Gray Level Emphasis	0.000	$\ln(X)$	0.058	-5.907	1.038
	Short Run High Gray Level Emphasis	0.000	No transformation found	/	918.47	448.93
	Long Run Low Gray Level Emphasis	0.000	No transformation found	/	0.014	0.021
	Long Run High Gray Level Emphasis	0.419	/	/	3044.2	1676.7
NGTDM	Coarseness	0.000	$\ln(X)$	0.754	-6.990	1.097

	Contrast	0.000	$\ln(X)$	0.342	-3.612	1.254
	Busyness	0.000	$\ln(X)$	0.650	-0.367	1.087
	Complexity	0.025	$\sqrt{X}$	0.293	32	15.744
	Strength	0.000	$\ln(X)$	0.444	0.546	1.274
GLDM	Small Dependence Emphasis	0.064	/	/	0.106	0.050
	Large Dependence Emphasis	0.379	/	/	86.426	33.490
	Gray Level Non-Uniformity	0.000	$\ln(X)$	0.729	7.300	1.115
	Dependence Non-Uniformity	0.000	$\ln(X)$	0.237	1.824	0.200
	Dependence Non-Uniformity Normalized	0.081	/	/	0.069	0.018
	Gray Level Variance	0.000	$\ln(X)$	0.518	2.561	1.401
	Dependence variance	0.228	/	/	21.284	8.380
	Dependence Entropy	0.321	/	/	6.485	0.481
	Low Gray Level Emphasis	0.000	No transformation found	/	0.005	0.008
	High Gray Level Emphasis	0.000	$\ln(-X - \min(-X) + 1)$	0.059	6.294	0.984
	Small Dependence Low Gray Level Emphasis	0.000	$\ln(X)$	/	-7.475	0.997
	Small Dependence High Gray Level Emphasis	0.320	/	/	106.250	71.746
	Large Dependence Low Gray Level Emphasis	0.000	$\ln(-\ln(\frac{1}{X}) - \min(-\ln(\frac{1}{X})) + 1)$	0.114	1.338	0.316
	Large Dependence High Gray Level Emphasis	0.681	/	/	102330.937	63824.152

Table 5.4: List of transformed variables in the data set. The p-values given in the table are those belonging to the Kolmogorov–Smirnov test, where the null-hypothesis states that the data follows a normal distribution.

5.7.1 Confusion matrices for the state-of-the-art classifiers

		Predicted	
Obs.		90	7
		17	27

(a) SVM (linear) - Evaluation

		Predicted	
Obs.		65	32
		36	8

(b) SVM (linear) - LOOCV

		Predicted	
Obs.		96	1
		39	5

(c) SVM (radial) - Evaluation

		Predicted	
Obs.		95	2
		44	0

(d) SVM (radial) - LOOCV

		Predicted	
Obs.		97	0
		0	44

(e) logit - Evaluation

		Predicted	
Obs.		57	40
		20	24

(f) logit - LOOCV

		Predicted	
Obs.		94	3
		4	40

(g) LDA - Evaluation

		Predicted	
Obs.		56	41
		23	21

(h) LDA - LOOCV

		Predicted	
Obs.		87	10
		2	42

(i) tree - Evaluation

		Predicted	
Obs.		60	37
		30	14

(j) tree - LOOCV

		Predicted	
Obs.		54	43
		28	16

(k) hierarchical clustering - LOOCV

		Predicted	
Obs.		85	12
		37	7

(l) knn ($K = 11$) - LOOCV

Table 5.5: Non-normalized confusion matrices for the state-of-the-art classifiers.

5.7.2 Confusion matrices for the ensemble methods

		Predicted	
Obs.		82	15
		40	4

(a) Breiman - Evaluation

		Predicted	
Obs.		82	15
		40	4

(b) Breiman - LOOCV

		Predicted	
Obs.		97	0
		0	44

(c) Tin Kam Ho - Evaluation

		Predicted	
Obs.		73	24
		36	8

(d) Tin Kam Ho - LOOCV

		Predicted	
Obs.		71	26
		8	36

(e) Evolutionary Random Subspace Forest - Evaluation

		Predicted	
Obs.		64	33
		14	30

(f) Evolutionary Random Subspace Forest - LOOCV

Table 5.6: Non-normalized confusion matrices for the ensemble methods.

Chapter 6

Bridging the Gap Between Machine Learning and Medicine

This chapter is based on the accepted research paper:
C. Raets, C. El Aisati, A. Rifi, M. De Ridder, K. Putman, J. De Mey, A. Sermeus, and K. Barbé, C., "Bridging the gap between machine learning and medicine: A critical evaluation of the dworak regression grade in rectal cancer," **IEEE Open Journal of Instrumentation and Measurement**, vol. 3, pp. 1–12, 2024

Abstract

The growing popularity of AI has increased its widespread adoption in medicine. However, the relationship between AI and medical experts' opinions remains elusive. This study investigated the consistency between Random Forest's prediction for rectal cancer regression grades and doctors' opinion based on clinical data. We examined the impact of grading system subjectivity on the algorithm.

Analyzing clinical parameters and medical notes from 85 rectal cancer patients, we identified patients with ambivalent grades, the "gray-zone patients", and explored the algorithm's difficulty in predicting their regression grade. We also introduced a regularization parameter to test if some patients could still correctly be predicted when some statistical information is suppressed.

Our results demonstrated that the gray-zone patients were significantly more difficult to classify using the algorithm, suggesting that such patients should be reviewed twice to reduce errors. Additionally, we observed that the regularization parameter did not benefit gray-zone patients as much as others.

Our findings emphasize the need for AI and clinical experts to work collaboratively since the algorithm cannot consider the subjectivity that medical experts can identify. Further research is necessary to incorporate subjectivity into AI algorithms to enhance their predictive capabilities and further bridge the gap between medicine and AI.

6.1 Introduction

Artificial intelligence (AI) refers to a set of computer algorithms that perform tasks that usually require human intelligence [100, 101]. In the medical field, AI is used for various purposes such as diagnosing patients, providing prognoses, and developing drugs, among others. Some companies like PathAI use machine learning (ML) technologies to assist pathologists in making more precise cancer diagnoses and developing customized treatments for patients [102]. Others focus on diagnosing patients using AI technology to alleviate administrative pressures or to diagnose and treat patients by examining their symptoms and suggesting a cure [103, 104]. In recent years, many more AI companies have been established to create AI solutions for medicine.

However, caution must be exercised when using AI algorithms specifically designed for medical applications. Not all clinicians trust and accept the use of AI in real medical settings, as the algorithms are often too complex and lack transparency (black-box models), making them difficult to interpret or question when necessary [6, 7]. Furthermore, if clinicians do not understand the algorithm and its choices, communicating and explaining those choices to the patients becomes challenging, and patients may hesitate to trust those medical decisions. This issue is highlighted by Visar Berisha and Julie Liss, two co-founders of Aural Analytics, who have referred to AI in medicine as overhyped [8]. They emphasize that mistakes made by AI algorithms in medicine can lead to life-or-death situations and that it is essential to comprehend how AI renders decisions before using it in hospitals and other medical settings: this is called AI explainability.

Instead of completely replacing the clinical expert with a set of AI algorithms, we propose using them as a second-opinion tool to assist the clinician. Additionally, we aim to open up black-box models to establish a connection between biomedicine and the decisions made.

Even though doubts about AI in medical settings have been raised, they can still provide significant value. It is essential to thoroughly test an AI model to determine whether it reaches the same conclusions as doctors and whether doctors can accept the AI's findings. For example, Bum-Sup Jang *et al.* developed a deep learning (DL) model for predicting pathological responses (complete response and good response) in rectal cancer patients using post-chemotherapy MRI images. They not only created a DL model but also assessed whether doctors agreed with the pathological responses generated by the AI model [9]. This type of research is crucial for stablishing a trustworthy model that doctors can have confidence in.

In the previous chapter, we introduced our prediction model, ERSF, for the TRG of patients with rectal cancer using quantitative data obtained from CT images taken before the start of their (chemo)radiotherapy. The radiomics used for the ML algorithm do not have any known biological interpretation. However, they can allow characterizing the tumor phenotype, which is not visible with the naked eye [99]. However, it is by establishing the connection between the radiomics utilized in our ERSF and the clinical data employed by the clinicians that we can determine where they can efficiently work together and where they can learn from each other. Further research aiming at unraveling the biological meaning of radiomic features is conjointly being conducted at the VUB (Free University of Brussels) by the research groups BISI and TROP [99].

The Dworak TRG is highly influenced by the pathologist's opinion, resulting in a RF algorithm that must predict a subjective and unstable variable. In this chapter, our objective is to examine the Dworak TRG more closely and establish a connection between the decisions made by the RF and those of the medical world, thereby creating the first cracks to open our black-box model. We expect that the ERSF will be severely impacted by Dworak's instability, and patients with an uncertain Dworak TRG are much more difficult to classify than others. Luckily, clinicians possess a unique ability to think outside the box, analyzing subtle symptoms, patient

backgrounds, and other contextual factors often missed by data. Therefore, if the algorithm is severely impacted by the difficulty in grading TRGs, we could offer these patients to the clinicians for an additional opinion. The ERSF is merely used as an additional opinion tool to assist the doctors.

In a preliminary step, we will examine the instability of the Dworak TRG and determine its potential impact on the RF's predictions. To accomplish this, we plan to reassign the Dworak TRG for each patient using significant clinical parameters and pathological notes obtained during the biopsy. Through this process, we aim to identify patients with an unstable regression grade, herein referred to as gray-zone patients. Once identified, we will provide their CT features to the algorithm to determine if they are systematically misclassified by the algorithm. If so, we can conclude that both the clinician and the algorithm have difficulty determining the regression grade of these patients.

In a subsequent step, we intend to introduce a regularization parameter on the covariance matrix used in the classification method to assess the algorithm's capability to correctly classify patients who were previously misclassified. Patients who can be correctly classified with some regularization can be seen as atypical patients who do not fit the general pattern observed during the AI's training phase. We want to investigate whether these unique patients are the ones presenting difficulty in grading.

In many ML studies, the stability of the data is not thoroughly examined, nor are any comparisons made between unstable data and the ML algorithm investigated and discussed. Our study presents a multifaceted approach that combines ML with clinical insights to address classification challenges. By elucidating the complexities of the Dworak TRG and its impact on algorithmic predictions, our research contributes to advancing the field of medical imaging analysis and personalized patient care.

6.2 Dworak Regression Grade

Several grading systems have been proposed to score rectal tumors, but there is currently no universally recognized grading system. None of them are purely quantitatively graded, leaving us with grading systems that are always partially subjective [84, 83]. The same holds for the Dworak TRG, where subjectivity is present across all different grades. The Dworak TRG can be viewed as a categorical variable with an underlying numerical scale, where the thresholds of the different groups overlap. Even when regrouping the Dworak TRG into bad (TRG 0-2) and good (TRG 3-4) responders, there will still be some subjectivity remaining, which may pose difficulties for any prediction algorithm. Figure 6.1 illustrates the overlap region, or gray-zone, where multiple categories intersect, for both the Dworak TRG and the regrouped bad and good responders. These overlapping regions as small as they might be, can be extremely damaging to the ML algorithm. By regrouping the Dworak TRG the overlap is minimized but will always still be present between the bad and good responders and will pose challenges when constructing the ERSF or any other prediction algorithm.

