UNIVERSIDADE D
COIMBRA

Heloisa Barbosa da Silva

Magnetic Resonance Imaging Biomarkers in Liver Metastases

"Machine Learning" Applications

VOLUME 1

Dissertation submitted to the Physics Department of the Faculty of Science and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialization in Biomedical Instrumentation, supervised by Doctor Henrique Alexandrino and PhD João Valente Duarte

September 2022

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Heloisa Barbosa da Silva

Magnetic Resonance Imaging Biomarkers in Liver Metastases

"Machine Learning" Applications

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Dr João Duarte (Advisor)
Dr Henrique Alexandrino (Advisor)

Faculty of Medicine, University of Coimbra, Coimbra, Portugal
Faculty of Sciences and Technology, University of Coimbra, Coimbra, Portugal
Institute of Nuclear Sciences Applied to Health (ICNAS), University of Coimbra, Coimbra, Portugal

Coimbra, 2022

This work was developped in collaboration with:

Coimbra Hospital and Universitary Centre

Coimbra Institute for Biomedical Imaging and Translational Research

Resumo

O terceiro tipo de cancro mais comumente diagnosticado no mundo, e que ocupa o segundo lugar em mortes relacionadas com o cancro, é o cancro colorretal (CRC). O fígado é o órgão mais comum para a disseminação das metástases desse tipo de cancro. A ressecção hepática é atualmente o tratamento mais efetivo para os doentes com CRC que apresentam metástases hepáticas (MHCRC). No entanto nem todos os doentes estão aptos para serem submetidos a este procedimento. O ideal seria escolher um tratamento personalizado baseado na biologia do doente.

Diversos estudos demonstraram que o padrão de crescimento histológico (PCH) são relevantes na escolha de tratamentos. O PCH é a margem de tecido entre as metástases e o parênquima de fígado não-tumoral. Esse padrão pode ser classificado em três tipos: desmoplástico (dPCH), infiltrativo (rPCH) ou expansivo (pPCH). Contudo, apesar da sua importância, estes só podem ser determinados após a análise histológica da peça removida cirurgicamente. A ressonância magnética (RM) é o exame de escolha na avaliação pré-operatória dos doentes com MHCRC, uma vez que possibilita estudar de forma detalhada a interface tumor-fígado. O trabalho desta tese tem como objetivo desenvolver uma aplicação de inteligência artificial para prever os PCH de forma não invasiva, a partir de imagens adquiridas por RM.

Foram analisadas imagens adquiridas por RM com a abordagem de radiómica, que permite extrair informações biológicas a partir de uma imagem, e testaram-se diversas técnicas de machine learning com o intuito de criar um modelo capaz de classificar os tipos de PCH a partir da informação extraída das imagens. Entre as categorias de abordagens de machine learning, foram consideradas duas abordagens de classificação. A primeira abordagem, chamada de classificador multiclasse, considerou três classes a serem classificadas: PCH desmoplástico, infiltrativo e expansivo. Já a segunda abordagem, de classificação binária, considerou um modelo com apenas duas classes: PCH desmoplástico ou não desmoplástico (incluindo infiltrativo e expansivo).

As informações extraídas das imagens (*features*) foram selecionadas testando dois métodos diferentes. Criou-se um modelo de classificação para cada fase de

aquisição de imagem no protocolo clínico de avaliação de MHCRC e avaliou-se a capacidade preditiva de cada modelo utilizando as métricas de revocação, precisão, *f1-score*, curva ROC e área sob a curva (AUC), e ainda a acurácia. As probabilidades de uma lesão pertencer a uma classe em particular foram extraídas dos modelos com melhor desempenho para cada fase, sendo consideradas como novas *features* num modelo final que também foi avaliado utilizando as métricas mencionadas.

Os melhores resultados foram obtidos com o classificador binário. O *f1-score* para a previsão do padrão desmoplástico foi de 0.84 para os dados de treino e 0.80 para os dados de teste. Os dados de teste também apresentaram uma AUC de 0.83. A acurácia foi de 0.85 para o treino e de 0.82 para o teste.

Palavras-Chave: cancro colorretal, padrões de crescimento histológico, biomarcadores de imagem, ressonância magnética, radiómica, machine learning, multiclasses, classe binária

Abstract

The third most commonly diagnosed cancer worldwide and the second leading cause of cancer-related deaths is colorectal cancer (CRC). The liver is the organ most commonly affected by the spread of metastases from this type of cancer. While liver resection is currently the most effective treatment for patients with CRC, that have liver metastases (CRCLM), however not all patients are able to undergo this procedure. Therefore, it would be ideal to choose a personalised treatment based on the patient's biology.

Several studies have shown that the knowledge of type of the histological growth pattern (HGP)is important in the choice of treatments. The HGP is the tissue margin between the metastases and the non-tumour liver parenchyma. These patterns can be divided into three types: desmoplastic (dHGP), replacement (rHGP) or pushing (pHGP). However, despite their importance, it can only be determined after histological analysis of the surgically removed piece. Magnetic resonance imaging (MRI) is the exam of choice in the preoperative evaluation of patients with CRCLM, as it allows detailed examination of the tumour-liver interface. The aim of this work is to develop an artificial intelligence application for non-invasive prediction of HGPs from MR images.

Images acquired with MRI were analysed using a radiomics approach, which allows biological information to be extracted from an image. Different machine learning techniques were tested to build a model capable of classifying the types of HGP based on the textural information extracted from the images. Among the categories of machine learning approaches, two classification approaches were considered. The first approach, called the multiclass classifier, considered three classes to be classified: desmoplastic HGP, replacement and pushing. The second approach, the binary classifier, considered a model with only two classes: desmoplastic or non-desmoplastic HGPs (including replacement and pushing).

The image information (features) to be included in the prediction models was selected by testing two different methods. A classification model was created for each phase of image acquisition in the CRCLM clinical assessment protocol and the

predictive ability of each model was assessed using the metrics of recall, precision, f1-score, accuracy, ROC curve and area under the curve (