6.3 Gray-zone patients

It is apparent that any ML algorithm will be impacted by the subjectivity of the Dworak TRG, even when the regrouped TRG version is used. In this chapter, we will define the patients who fall within the gray-zone area, referring to them as gray-zone patients. Since we reduced variability by regrouping patients into only bad and good responders, we need only examine those with a Dworak TRG of 2 or 3, as they may fall within the gray-zone area. To do so, we performed

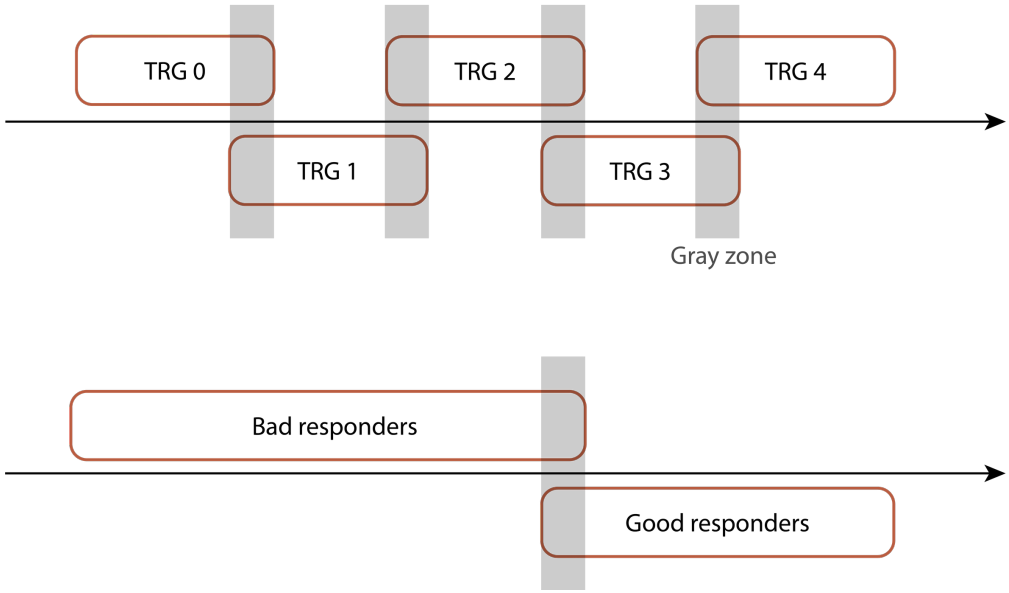


Figure 6.1: Overlapping region (gray-zone) demonstrating the subjectivity problem between the different Dworak TRGs and the regrouped bad and good responders.

a semi-qualitative post-hoc analysis on clinical data that we believed had predictive relevance. The variables extracted were: gender, age at the start of the therapy, histological type, tumor grade, distance of the tumor from the anal verge, neoadjuvant therapy, surgical procedure, post-operative follow-up treatment, recurrence of the tumor, TNM score, pTNM score, proximal-, distal- and circumferential resection margin, lymphatic-, perineural-, and venous invasion. The gray-zone patients were identified using a tree structure based on the significant variables from the post-hoc test, supervised by an expert with pathological and histological knowledge. Furthermore, all patients with surgical notes indicating doubt or containing contradictory notes were classified as gray-zone patients

Due to low sample sizes, the significance of categorical variables was tested using Fisher's exact, while an independent t-test was used to test for significance of any continuous variables. As our ERSF is trained using the regrouped Dworak TRG, we are only interested in testing the significance between its two levels. All tests used a significance level of $\alpha = 0.1$.

6.4 Regularization of the ERSF

Regularization or penalization techniques are frequently used in regression and classification problems. The most popular technique among them are lasso and the ridge regression, both of which penalize the coefficients of the regression or classification problem. The idea of Ridge regression can also be applied to LDA. The penalization in the LDA will be on the estimated covariance matrix denoted by $\hat{\Sigma}$. Regularizing the covariance matrix has an advantage in penalizing the matrix, potentially voiding its singularity, and accounting for the multicollinearity of the data set. Moreover, it minimizes the impact of the covariance matrix on the LDA.

Numerous studies have been conducted to find an effective regularization method for LDA. Mkhadri proposed two forms of penalization [105]. The first type was introduced by Peck and Van Ness, who presented a new formula for the inverse of the covariance matrix, $\hat{\Sigma}^{-1}(\lambda) =$

$(1-\lambda)a\hat{\Sigma}^{-1} + \left(b\lambda/tr(\hat{\Sigma})\right) I_p$ where a and b are positive constants, $tr(\cdot)$ represents the the trace function, and $\lambda \in [0,1]$ the regularization parameter. Mkhadri's second proposition was a regularization method from Di Pillo and Campbell that employed a regularized covariance matrix given by $\hat{\Sigma}^{-1}(\lambda) = \hat{\Sigma}^{-1} + \lambda I_p$. Friedman proposed a method called regularized discriminant analysis (RDA) in his paper [93]. However, this method uses the information of the covariance matrices per prediction group (i.e. bad or good responder).

6.5 Medical data

Under the ethical committee approval EC1010135, we initially had a total of 172 patient files available. However, upon close inspection, we discovered that two patients were added twice, resulting in a total of 170 unique patient files. Since we are working with a retrospective data set, we do not have all the desired information for each patient. Our patients were initially admitted to the UZB hospital in Brussels, but not all patients completed their treatment there (see Table 6.1). Among the 170 patients, we identified 141 patients who had both a CT scan and a Dworak TRG available, qualifying them for the ERSF study. However, our study now aims to examine the influence of Dworak TRG instability on the ERSF prediction. To identify factors influencing the Dworak TRG, we need to examine the surgical notes and histological parameters. In our study, we have 106 patients with both surgical notes and Dworak TRG available, suitable for post-hoc analysis. However, not all 106 patients with surgical notes also have a CT scan available. We only have 85 patients overlapping with Dworak TRG, CT scan, and surgical notes. Therefore, these 85 patients are the only ones we can use to assess the influence of the Dworak subjectivity on the ERSF

Hospital	**Full data**	**Gray-zone data set**	**CT data set**
Hospital 1	41	9	36
Hospital 2	4	2	4
Hospital 3	12	5	7
Others	2	1	2
UZB	111	89	92
Total Patients	170	106	141

Table 6.1: Different hospitals in the full data set of $n = 170$ patients, the subgroup of $n = 106$ patients for the gray-zone study, and the $n = 141$ patients with both CT scan and Dworak TRG for the ERSF construction.

The youngest patient in our study was 32 years old while the oldest was 85 years old with a sample average age of 65 years old. The data set consisted of more male patients (111) than female patients (59). The distribution of the Dworak TRG was not balanced across the data set, with 2 patients classified as grade 0, 54 as grade 1, 58 as grade 2, 39 as grade 3, and 7 as grade 4. Additionally, 10 patients had an unknown Dworak grade and were consequently excluded from the analysis.

Parameter		Full data	Gray-zone data set	CT data set
Gender	Male	111	65	93
	Female	59	41	48
Dworak	Grade 0	2	2	2
	Grade 1	54	37	45
	Grade 2	58	41	50
	Grade 3	39	19	38
	Grade 4	7	7	6
	NA	10	-	-

Table 6.2: Patient characteristics for the full data set of $n = 170$ patients, the subgroup of $n = 106$ patients for the gray-zone study, and the $n = 141$ patients with both CT scan and Dworak TRG for the ERSF construction.

We can now analyze the ERSF constructed using the data from 141 patients, for whom we have both the planning CT images and the Dworak TRG. Subsequently, we will examine the ERSF prediction for the 85 patients, for whom we also possess surgical notes. Our primary focus lies in exploring the intersection between machine learning and medicine. To further investigate this connection, we will enhance our ERSF by introducing a regularization parameter. This parameter modifies the influence of one or more parameters, potentially impacting the prediction process. Through the introduction of this regularization parameter, we aim to ascertain whether patients deemed difficult to classify by pathologists are also more challenging to predict accurately using the algorithm.

6.6 Results

6.6.1 Determination of the gray-zone patients

To identify the gray-zone patients, we need to examine the surgical notes reported by the pathologists. However, we only have access to surgical notes for a limited number of 106 patients. Consequently, we can only use these patients for the gray-zone study.

To begin with, we tested the significance of clinical variables and surgical parameters found in the clinical notes. Table 6.3 lists the p-values for all significant categorical and numerical parameters. Table 6.8 provides more information about non-significant parameters. Due to the low number of observations per category (bad vs good response grades) in the regrouped Dworak TRG, we used the non-parametric Fisher's exact test instead of the commonly used chi-square test. The significance of the numerical variables was tested using a t-test. Using a significance level $\alpha = 0.1$, we identified ten significant variables for the regrouped grading system.

Variable name		**Number of bad - good responders**	**p-value**
Sex	Male female	$54-11$ $26-16$	0.0357
Post-operative therapy	None Chemotherapy Aggressive Treatment	$22-16$ $45-9$ $10-1$	0.0104
Surgical resection margin	Positive Negative No Tumor	$9-0$ $68-20$ $0-6$	$7.562e-05$
Tumor grade	Low Mediocre High Indeterminable	$50-3$ $16-3$ $11-5$ $1-4$	0.0283
Pathological T stage	pT(0-2) pT(3-4)	$27-17$ $50-9$	0.0110
Pathological N stage	pN0 pN(1-2)	$51-22$ $26-3$	0.0424
Lymphatic invasion	Yes No	$29-2$ $45-22$	0.0049
Venous invasion	Yes No	$27-3$ $47-20$	0.0399
Perineural invasion	Yes No	$26-2$ $50-21$	0.0179
CRM			0.0554

Table 6.3: Significant categorical parameters for regrouped bad vs good response grading system.

We used the significant clinical and surgery-related variables available to determine whether or not a patient (with Dworak 2 or 3) fell into the gray-zone category. These patients were characterized as such when some of the variables indicated a response that was contrary to that given by the Dworak score. For instance, some patients who were identified as Dworak grade 2 (indicating poor response), based on microscopic examination of tumor specimens, exhibited a negative resection margin, no lymphatic-, perineural-, or venous invasion, a low tumor grade, and low T- and N-stages. These parameter values typically indicate a good response, leading to the classification of these patients as gray-zone patients. Analogously, some patients who identified as Dworak grade 3 (indicating good response) had a high tumor grade, presence of lymphatic, venous, or perineural invasion, and higher (3-4 and 1-2) pathological T- and N-stages. These patients displayed poor parameter values and were also classified as gray-zone patients.

Since we are only interested in finding out whether or not there are patients for which it is not clear if they are a bad or a good responder, we only focused on the Dworak grades 2 and 3 as these grades form the boundary between the regrouped regression grade (see Table 6.4). There were 46 patients in the study with Dworak grades 0, 1, or 4. These patients were excluded from

the gray-zone analysis. From the remaining 60 patients, 25 patients were found to be gray-zone patients and the remaining 35 were non gray-zone patients.

Dworak TRG		**Number of gray-zone patients**	**Number of CT images available**
Grade 0	Excluded	2	2
Grade 1	Excluded	37	25
Grade 2	Yes	17	13
	No	24	20
Grade 3	Yes	8	8
	No	11	10
Grade 4	Excluded	7	6

Table 6.4: Overview of the gray-zone patients. Dworak grades 0,1, and 4 are excluded from the gray-zone analysis.

6.6.2 Impact of gray-zone patients to the ERSF

We constructed an ERSF using the 141 patients for whom we have both a CT scan and Dworak TRG. We created $n_T = 200$ trees per iteration and performed K-fold CV within each tree ($K = 6$) to assess overfitting. After generating n_T trees in each round, poorly performing trees were pruned, resulting in a variable number of trees in the forest across different iterations. In total, we executed $n_F = 1000$ iterations.

The ERSF model trained on CT images for all the 141 patients in the data set resulted in a forest with a final number of trees, $n_T = 199$. We opted to evaluate our ERSFs using a performance metric that considers the imbalance in our data set. When dealing with unbalanced data, more predominant classes may influence the outcomes, potentially leading to more positive or negative results. As our goal is to accurately predict both bad and good responders, we normalize our confusion matrix by dividing each entry by the sum of its corresponding row, thereby eliminating any influence of the output frequency. The accuracy obtained from this normalized confusion matrix is referred to as proportional accuracy (pACC). We obtained a training pACC of 0.760 and a LOOCV pACC of 0.654 for the full data set of 141 patients. Although pACC is the primary evaluation metric in our study, we also reported the precision, recall, F1-score, and the AUC for all results, including the full data set of 141 patients, the gray-zone study, and the added regularization to the gray-zone study (see Table 6.5). Additionally, the ROC curves for all results can be found in Appendix D.

Focusing solely on the 85 patients available for the gray-zone study, we examined the validation of the ERSF and found a LOOCV pACC of 0.603, which is slightly lower compared to the performance on the full data set of 141 patients.

Table 6.6 shows confusion matrices for the three different patient groups: gray-zone, non gray-zone and excluded patients. We observe that the ERSF algorithm struggles the most with accurately predicting the gray-zone patient group. The LOOCV pACC for these patients is only 0.380, while it is 0.700 and 0.673 for the non gray-zone and excluded patient groups, respectively.

Case	pACC	Precision	Recall	F1-score	AUC
Full ERSF - Training ($n = 141$)	0.760	0.895	0.701	0.786	0.837
Full ERSF - Validation ($n = 141$)	0.654	0.808	0.649	0.720	0.698
gray-zone ($n = 21$)	0.380	0.500	0.385	0.435	0.490
Non gray-zone ($n = 30$)	0.700	0.824	0.700	0.757	0.785
Excluded ($n = 34$)	0.673	0.905	0.679	0.776	0.655
Regularization - gray-zone ($n = 21$)	0.697	0.769	0.760	0.762	0.837
Regularization - non gray-zone ($n = 30$)	0.925	0.950	0.950	0.950	0.960
Regularization - excluded ($n = 34$)	0.928	1.000	0.857	0.923	0.929

Table 6.5: Results for the different prediction cases.

Gray-zone

Observed	Predicted: Bad	Predicted: Good
Bad	5	8
Good	5	3

Non gray-zone

Observed	Predicted: Bad	Predicted: Good
Bad	14	6
Good	3	7

Excluded

Observed	Predicted: Bad	Predicted: Good
Bad	19	9
Good	2	4

Table 6.6: Confusion matrices for the three different subgroups.

To incorporate the regularization parameter into the CV analysis, we followed a process where we began with no regularization on the covariance matrix and gradually increased the amount of regularization as long as the predictions remained incorrect. We adopt regularization based on Friedman's theory, where a regularization parameter $\lambda = 1$ corresponds to no regularization and $\lambda = 0$ eliminates the entire covariance matrix replacing it with only the identity matrix. We employ this type of regularization to reflect the opinions of clinicians. Patients who have unique characteristics and do not fit the common pattern might benefit from this type of regularization when suggested by a clinician. By incorporating a clinician's opinion and the ability to adjust the regularization parameter, we aim to move towards a more personalized treatment approach.

We see that some patients that were originally wrongly classified can correctly be classified when choosing the optimal regularization parameter. However, some patients cannot be predicted correctly even though we change the regularization parameter. Table 6.7 shows that after introducing a regularization parameter we were able to correctly classify 0.697 of the gray-zone patient, 0.925 of the non gray-zone patients, and 0.928 for the excluded patients. It is evidently clear that the gray-zone patients are the most difficult to correctly classify.

Gray-zone

Observed	Predicted: Bad	Predicted: Good
Bad	10	3
Good	3	5

Non gray-zone

Observed	Predicted: Bad	Predicted: Good
Bad	19	1
Good	1	9

Excluded

Observed	Predicted: Bad	Predicted: Good
Bad	24	4
Good	0	6

Table 6.7: Confusion matrices for the three different subgroups after regularization.

6.7 Discussion

6.7.1 Gray-zone patients

There were no differences found among the three different patient groups regarding the algorithm's sureness parameters. It is logical for the algorithm not to show a difference in certainty parameters for patients excluded from the gray-zone study and patients considered non-gray zone patients. However, we expected to see a difference in the sureness parameter for gray-zone patients since it's uncertain whether they are good or poor responders. Nevertheless, the ERSF is a "fictional" doctor that has its own opinion and makes predictions for gray-zone patients based on all the factors presented to it, just like pathologists do.

We must not overlook the invaluable capability of doctors to think outside the box, interpreting nuanced symptoms, considering patient histories, and taking into account other factors that may not be explicitly captured in data. This ability is a strong advantage that ML lacks. The subjectivity that doctors bring to their practice enables them to make informed decisions that consider the uniqueness of each patient. Therefore, it is crucial to integrate and harmonize the knowledge that ML can provide, which may not be apparent to the doctor, with the expertise and sound judgment of the doctor.

We observed significant differences in the accuracy prediction for the three groups. As anticipated, the group with patients excluded from the gray-zone study performed well in the LOOCV, with an accuracy of 67.26%, more or less the same as that obtained by the ERSF on the 141 patients. It was surprising to see that non-gray zone patients had an even better performance, with 70% accuracy. For gray-zone patients, we expected a lower prediction accuracy, and we indeed got only 37.98%, indicating that our algorithm's opinion differed more from the observed grades compared to the other two groups. This is not surprising since the real regrouped regression grade of these patients is debatable.

6.7.2 Regularization analysis

Regarding the regularization analysis, we observed that more patients in the gray-zone group required regularization than the other two groups, as there were more patients predicted incorrectly in this group. Moreover, we found that reducing the regularization parameter and thereby reducing the information from the covariance matrix resulted in more patients in the gray-zone group who could not be predicted correctly compared to patients in the other two groups. Even when we added the regularization parameter, we could not correctly classify 6 out of 21 (28.57%) gray-zone patients. For non-gray zone and excluded patients, the percentage of patients who were not predictable was only 2 out of 30 (6.67%) and 4 out of 34 (11.76%), respectively. We, therefore, saw that the accuracy of the gray-zone patients remained significantly lower than the other two groups even after regularization, with a pACC of only 0.697 for the gray-zone patients and nearly 0.93 for the other two groups. The algorithm may disagree more often with the doctor's opinion in the gray-zone group than in the other two groups. This could be because of the subjectivity in the Dworak grades 2 and 3 or because these patients are wrongly predicted by either the algorithm or the human doctor.

6.7.3 Mitigating subjectivity in Dworak grading system

Moving forward, it is crucial to address the subjectivity inherent in the grading of the Dworak system. One obvious solution would be to involve multiple pathologists in grading the Dworak scores, averaging their assessments to mitigate subjectivity. However, this approach is neither cost-effective nor a good allocation of time. Alternatively, one could employ a single pathologist

to grade all patients. Yet, this is impractical for several reasons – for instance, the pathologist might leave the study or hospital, and even if they remain active, retirement would eventually become a factor.

Other grading systems with reduced subjectivity could potentially be adopted instead of the Dworak. Unfortunately, all current grading systems for rectal cancer, such as the Mandard, Ryan, or (Modified) Dworak systems, are subjectively graded and share similar characteristics and therefore limitations [106]. One possible approach is to combine these various subjective grading systems, evaluating their concordance to reduce overall subjectivity. However, it might be more beneficial to establish better-defined rules for each grading system to minimize subjectivity as much as possible.

A potentially more efficient solution would be to develop a predictive model for the Dworak regression grade. This model could incorporate laboratory values, endoscopic images, surgical and pathological notes, and other data collected immediately before, during, or from the biopsy to estimate the Dworak grade. Unlike the ERSF, which aims to predict the Dworak grade before the start of any therapy to avoid unnecessary surgeries potentially, this model would use data collected after therapy, closer to the time of surgery, and even from the surgical procedure itself. Developing such a model could be the most time- and cost-efficient option. However, additional research is required to evaluate and implement it effectively.

6.8 Conclusion

The study results presented here take a first step in bridging the gap between medicine and the ERSF algorithm in predicting rectal cancer patients' response to (radio)chemotherapy based on planning CT images. It highlights an uncertainty measure between the Dworak TRGs 2 and 3 that causes the algorithm to misclassify more patients than any other patients in the data set. We confirmed in our study that the instability in the Dworak TRG indeed impacts the ERSF algorithm, where the gray-zone patients are more difficult to correctly classify by the ERSF than others. With this, we showed that the quality of classification does not only depend on the quality of the ML algorithm solely but also depends on the stability of the data. If the data is unstable, one cannot expect the ML algorithm to produce superb results. As we are already faced with some repeatability and reliability issues of the radiomics features, it is normal that the instability of the Dworak TRG will impact the prediction performance of the ERSF.

It is therefore crucial to find a method to incorporate this uncertainty of the regression grade into the ML algorithm. By doing so, we could gain more insight into the algorithm, opening up the black box. Additionally, incorporating more subjective information from doctors could improve the algorithm's predictions and give doctors more confidence in using it.

6.9 Appendix

6.9.1 Significance between the regrouped Dworak regression grade and clinical/pathological variables

Results for non-significant variables using Fisher's exact test for categorical variables and the ANOVA test for numerical variables are given in Table 6.8.

Variable name		**Number of bad - good responders**	**p-value**
Histological Type	Adenocarcinoma Mucinous carcinoma	$73-21$ $7-4$	0.288
Tumor Location	Low rectum Mid rectum High rectum	$26-11$ $30-7$ $9-2$	0.602
Pre-operative therapy	Radiotherapy Chemoradiotherapy	$59-20$ $21-6$	1
T stage	$T(0-2)$ $T(3-4)$	$6-2$ $73-24$	1
N stage	$N(1-2)$ N0	$73)25$ $6-1$	0.678
M stage	M0 M1	$72-26$ $!-1$	0.448
Age			0.593
Distance Anal Verge			0.388

Table 6.8: Non-significant categorical parameters for the regrouped bad vs good response grading system.

6.9.2 ROC-curves

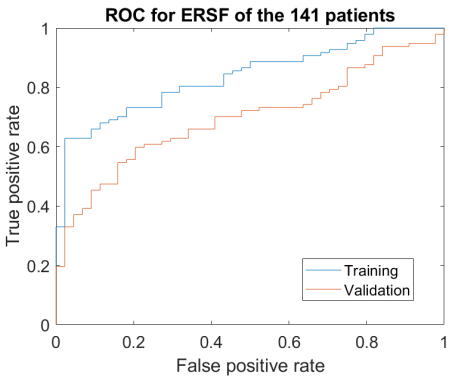


Figure 6.2: ROC curve for both the training and validation of the entire 141 patients data set.

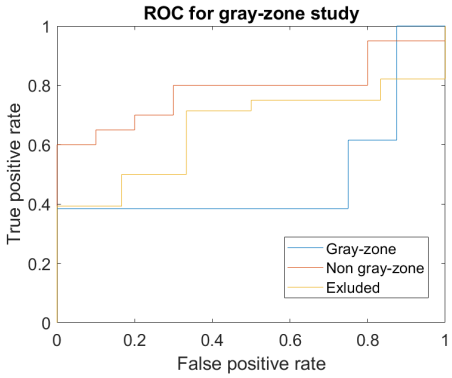


Figure 6.3: ROC curve for the validation of the gray-zone study.

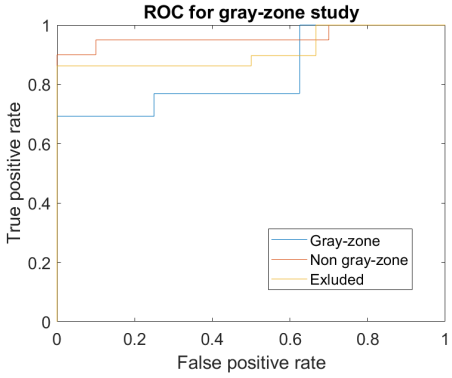


Figure 6.4: ROC curve for the validation of the gray-zone study using regularization.

Chapter 7

Deep Fourier Features

This chapter is based on the publication:
C. Raets, C. Aisati, A. Rifi, M De Ridder, K. Putman, J. De Mey, A. Sermeus, K. Barbé, "Optimizing rectal cancer patient care: Dworak TRG prediction via bayesian evolutionary fourier-domain random subspace forest," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

7.1 Introduction

Often in medical imaging, spatial features like radiomics are used for prediction. These features describe the distribution of the voxel intensities of the image, which can include intensity, shape, and/or texture features.

However, in Chapters 3 and 4, we showed that Fourier plays an important role in medical imaging. We hence thought it to be interesting to look at the predictive value of features extracted using the Discrete Fourier Transform (DFT). Fourier features, in contrast to the radiomic features, capture features from the frequency domain rather than in the spatial domain.

In this chapter, we want to see whether frequency domain features might be interesting to add into our prediction model. These features might be able to capture complex patterns in the frequency domain that the spatial features cannot capture. We will develop a novel approach of extracting custom deep Fourier features from the same planning CT images used for the radiomics extraction. We used the 3D Discrete Fourier Transform (DFT) on the voxel values from the planning CTs. Since the number of voxels in the images is extremely larger than the number of samples in the data set, we need to rescale the image to a smaller size such that the number of features is smaller. Moreover, we need to rescale the images to get the same number of voxels over all images. This is necessary as the images all differ in sizes.

Furthermore, in a later chapter, we will see if adding the information gained from the radiomic features in the spatial domain and the Fourier features in the frequency domain together can improve the results even further.

Additionally we will do some transformations of the voxels such that the pixel size and slice thickness settings are more uniform over the different images. Here we will look at the different setting options present in the data set and see whether some settings are better than others.

7.1.1 Problem statement

With the ethical committee approval EC1010135, we had a total of 141 patient files available for our study. However, after some inspection we noticed that we had one patient who was added twice to the data set and another who was wrongly included to our data set, leaving us with 139 patients. The remaining 139 patients were all treated for rectal cancer, and all of them had a planning CT scan taken before the start of the therapy and underwent surgery from which the regression grade of the cancer after the treatment was determined.

7.2 Deep Fourier Features

In this chapter, we aim to extract Discrete Fourier Transform (DFT) features from the same CT images to explore potential insights when incorporating this new data set into our ERSF. Fourier transforms are a mathematical technique used to convert signals into their frequency domain, generating a complex valued output. When working with discrete data, such as medical images, we refer to this process as the DFT.

Since every tumor varies in size, the resulting number of DFT frequencies obtained by the DFT will not be consistent across different patients. Therefore, we need to preprocess the images to ensure a uniform number of features for all patients. From the three-dimensional CT images, we extract the HU and store the gray values in a three-dimensional matrix. Here, we first crop the image to a three-dimensional matrix containing only the ROI, which in our study is the GTV, before extracting the deep Fourier features from it (see Figure 7.1).

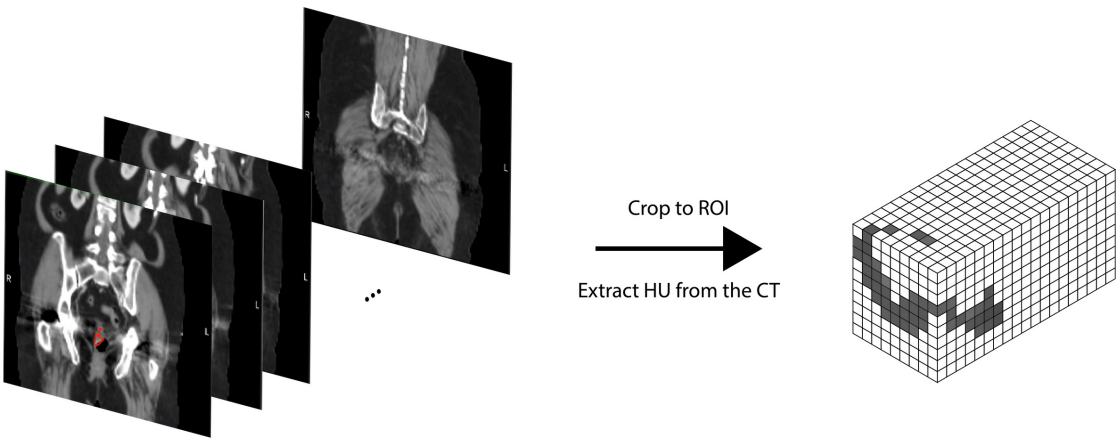


Figure 7.1: Store the ROI from the CT scan as gray-values in a 3D matrix.

Assume that we want our images, after DFT, to have a final size of $n_x \times n_y \times n_z$, resulting in a consistent number of features equal to $n_x n_y n_z$. Without loss of generality, let X represent a matrix of the CT scan of a patient in our data set, where each element $X_{l,k,m}$ represents the intensity value of a voxel at the lth row, kth column, and mth slice of the 3D scan. Furthermore, let $N_x \times N_y \times N_z$ denote the size of the scan in the three dimensions, such that the matrix X consists of $N_x N_y N_z$ gray-valued voxels. A schematic overview of the entire deep Fourier features extraction process is given in Figure 7.2.

As we want to resize the DFT of each image to the size $n_x \times n_y \times n_z$, it is essential to ensure that the dimensions of the images are multiples of the desired final size. We obtain this by

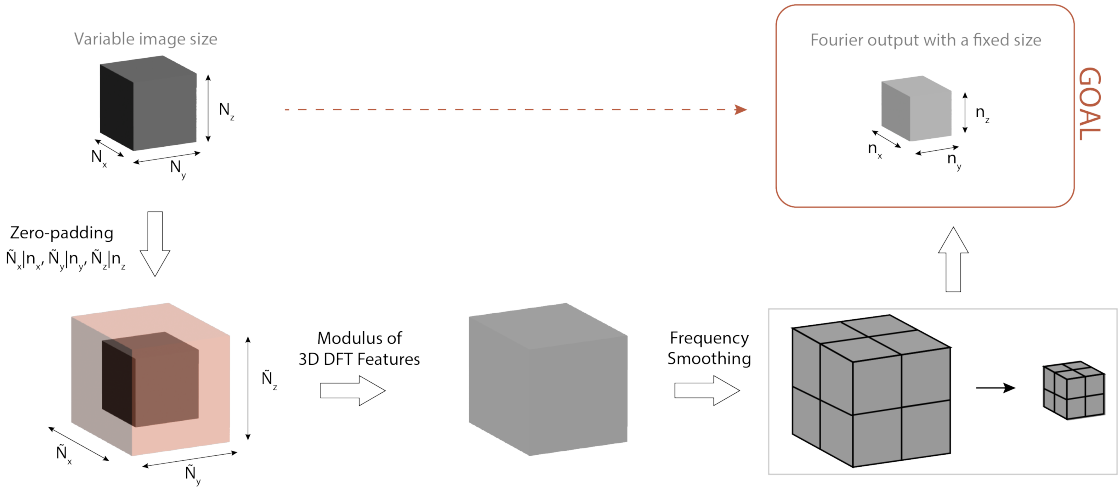


Figure 7.2: Deep Fourier features extraction.

employing zero-padding, where we add zeros in all necessary dimensions to create a matrix of the correct size. The resulting zero-padded matrix, denoted by $\tilde{X} \in \mathbb{R}^{\tilde{N}_x \times \tilde{N}_y \times \tilde{N}_z}$, now has new dimensions $\tilde{N}_x \times \tilde{N}_y \times \tilde{N}_z$. In each dimension, the new size is the smallest multiple of the final size that is greater than or equal to the original size.

Concretely, in the x direction, we add $\lceil (\tilde{N}_x - N_x)/2 \rceil$ zeros to the left side of the image and $\lfloor (\tilde{N}_x - N_x)/2 \rfloor$ zeros to the right side of the image. If the number of zeros that need to be added is even, the zeros are evenly distributed on both sides; if not, we add one more zero on the left than on the right side. The process is analogous for all the other dimensions. It's important to note that by adding a zero boundary, our image becomes periodic, which is a requirement for the DFT

After adding zero boundaries where necessary to ensure the image is of the correct size, we can compute the three-dimensional DFT of the zero-padded image $\tilde{X} \in \mathbb{R}^{\tilde{N}_x \times \tilde{N}_y \times \tilde{N}_z}$ using the following formula,

$$\hat{\tilde{X}}(u_x, u_y, u_z) := \sum_{k_x=0}^{\tilde{N}_x-1} \sum_{k_y=0}^{\tilde{N}_y-1} \sum_{k_z=0}^{\tilde{N}_z-1} c_{\tilde{N}_x}^{u_x k_x} c_{\tilde{N}_y}^{u_y k_y} c_{\tilde{N}_z}^{u_z k_z} \tilde{X}_{k_x, k_y, k_z}, \tag{7.1}$$

where $c_j := \exp\left\{-\frac{u_j k_j}{\tilde{N}_j} 2\pi i\right\}$ for $j \in \{x, y, z\}$ [107]. The DFT elements $\hat{\tilde{X}}(u_x, u_y, u_z)$ form a complex three-dimensional matrix $\mathcal{F}(\tilde{X})$. As working with complex features can be difficult, we take the modulus of the DFT matrix to obtain a real-valued three-dimensional matrix, denoted as $\left|\hat{\tilde{X}}(u_x, u_y, u_z)\right|$.

We are now left with a three-dimensional matrix that varies in size across different patients in the data set. To ensure that we have the same number of features for all patients, we need to transform this matrix such that it has the desired size of $n_x \times n_y \times n_z$. This transformation involves frequency smoothing, where blocks of data are taken together, and the square root of the average of the quadratic values is computed. Since we already ensured that the size of each dimension is a multiple of the desired final size, we can combine even-sized blocks of data together.

An element at index position (u, v, w) of the smoothed matrix $M \in \mathbb{R}^{n_x \times n_y \times n_z}$ is calculated

by

$$M(u,v,w) = \left(\frac{n_x n_y n_z}{\tilde{N}_x \tilde{N}_y \tilde{N}_z} \sum_{k_x=1}^{\frac{\tilde{N}_x}{n_x}} \sum_{k_y=1}^{\frac{\tilde{N}_y}{n_y}} \sum_{k_z=1}^{\frac{\tilde{N}_z}{n_z}} \left| \hat{\tilde{X}} \left((u-1)\frac{\hat{\tilde{N}}_x}{n_x} + k_x, (v-1)\frac{\tilde{N}_y}{n_y} + k_y, (w-1)\frac{\tilde{N}_z}{n_z} + k_z \right) \right|^2 \right)^{1/2}. \tag{7.2}$$

7.3 Pixel spacing and slice thickness alterations

As we are working with a retrospective data set, it is not surprising to find different CT scanners and/or different scanner settings among the patients in our sample. Consequently, there are differences in pixel spacing and slice thickness to be found among the patients. However, from a mathematical perspective, each voxel is represented as an element, making it difficult to detect differences between pixel spacing and slice thickness settings.

In image processing, pixel alterations are often necessary to achieve the desired pixel spacing. When the original pixel value is a multiple or divisor of the desired pixel spacing, the adjustment process is straightforward. In such cases, we can simply divide or calculate the average over multiple pixels to achieve the desired spacing. However, when the original pixel value is not a multiple or divisor of the desired spacing, we can alter the original pixel spacing such that it is as close to the desired pixel spacing as possible.

For each slice, the pixel spacing in the x and y direction is the same. To simplify the explanation, we will focus on the one-dimensional case, where we aim to change a pixel spacing P to the desired pixel spacing S. The alterations in higher dimensions follow the same principles. There are several cases to consider:

1. If $P = S$, then no alterations are necessary.
2. $P \mid S$ or $S \mid P$
   (a) If $P \mid S$, i.e. $S = kP$ for some $k \in \mathbf{N}$, then the pixel spacing P is smaller than the desired spacing S. In this case, we can alter the pixel spacing P by taking the average of k consecutive pixels.
   (b) If $S \mid P$, i.e. $P = kS$ for some $k \in \mathbf{N}$, then the pixel spacing P is larger than the desired spacing S. In this case, we can just replace each pixel by k consequential pixels with the same pixel value.
3. If $P \nmid S$ and $S \nmid P$, then we need to find the closest pixel spacing S' to S for which $P \mid S$.
   (a) If $P < S$, then we can find a k such that $S' := kP$ for which $k := \operatorname{argmin}_{\lambda \in \mathbf{N}} |S - \lambda P|$. The problem now boils down to case 2a.
   (b) If $P > S$, then we can find a k such that $S^{:} = P/k$ for which $k := \operatorname{argmin}_{\lambda \in \mathbf{N}} |S - P/\lambda|$. The problem now boils down to case 2b.

This approach allows for precise pixel alterations, even when the desired spacing is not a multiple or divisor of the original pixel value. It ensures that the resulting image maintains the desired spacing characteristics while preserving image quality. Our hypothesis is that by standardizing the pixel spacing and slice thickness across patients, the prediction precision will be improved.

7.4 Results

Remember that our ERSF using the radiomic features gave a proportional training accuracy of 77.51% and proportional validation accuracy of 67.08%. We now additionally extracted additionally deep Fourier features from the same planning CT images. As we wanted to obtain a features size that is similar to the 109 radiomic features and since on average the slice thickness was twice the pixel setting, we chose a final matrix size after frequency smoothing to be set equal to $6 \times 6 \times 3$ resulting in 108 deep Fourier features.

Subsequently, we created an ERSF using this new data and obtained a proportional training accuracy of 82.40% and an AUC of 88.16%. During the LOOCV, we obtained a proportional accuracy of 59.13% and an AUC of 63.56%.

Additionally, we are interested in examining the influence of the different pixel spacing settings and slice thickness values used. Since we are working with a retrospective data set, we have encountered a wide range of pixel spacing and slice thickness combinations in our data. Upon close inspection, we identified 14 different combinations within our data. Table 7.1 shows the 14 combinations options, labeled with numbers 1 to 14 for easy reference.

Option Nr	**Slice Thickness**	**PixelSpacing**
1-108	1.500	1.172
2-332	1.500	0.977
3-479	1.500	1.953
4-350	2.500	0.855
5-469	2.500	1.953
6-309	3.000	0.977
7-310	3.000	1.953
8-340	3.000	3.906
9-363	3.000	1.711
10-415	3.000	1.387
11-456	3.000	1.621
12-362	5.000	1.172
13-391	5.000	1.523
14-470	5.000	1.953

Table 7.1: Different pixel spacing and slice thickness combinations (in mm) present in the data.

We altered the pixel spacing and slice thickness settings for all patients' CT scans according to the 14 different combinations options as described in Section 7.3. This resulted in the creation of 14 new subsets of CT data, each using the 14 different options of pixel spacing and slice thickness settings. Subsequently, we extracted deep Fourier features from these 14 different subsets, maintaining again a final matrix size of $6 \times 6 \times 3$ after frequency smoothing.

We created different ERSFs for all 14 different data sets. The results are summarized in Table 7.2. We observed that the proportional training accuracy (pACC) and AUC for the data sets using option numbers 9, 11, and 12 were lower than those of the ERSF created with the original deep Fourier features, without pixel spacing and slice thickness alterations. However, for most altered data sets, the training performance improved with these alterations compared to the original data. In terms of validation, most altered data sets yielded better results than the original data. Notably, the data set constructed using setting option 14 produced outstanding results for both training and validation, where the pixel spacing and slice thickness were set relatively large compared to the other combinations.

	Training		Validation	
Ref ID	**pACC**	**AUC**	**pACC**	**AUC**
1	0.851	0.935	0.559	0.638
2	0.884	0.953	0.614	0.686
3	0.962	0.988	0.615	0.674
4	0.917	0.986	0.585	0.596
5	0.858	0.903	0.671	0.690
6	0.868	0.943	0.692	0.703
7	0.858	0.903	0.671	0.690
8	0.950	0.996	0.587	0.636
9	0.701	0.717	0.650	0.674
10	0.851	0.920	0.686	0.700
11	0.673	0.699	0.600	0.597
12	0.709	0.734	0.671	0.667
13	0.895	0.969	0.604	0.629
14	0.906	0.957	0.727	0.724

Table 7.2: Prediction results in percentages for the data sets with altered pixel spacing and slice thickness using newly created forests without prior information for each new data set.

7.5 Discusion

We extracted a total of $p = 108$ deep Fourier features from the planning CT images, where we set the final size after frequency smoothing to $6 \times 6 \times 3$. We chose this final size as we observed that on average the tumor dimension was twice as large in the xy-plane than in the z-plane. Furthermore, with this combination, the number of features was similar to the number of radiomics features. Obviously, more research has to be done to find the ideal combination for the final size. However, one also has to keep in mind that when increasing the feature size, the sample size has to be sufficiently large to accommodate the feature size. We saw that the ERSF with the radiomics data performed better in terms of validation than the ERSF using the deep Fourier data.

The ERSF results, in terms of ACC and AUC, for the altered pixel spacing and slice thickness for the 14 different options settings showed that the uniformization of the settings might be beneficial for the training. However, we saw that for a few options, the alterations were not benefiting the training results. For the LOOCV, we also saw that most results were up to 10% better with the alterations than without. However, some options could be seen showing lower LOOCV results after the alterations. The performance accuracy of the validation is extremely important for clinical use, as many different patients from outside the training data will be validated.

It is not entirely clear why some pixel spacing settings perform better than others, and why some do not perform well at all. Furthermore, we did see that for option number 14 the results were extremely good for both training and validation results.

7.6 Conclusion

Overall, this chapter demonstrates the significance of altering the pixel spacing and slice thickness settings in medical image analysis for predicting the response to rectal cancer treatment. We have

shown that uniformizing these settings can enhance the prediction performance of our ERSF. Importantly, we introduce an innovative approach of adding yet another additional layer to our ERSF by first applying pixel spacing alterations before extracting the DFT data and constructing the ERSF and then extracting the prior information that can be used to create a forest with another data set.

Furthermore, we recommend exploring similar alterations in pixel spacing and slice thickness for radiomics data, as this may further enhance predictive accuracy. Our study opens the door to more research questions regarding scanner settings and their influence on prediction accuracy, promising exciting avenues for future investigations.

Chapter 8

Bayesian Approach in the ERSF

This chapter is based on the publication:
C. Raets, C. Aisati, A. Rifi, M De Ridder, K. Putman, J. De Mey, A. Sermeus, K. Barbé, "Optimizing rectal cancer patient care: Dworak TRG prediction via bayesian evolutionary fourier-domain random subspace forest," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

8.1 Introduction

Research in both medicine and AI is a rapidly evolving domain, where new progress is made each day. Over the years, substantial progress has been achieved in the realm of cancer studies using AI techniques, including areas such as screening and treatment. Among the array of tools aimed at refining cancer detection and treatment, machine learning, feature selection algorithms, and predictive modeling make a strong impression [108]. In a rapidly evolving domain, we have to be flexible and add new data whenever possible.

The method of incorporating data is crucial for the algorithm's prediction accuracy. Adding new data to the existing data set would further expand the feature set, potentially causing overfitting issues, especially with smaller sample sizes. On the other hand, if we were to use the data sets separately for prediction, we could overlook valuable information that becomes apparent when they are used together.

Given that we already have two different data sets for predicting our TRG, we were interested in the potential benefits of merging them. However, both data sets have an extensive number of features, with 109 radiomic features and 108 deep Fourier features. Consequently, merging them into a single large data set comprising 217 features for only 139 patients is not advisable. Nevertheless, our utilization of LDA in the ERSF provides an opportunity to integrate prior information into the model. Instead of merging the data sets entirely, we propose extracting prior information from the first data set and incorporating it into the ERSF constructed with the second data set. This approach allows us to exploit the information of both data sets without necessitating their combination.

8.2 Bayesian ERSF

In this chapter, we aim to enhance our ERSF algorithm with an additional step involving Bayesian inference. Bayesian inference employs Bayes' theorem to calculate a posterior probability using a prior probability with a likelihood function [94]. Given that we already employ Bayes' theorem in constructing our LDA tree, integrating the prior distribution to create a Bayesian ERSF is a straightforward extension. Since the Bayesian inference will be utilized at our tree level, our focus in this section lies solely on the construction of the trees.

In our algorithm, each tree is constructed using LDA, which assumes that the predictor variables are normally distributed. Although this technique was originally introduced by Fisher in 1936, we follow the mathematical notation as described by Friedman in 1989 [80, 93].

Let's assume again that $\underline{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})$ denotes a random vector consisting of $p \in \mathbb{N}$ normally distributed random variables, and let $m \in \{1, \ldots, M\}$ represent the different possible outcome groups. In our study, we started our analysis with Bayes' rule and the rule of conditional probability, which can be expressed as follows:

$$\mathbb{P}(m|\underline{X}) = \frac{\mathbb{P}(m)\mathbb{P}(\underline{X}|m)}{\mathbb{P}(\underline{X})}, \tag{8.1}$$

where $\mathbb{P}(m)$ is the prior distribution and $\mathbb{P}(m|\underline{X})$ is the posterior distribution or data distribution [94].

In Section 5.3.1 of Chapter 5 we looked at the LDA tree and obtained the following equation (see eq. (5.4)):

$$\hat{m} = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}} \frac{\mathbb{P}(m)}{\sqrt{(2\pi)^p|\Sigma_m|}} e^{-\frac{1}{2}(\underline{X}-\underline{\mu}_m)\Sigma_m^{-1}(\underline{X}-\underline{\mu}_m)'}, \tag{8.2}$$

where $\mu_m := \mathbb{E}[\underline{X}|m]$ represents the population mean vector for the m-th class and $\Sigma_m := Cov(\underline{X}|m)$ is the population covariance matrix for that m-th class. In Chapter 5, we assumed that the prior distribution $\mathbb{P}(m)$ was equal across all groups. Now, we will not make this assumption; thus, we must slightly adapt our formula from the final one given in Section 5.3.1.

By taking the natural logarithm of equation (5.4), we can simplify the formula. Furthermore, in LDA, we assume that the covariance matrices do not vary across different classes, i.e. $\Sigma := \Sigma_m$ for all $m \in \{1, \ldots, M\}$. Therefore, we arrive at the following equation:

$$\hat{m} = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}} \left(\underline{X} - \frac{1}{2}\underline{\mu}_m\right)\Sigma^{-1}\underline{\mu}_m' + \log\left(\mathbb{P}(m)\right) \tag{8.3}$$

8.2.1 Extraction of the Prior Information

We can now first construct an ERSF using the initial formula and extract the prior information from that ERSF. Then, we can build a new ERSF, this time using eq. (8.3), and incorporate the prior information extracted from the first ERSF. Once an ERSF is fully trained and n_F trees remain in the final forest, we can extract the prior information for each patient though LOOCV.

For each of the n_F trees in the forest, we make a prediction for the validation patient using the mean vector and covariance matrix computed using the training data. Let $\hat{y}_i$ be the prediction of the i-th tree for the validation patient, where $i \in \{1, \ldots, n_F\}$. The prior probability for the entire forest can be calculated for all groups as follows:

$$\mathbb{P}(m) = \frac{\sum_{i=1}^{n_F} \mathbf{1}_{\{\hat{y}_i=m\}}}{n_F}. \tag{8.4}$$

We can perform LOOCV for each patient, using them once as a validation patient, and calculate all priors for all patients in this manner.

8.3 Results

We chose to extract prior information from the deep Fourier data and use it to model our radiomic features data. This way we would be able to first get some information and prediction pattern for the frequency domain and use the information into the spatial features domain.

Remember that in the previous chapter we already obtained results for both the radiomics data and the deep Fourier features. Our ERSF using radiomics obtained a proportional training accuracy of 0.775 and proportional validation accuracy of 0.671. Using the deep Fourier, we obtained a proportional training accuracy of 0.824 and proportional validation accuracy of 0.591. Furthermore, we obtained ERSF results for the 14 different setting options using the deep Fourier features (see Table 7.2).

We extracted prior information from all 14 different deep Fourier features data sets with altered spacing and created 14 new ERSFs using the radiomics features and the 14 different deep Fourier features sets. The results for all 14 different test cases are given in Table 8.1.

We observed that in all cases, the training performance was inferior to when no priors were used for creating the ERSF with the radiomics data. For the forests with priors from the data sets using option numbers 4, 8 and 12, the validation performance was also less well compared to using no priors. However, for all other options, the validation results were better with the prior information. It was also notable that the difference between the training and validation results was much smaller when using the priors.

The ERSF with prior information from the data set using option number 10 showed ideal results, achieving a decent training accuracy of 73.78% and a validation accuracy of 72.65%. Interestingly, we observed peculiar results for the ERSF using option number 14, where the validation accuracy was higher than the training accuracy, with 71.93% for validation and 71.42% for training. In this case, one more patient was correctly classified in the validation set compared to the training.

8.4 Discussion

When testing the radiomics data again, but this time using the different setting options as prior, we saw that for most cases the training results were lower than the training accuracy of the ERSF with radiomics data and no priors. However, the validation results were better for most test cases. We also saw that the results for training and validation were closer together when using the priors, indicating that these ERSFs could be generalized better. This generalization is important to ensure good clinical usage. From the validation results, it is clear that the prior gives more information to the forest, making the validation better.

There was also option number 14, where the validation outperformed the training by achieving one more correct prediction. Lastly, the best case was that of option numbers 2 and 10 where the training and validation both were above 70%.

8.5 Conclusion

We introduced an innovative approach of adding yet another additional layer to our ERSF by first applying pixel spacing alterations before extracting the DFT data and constructing the

	Training		Validation	
Ref ID	**Acc**	**AUC**	**Acc**	**AUC**
1	69.6	74.6	66.8	69.7
2	72.8	78.9	70.0	72.2
3	74.3	76.4	66.8	70.3
4	70.0	76.6	62.0	66.2
5	72.8	77.0	69.5	73.2
6	71.6	74.2	69.5	71.4
7	72.8	77.0	69.5	73.2
8	72.2	77.2	63.0	69.9
9	72.3	75.8	67.4	70.6
10	73.8	75.9	72.7	74.4
11	72.8	76.7	67.2	71.6
12	71.6	78.0	57.2	66.2
13	69.6	75.0	67.5	69.6
14	71.4	72.1	71.9	72.1

Table 8.1: Prediction results in percentages for the radiomics data using newly created forests with prior information extracted from each altered data set.

ERSF and then extracting the prior information that can be used to create a forest with another data set.

It must also be noted that even though the training results of the ERSFs with priors were lower than the ERSF created without priors, the validation was in most cases better. Furthermore, this novel approach not only improves the algorithm's generalization but also offers valuable insights into the impact of standardization in medical imaging before making the ERSF. The enhanced generalization is particularly promising for the practical application of our algorithm in a hospital setting, this characteristic is extremely important.

In summary, our ERSF, with its added layer of pixel spacing alterations and prior information extraction, can be seen as a Convolutional Neural Network (CNN) with a unique hidden layer, while also embodying Bayesian principles into the ERSF. These innovations highlight the potential of our approach as a valuable asset in predictive modeling for healthcare.

Conclusion

Discussion

Over the years, we have created a custom-built ML algorithm tailored to our specific prediction problem. We were faced with a challenging problem of predicting the Dworak TRG for rectal cancer patients using only the information obtained from the planning CT images taken before the start of the therapy. From the start we already knew that the state-of-the-art classification algorithms and feature selection algorithms wouldn't work as both the output variables and the input variables were either subjectively graded are not completely repeatable and /or reproducible.

Radiomic features are prone to subjectivity. It has been shown that the equipment and settings used influence the quantitative features where sometimes there are significant differences between different equipment and/or settings used [70]. It was also shown that even when using the same equipment and settings, the features can vary significantly when performing a coffee break study where the subjects were scanned twice with a 15-minute break in between them. How severe the differences are, depends on the scanner and settings used. As we are working with retrospective data that was collected over a long period, we have multiple different scanners with different settings in our data. Furthermore, we have in many cases no knowledge of the scanner or settings used as this information was not saved

Just like the Radiomics, the Dworak TRG is prone to subjectivity as it is a semi-quantitative measure. It is used as a measure of how dominantly the tumor is still present after the treatment and is determined by a pathologist. The given TRG will be influenced by many different factors like the risk aversion of the pathologist, the mental and physical health of the pathologist, or even other factors like the environment and equipment used. The Dworak TRG consists of five different grades with 0 being no regression and 4 being a patient with no tumor cells. There are therefore four thresholds between the different grades where errors due to subjectivity can occur. Suggestions to regroup the Dworak TRG are therefore made, where it was suggested to regroup the Dworak TRG into the bad responders (TRG 0-2) and good responders (TRG 3-4) [82, 83, 84].

We started our prediction of the regrouped Dworak TRG using the radiomics features using state-of-the-art classification methods. We noticed that either both training and validation were performing poorly or that they were overfitting. Most of the validation sets did not even reach 0.50 accuracy. The need for a more tailored and customized prediction method was therefore quite clear.

In a first step, we created a ML algorithm that combines methods from Leo Breiman and Tin Kam Ho, both pioneers in the field of RFs. We used the idea of Tin Kam Ho of using subspaces to create our forest and did pruning like proposed by Breiman. Additionally, we added more details to the algorithm like changing the standard decision trees for LDA trees and performing K-fold validation within the construction of a tree. All together, we constructed a unique algorithm

specifically tailored to tackle the challenges of our data. Far too often, researchers use standard methods that do not perfectly match their problem. Instead, one should look for alternative methods that suit the problem as clearly as possible. Our ERSF is one such method. Our ERSF obtained very good results considering the subjectivity of the regrouped TRG and the instability issues of the radiomic features. Furthermore, we saw that our method did not overfit like some of the state-of-the-art methods did. As the goal is to create a method that can be used in a real hospital, our method must not only work well for the training but also for the validation data. The lack of reporting validation results is yet another issue that we frequently encounter in research papers.

Secondly, we were interested in how severely the ERSF is impacted by Dworak TRGs that are wrongly classified or that cannot be classified without a doubt. We find it crucial to establish a connection between AI and doctors. Nowadays, AI is more popular than ever, but in medicine, it is important that the AI must be explainable by humans. This means that we need to be able to explain why an algorithm makes certain decisions. We were interested to see if the patients who are difficult to grade by the pathologist also tend to be the patients who are difficult to grade by the algorithm. This comparison between AI and medicine gives a unique insight of how the model works and is not often done. We did indeed find that the more difficult to grade patients by the pathologist, i.e. the patients for whom no clear Dworak TRG can be given, are much more likely to be misclassified. We also found that when adding a regularization parameter, which can be seen as a parameter that can be altered by the doctor to specify how peculiar or special a patient is, we can increase the number of correctly classified patients greatly. However, we still noticed that for the patients with an unclear TRG, the regularization helped less than for the other patients. We therefore concluded that these difficult patients, the algorithm and the pathologist do not always share the same opinion. As these patients have no clear TRG, it is perfectly possible for two opinions to differ from each other. In this case, it all depends on what information is more important for the one determining the grade.

As mentioned earlier, the radiomics are also not the perfect tool to use as we have no idea how reproducible and repeatable the features are. As a third step, we were therefore interested in whether we could find new features that could be used instead of the radiomics. As the radiomic features there is a feature class that gives texture features, we had the idea to look at Fourier features. Fourier transforms data in the spatial domain to the frequency domain. In this frequency domain, we might find useful information. However, we were not able to just take the Fourier transform of the data and use them as features as the images and the ROIs of the different patients have different sizes. The different sized images would lead to a different number of features for each patient. We therefore constructed once again a custom-built algorithm, with zero-padding and frequency smoothing, to extract Fourier features. The Fourier features did not provide superb results. We also noticed that our images are not evenly spaced, meaning that for instance one patient can have double the pixel space setting in comparison with another. We therefore looked at different pixel spacing settings to see whether the setting had an influence on the prediction accuracy and whether there were many differences between the chosen settings. We noticed that some settings indeed performed better than others. Most of them gave better results than the Fourier features did without rescaling the pixels. We even had one setting that delivered superb results with more than 0.7 prediction accuracy for the validation. All together, we saw that it was very important to look at the pixel settings and find an optimal setting for the prediction. This rescaling of the pixels is nearly never done and more, usually no information as to whether or not the pixels are evenly scaled is given. It is crucial to look at all the information in the data and try to overcome all the challenges and problems that it presents.

Given the two distinct feature sets we used to predict the Dworak TRG, we considered combining all the feature information beneficial. However, we encountered a significant challenge:

when considered separately, the number of radiomic features and Fourier features was substantial relative to the number of patients. Consequently, merging both feature sets was not feasible. Nevertheless, the LDA has the option to add prior information to the classification method. In the previous research we always assumed the prior information to be equal over all different prediction groups. We made this assumption as we had no information about the prior. Now, with the two different data set, we could first make a ERSF using one data set and then incorporate for each patient the prior information extracted from one data set into the other. If in this case the first ERSF with the Fourier features was extremely confident that a patient was a bad responder, the radiomics needed to be extremely certain that the bad responder was incorrect for the opinion to change. On the other hand, if the first ERSF was not very confident about its decision, the second one could give more information and therefore more clarity to whether or not the decision was made correctly. The results showed that adding prior information to the radiomics ERSF indeed improved the results. We believe that adding new information through iterative building of Forests will give unique additional information for better classification and will overcome the problem of high dimensionality.

With this fourth and last step, we concluded our research where we created a complicated yet quite intuitive process where we have some unique more hidden layers, like normalizing data, extracting Fourier features, rescaling the pixel. On top of these more hidden preprocessing layer we have a unique custom-built ERSF algorithm specifically tailored to tackle all our data problems.

Limitations and Future Work

In this thesis, we laid the foundations of the ERSF for rectal cancer patients. As a next step, the model could be employed by hospitals to assist in clinical decision-making. Simultaneously, we could continue refining the model by consistently evaluating its performance and improving it as needed.

While significant progress has been made and promising results have been achieved, there remains considerable room for improvement in the coming years. There are many more issues with the data set left that need to be tackled before the algorithm can be used in the hospital with confidence. Additionally, it's crucial to acknowledge certain shortcomings of our research that need to be addressed before employing the model.

Given that we're working with a retrospective data set, we lack control over the quality of the radiomics. Some studies have indicated poor reproducibility and repeatability in radiomics. Furthermore, the quality of radiomics is dependent on the scanner model and settings used. Therefore, it's advisable to assess the quality and maintain records of these CT scanner models and settings, as well as the Python version used for data extraction. Moreover, efforts should be directed toward establishing a more standardized approach to data acquisition, ensuring greater stability in radiomic features. Standardization could involve implementing uniform rules and guidelines to obtain well-documented images and more consistent image settings.

Settings play a crucial role in extracting quantitative data. Pixel spacing and slice thickness settings have been demonstrated to affect prediction accuracy. Future research could explore optimal pixel spacing and slice thickness settings with more data. Additionally, exploring alternative interpolation techniques instead of our current method could be beneficial. It's also important to evaluate other settings, as they may vary among patients.

The Dworak regression grade, despite its frequent and confident usage by physicians, presents challenges in mathematical prediction models due to its variability. This variability inevitably impacts the prediction accuracy of the model. Unfortunately, there is no straightforward so-

lution to correct for this variability. While ideally, a new TRG that is purely quantitatively graded would resolve this issue, it's unrealistic to expect such a solution to be attainable. An alternative approach could involve seeking second or even third opinions from other pathologists or having a pathologist grade the same patient twice to limit variability across different assessments. However, it's important to recognize that pathologists are already overloaded with work and do not have the capacity for additional tasks. Additionally, obtaining multiple pathologist opinions would likely incur additional costs, which would ultimately be passed on to the patient. In conclusion, as previously mentioned, there is currently no straightforward solution available for this problem.

Incorporating additional data sources such as blood values and treatment type information is essential. While it wasn't feasible to include blood values in our retrospective data set due to extensive missing data, future data sets should monitor blood values more rigorously. Establishing guidelines for the timing of blood value measurements is crucial, given the potential fluctuations in these values over time. Adding treatment-type information is also necessary for predicting treatment outcomes accurately and optimizing personalized treatment strategies. However, a larger patient cohort is required to integrate treatment information into the model effectively.

Finally, we recognize that ML algorithms are often considered black-box algorithms and that it's crucial for their methods to be more explainable for physicians to confidently utilize them. Therefore, it will be necessary to further strengthen the connection between ML and medicine in the future. Our previous findings demonstrated that the ERSF struggles with gray-zone patients as much as physicians do. However, we can also focus on enhancing the explainability of radiomics. This ongoing research is being conducted within the BISI and TROP research groups. It's essential to emphasize that explainable AI is vital for physicians, and as such, we must allocate more resources to research in this area.

Conclusion

During our research, we successfully created a custom-built ML algorithm tailored to predict the Dworak TRG for rectal cancer patients using quantitative features extracted from the planning CT images. Here, we addressed subjectivity and variability in both the radiomic features and the Dworak TRG while creating our ERSF algorithm. Our method integrates ideas from pioneers in the field, enhances explainability, and robustly handles the challenges presented by our data.

We demonstrated that standard classification methods were insufficient for our data's needs, leading to the creation of our ERSF algorithm, which avoided overfitting and achieved promising results despite the data's subjectivity and instability. Additionally, we looked into Fourier features and pixel spacing standardization, identifying good settings that significantly improved prediction accuracy.

The algorithm's performance could be further improved by iteratively incorporating the prior information. By combining multiple feature sets iteratively, we showed that the classification accuracy could be enhanced without complicating the high dimensionality even more than it already was.

Even though our research shows significant progress in the prediction of rectal cancer, we are aware that further research and optimization of the algorithm are necessary before implementing it in the clinic.

While our work shows significant progress, we are aware that further research and optimization are necessary before clinical application. Future efforts should focus on addressing data quality issues, optimizing settings, incorporating additional data, and improving the explainability of the model for clinical use.

In conclusion, our research shows a novel and promising approach to predicting the Dworak TRG, laying the groundwork for future advancements in personalized cancer treatment and the integration of AI in medical practice. Moreover, in the future, the study could be expanded to more types of cancer.

Bibliography

[1] World Health Organization (WHO). (2022, Feb.) cancer. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] International Agency for Research on Cancer (IARC). World factsheet. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf

[3] International Agency for Research on Cancer (IARC). Colorectum factsheet. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/cancers/41-colorectum-fact-sheet.pdf

[4] World Health Organization (WHO). Cancer today. [Online]. Available: https://gco.iarc.fr/today/en

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis et al., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, Nov. 2015.

[6] W. L. Bi, A. Hosny, M. B. Schabath, and et al., "Artificial intelligence in cancer imaging: Clinical challenges and applications," CA Cancer J Clin, vol. 69, no. 2, pp. 127–157, Mar. 2019.

[7] S. Reddy, "Explainability and artificial intelligence in medicine," Lancet Digit Health, vol. 4, no. 4, pp. e214–e215, Apr. 2022.

[8] J. L. V. Berisha. (2022) AI in Medicine Is Overhyped. [Online]. Available: https://www.scientificamerican.com/article/ai-in-medicine-is-overhyped

[9] B. Jang, Y. J. Lim, C. Song, and et al., "Image-based deep learning model for predicting pathological response in rectal cancer using post-chemoradiotherapy magnetic resonance imaging," Radiotherapy and Oncology, vol. 161, pp. 183–190, Jun. 2021.

[10] H. Zhang, L. Li, K. Qiao, L. Wang et al., "Image prediction for limited-angle tomography via deep learning with convolutional neural network," ArXiv, Jul 2016.

[11] R. Morita, S. Ando, D. Fujita, S. Ishikawa et al., "Quantification of pediatric brain development with x-ray ct images using 3d-cnn," in 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems, 2022, pp. 1–3.

[12] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 4, pp. 1139–1155, Apr. 2009.

[13] M. Kunaver and J. Tasic, "Image feature extraction - an overview," in EUROCON 2005 - The International Conference on "Computer as a Tool", 2005, pp. 183–186.

[14] P. Lambin, E. Rios Velazquez, R. Leijenaar, S. Carvalho et al., "Radiomics: Extracting more information from medical images using advanced feature analysis," European journal of cancer, vol. 48, pp. 441–6, Mar. 2012.

[15] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings et al., "Radiomics: the bridge between medical imaging and personalized medicine," Nat Rev Clin Oncol, vol. 14, no. 12, Dec. 2017.

[16] T. Olson, Applied Fourier Analysis: From Signal Processing to Medical Imaging, 1st ed. New York (USA): Birkhäuser, 2010.

[17] K. Basu, R. Sinha, A. Ong, and T. Basu, "Artificial intelligence: How is it changing medical sciences and its future?" Indian J Dermatol, vol. 65, no. 5, p. 365–370, Sep. 2020.

[18] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood et al., "Machine Learning (ML) in Medicine: Review, Applications, and Challenges," Mathematics, vol. 9, no. 22, Nov. 2021.

[19] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi et al., "edoctor: machine learning and the future of medicine," J Internal Med., vol. 284, no. 6, pp. 603–619, Dec. 2018.

[20] S. Daley. (2024, Feb.) 46 artificial intelligence examples shaking up business across industries. [Online]. Available: https://builtin.com/artificial-intelligence/examples-ai-in-industry

[21] A. Pan. (2022, Dec. 15) A Gentle Introduction to Machine Learning Models. [Online]. Available: https://wandb.ai/wandb_fc/gentle-intros/reports/A-Gentle-Introduction-to-Machine-Learning-Models--VmlldzoyOTUxNjQw

[22] E. Avuçlu, "A new data augmentation method to use in machine learning algorithms using statistical measurements," Measurement, vol. 180, Aug. 2021.

[23] H. Chen, Z. Lin, H. Wu, L. Wang et al., "Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest," Spectrochimica acta Part A, Molecular and biomolecular spectroscopy, vol. 135, pp. 185–191, Jan. 2015.

[24] P. P. Ypsilantis, M. Siddique, H. M. Sohn, A. Davies et al., "Predicting response to neoadjuvant chemotherapy with pet imaging using convolutional neural networks," PLoS One, vol. 10, no. 9, Sep. 2015.

[25] B. He, T. Ji, H. Zhang, Y. Zhu et al., "Mri-based radiomics signature for tumor grading of rectal carcinoma using random forest model," J Cell Physiol, vol. 234, no. 11, pp. 20 501–20 509, Nov. 2019.

[26] C. Liang, Y. Huang, L. He, X. Chen et al., "The development and validation of a ct-based radiomics signature for the preoperative discrimination of stage i-ii and stage iii-iv colorectal cancer," Oncotarget, vol. 7, no. 21, Apr. 2016.

[27] C. Raets, C. E. Aisati, M. D. Ridder, A. Sermeus et al., "An Evolutionary Random Forest to measure the Dworak tumor regression grade applied to colorectal cancer," Measurement, vol. 205, pp. 112–131, Nov. 2022.

[28] C. Raets, C. El Aisati, A. L. Rifi, M. De Ridder et al., "Bridging the gap between machine learning and medicine: A critical evaluation of the dworak regression grade in rectal cancer," IEEE Open Journal of Instrumentation and Measurement, vol. 3, pp. 1–12, 2024.

[29] C. Raets, C. E. Aisati, A. L. Rifi, M. De Ridder et al., "Optimizing Rectal Cancer Patient Care: Dworak TRG Prediction via Bayesian Evolutionary Fourier-Domain Random Subspace Forest," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1–10, 2024.

[30] International Agency for Research on Cancer (IARC). Rectum factsheet. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/cancers/9-rectum-fact-sheet.pdf

[31] International Agency for Research on Cancer (IARC). Anus factsheet. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/cancers/10-anus-fact-sheet.pdf

[32] World Health Organization (WHO). Cancer tomorrow. [Online]. Available: https://gco.iarc.fr/tomorrow/en

[33] P. Rawla, T. Sunkara, and A. Barsouk, "Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors," Gastroenterology Review, vol. 14, no. 2, p. 89–103, Jan. 2019.

[34] S. L. Gearhart and N. Ahuja, Early Diagnosis and Treatment of Cancer Series: Colorectal Cancer. Saunders Elsevier, 2010.

[35] J. E. Niederhuber, J. O. Armitage, M. B. Kastan, J. H. Doroshow et al., Abeloff's Clinical Oncology, 6th ed. Philadelphia: Elsevier, 2020.

[36] Vrije Universiteit Brussel (VUB), "Klinische en maatschappelijke topics I," [Onsite Course], 2024.

[37] American Cancer Society. (2024, Jan.) Tests to diagnose and stage colorectal cancer. [Online]. Available: https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/how-diagnosed.html

[38] National Cancer Institute. (2024, Oct.) Screening tests to detect colorectal cancer and polyps. [Online]. Available: https://www.cancer.gov/types/colorectal/screening-fact-sheet

[39] U. for International Cancer Control (UICC), TNM Classification of Malignant Tumours, 8th ed., J. D. Brierley, Ed. Hoboken: Wiley, 2020.

[40] E. Kapiteijn, C. A. Marijnen, I. D. Nagtegaal, H. Putter et al., "Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer," New England Journal of Medicine, vol. 345, no. 9, pp. 638–646, Aug. 2001.

[41] Swedish Rectal Cancer Trial, "Improved survival with preoperative radiotherapy in resectable rectal cancer," New England Journal of Medicine, vol. 336, no. 14, pp. 980–987, Apr. 1997.

[42] R. Sauer, H. Becker, W. Hohenberger, C. Rödel et al., "Preoperative versus postoperative chemoradiotherapy for rectal cancer," New England Journal of Medicine, vol. 351, no. 17, pp. 1731–1740, Oct. 2004.

[43] O. Dworak, L. Keilholz, and A. Hoffmann, "Pathological features of rectal cancer after preoperative radiochemotherapy," Int J Colorectal Dis, vol. 12, no. 1, pp. 19–23, Mar. 1997.

[44] C. S. Denlinger and B. A. M., "The challenges of colorectal cancer survivorship," J Natl Compr Canc Netw, vol. 7, no. 8, p. 883–93, Sep. 2009.

[45] T. J. Ridolfi, N. Berger, and K. A. Ludwig, "Low anterior resection syndrome: Current management and future directions," Clin Colon Rectal Surg, vol. 29, no. 3, pp. 239–45, Sep. 2016.

[46] Y. Ziv, A. Zbar, Y. Bar-Shavit, and I. Igov, "Low anterior resection syndrome (LARS): cause and effect and reconstructive considerations," Tech Coloproctol, vol. 17, no. 2, pp. 151–162, Oct. 2013.

[47] National Guideline Alliance, Optimal management of low anterior resection syndrome: Colorectal cancer (update): Evidence review E2. London: National Institute for Health and Care Excellence (NICE), Jan. 2020.

[48] Surveillance Research Program, National Cancer Institute (SEER). (2023, Apr. 19) SEER*Explorer: An interactive website for SEER cancer statistics. [Online]. Available: https://seer.cancer.gov/statistics-network/explorer

[49] Steimel, Josh, "Instrumentation and Experimentation," Pacific Open Texts. 13., 2020. [Online]. Available: https://scholarlycommons.pacific.edu/open-textbooks/13

[50] R. S. Elias M. Stein, Fourier Analysis: An Introduction 1. Princeton University Press, 2003.

[51] R. Bracewell, The Fourier Transform & Its Applications, 3rd ed. McGraw-Hill, 1999.

[52] P. V. O'Neil, Advanced Engineering Mathematics, 7th ed. Stamford (USA): Cengage Learning, 2012.

[53] P. Suetens, Fundamentals of Medical Imaging, 2nd ed. Cambridge University Press, 2009.

[54] A. Nokhostin and S. Rashidi, "Covid-19 diagnosis by extracting new features from lung ct images using fractional fourier transform," Fractal and Fractional, vol. 8, no. 4, p. 237, Apr. 2024.

[55] Y. B. Luo, J. H. Cai, P. L. Qin, R. Chai et al., "Ffs-net: Fourier-based segmentation of colon cancer glands using frequency and spatial edge interaction," Expert Systems With Applications, vol. 262, Okt. 2024.

[56] T. Yoshimasu, M. Kawago, Y. Hirai, T. Ohashi et al., "Fast fourier transform analysis of pulmonary nodules on computed tomography images from patients with lung cancer," Annals of Thoracic and Cardiovascular Surgery, vol. 21, no. 1, pp. 1–7, Feb. 2015.

[57] F. Nüsslin, "Wilhelm conrad röntgen: The scientist and his discovery," Phys Med, vol. 79, pp. 65–68, Nov. 2020.

[58] The Editors of Encyclopaedia. (2023, Mar. 20) cathode-ray tube. [Online]. Available: https://www.britannica.com/technology/cathode-ray-tube

[59] T. Buzug, Computed Tomography: From Photon Statistics to Modern Cone-Beam CT. Heidelberg: Springer Berlin, 2008.

[60] E. Seeram, Computed Tomography: Physical Principles, Clinical Applications, and Quality Control, 4th ed. Missouri (US): Elsevier, 2009.

[61] P. N. T. Wells, "Sir Godfrey Newbold Hounsfield KT CBE. 28 August 1919 – 12 August 2004," Biogr. Mems Fell. R. Soc., vol. 51, pp. 221–235, Dec. 2005.

[62] T. Feeman, The Mathematics of Medical Imaging: A Beginner's Guide. New York: Springer, 2009. [Online]. Available: https://books.google.be/books?id=8uQwQv2wSBMC

[63] CT Imaging: Practical Physics, Artifacts, and Pitfalls.

[64] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny et al., "Computational radiomics system to decode the radiographic phenotype," Cancer Res, vol. 77, no. 21, pp. e104–e107, Nov. 2017.

[65] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," ACM SIGGRAPH Computer Graphics, vol. 21, no. 4, p. 163–169, aug 1987.

[66] G. Ajmera. (2024, Feb. 15) Feature Extraction of Images using GLCM (Gray Level Cooccurrence Matrix). [Online]. Available: https://medium.com/@girishajmera/feature-extraction-of-images-using-glcm-gray-level-cooccurrence-matrix-e4bda8729498

[67] MathWorks®. Create a Gray-Level Co-Occurrence Matrix. [Online]. Available: https://nl.mathworks.com/help/images/create-a-gray-level-co-occurrence-matrix.html

[68] H. Yu, X. Meng, H. Chen, J. Liu et al., "Predicting the level of tumor-infiltrating lymphocytes in patients with breast cancer: Usefulness of mammographic radiomics features," Front Oncol, vol. 11, Mar. 2021.

[69] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "An ensemble learning ap- proach for brain cancer detection exploiting radiomic features," Comput Methods Programs Biomed, vol. 185, pp. 105–134, Mar. 2020.

[70] B. A. Varghese, D. Hwang, S. Y. Cen, J. Levy et al., "Reliability of ct-based texture features: Phantom study," journal of Applied Clinical Medical Physics, vol. 20, no. 8, pp. 155–163, Jun. 2019.

[71] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," Phys Med Biol, vol. 61, no. 13, pp. R150–66, Jun. 2016.

[72] M. Chiara, D. Marfisi, A. Barucci, J. Del Meglio et al., "Collinearity and dimensionality reduction in radiomics: Effect of preprocessing parameters in hypertrophic cardiomyopathy magnetic resonance t1 and t2 mapping," Bioengineering, vol. 10, no. 1, Jan. 2023.

[73] R. T. H. M. Larue, G. Defraene, D. De Ruysscher, P. Lambin et al., "Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures," Br J Radiol, vol. 90, no. 1070, Feb. 2017.

[74] S. G. Mougiakakou, I. K. Valavanis, A. Nikita, and K. S. Nikita, "Differential diagnosis of ct focal liver lesions using texture features, feature selection and ensemble driven classifiers," Artificial Intelligence in Medicine, vol. 41, no. 1, pp. 25–37, May. 2007.

[75] R. Canellas, K. S. Burk, A. Parakh, and D. V. Sahani, "Prediction of pancreatic neuroendocrine tumor grade based on ct features and texture analysis," Amarican Journal of Roentgenology, vol. 210, no. 2, pp. 341–346, Feb. 2018.

[76] Y. Zeng, C. Xiang, and G. Wu, "Muscle ct radiomics is feasible in the identification of gout," Current Medical Imaging, vol. 20, Aug. 2024.

[77] H. Tin Kam, "Random decision forests," in Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278–282.

[78] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Chapman & Hall, 1984.

[79] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[80] R. A. Fisher, "the use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, no. 2, pp. 179–188, Sep. 1936.

[81] S. C. Hollensead, W. B. Lockwood, and R. J. Elin, "Errors in pathology and laboratory medicine: Consequences and prevention," Journal of Surgical Oncology, vol. 88, no. 3, pp. 161–181, Nov. 2004.

[82] L. Reggiani Bonetti, S. Lionti, F. Domati, and V. Barresi, "Do pathological variables have prognostic significance in rectal adenocarcinoma treated with neoadjuvant chemoradiotherapy and surgery?" World J Gastroenterol, vol. 23, no. 8, pp. 1412–1423, Feb. 2017.

[83] M. D. Santos, C. Silva, A. Rocha, E. Matos et al., "Prognostic value of mandard and dworak tumor regression grading in rectal cancer: study of a single tertiary center," ISRN Surg, vol. 2014, Mar. 2014.

[84] M. R. Siddiqui, J. Bhoday, N. J. Battersby, M. Chand et al., "Defining response to radiotherapy in rectal cancer using magnetic resonance imaging and histopathological scales," World J Gastroenterol, vol. 22, no. 37, pp. 8414–8434, Oct. 2016.

[85] J. C. Peeken, M. Bernhofer, B. Wiestler, T. Goldberg et al., "Radiomics in radiooncology - challenging the medical physicist," Phys Med, vol. 48, pp. 27–36, Apr. 2018.

[86] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," Knowledge and Information Systems, vol. 53, no. 3, pp. 551–577, May. 2017.

[87] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015, pp. 1200–1205.

[88] C. Cortes and V. Vapnik, "Support vector networks," Machine Learning, vol. 20, pp. 273–297, Sep. 1995.

[89] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 2nd ed. New York: Springer, 2021.

[90] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of Machine Learning. Cambridge: The MIT Press, 2012.

[91] C. Shalizi. (2024, Feb. 23) Advanced data analysis from an elementary point of view. Unpublished book. [Online]. Available: https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/

[92] B. D. Ripley, Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press, 1996.

[93] J. H. Friedman, "Regularized discriminant analysis," Journal of the American Statistical Association, vol. 84, no. 405, pp. 165–175, Oct. 1989.

[94] A. Gelman, J. Carlin, H. Stern, D. Dunson et al., Bayesian Data Analysis, Third Edition. Taylor & Francis, 2013.

[95] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[96] H. Tin Kam, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[97] T. Giannakopoulos and A. Pikrakis, Audio Classification, 1st ed. USA: Academic Press, Inc., 2014.

[98] N. Deo, Graph Theory with Applications to Engineering and Computer Science (Prentice Hall Series in Automatic Computation). Hoboken: Prentice-Hall, Inc., 1974.

[99] A. L. Rifi, I. Dufait, C. E. Aisati, M. D. Ridder et al., "Unraveling the biological meaning of radiomic features," in 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2022, pp. 1–6.

[100] K. Basu, R. Sinha, A. Ong, and T. Basu, "Artificial intelligence: How is it changing medical sciences and its future?" Indian J Dermatol, vol. 65, no. 5, pp. 365–370, Sep 2020.

[101] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood et al., "Machine learning (ml) in medicine: Review, applications, and challenges," Mathematics, vol. 9, no. 22, p. 2970, Nov. 2021.

[102] S. Daley. (2023) 46 ai in healthcare examples improving the future of medicine. [Online]. Available: https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare

[103] Regard. (2023) Regard. [Online]. Available: https://withregard.com/about

[104] Buoy Health, Inc. (2024) Buoy health. [Online]. Available: https://www.buoyhealth.com

[105] A. Mkhadri, "Shrinkage parameter for modified linear discriminant analysis," INRIA, Report, Nov. 1992. [Online]. Available: https://inria.hal.science/inria-00077033

[106] S. H. Kim, H. J. Chang, D. Y. Kim, J. W. Park et al., "What is the ideal tumor regression grading system in rectal cancer patients after preoperative chemoradiotherapy?" Cancer Research and Treatment, vol. 48, no. 2, pp. 998–1009, Oct. 2015.

[107] L. Tan and J. Jiang, Digital Signal Processing: Fundamentals and Applications, 2nd ed. Elsevier, 2013.

[108] K. Kourou, T. Exarchos, K. Exarchos, M. Karamouzis et al., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, Nov. 2014.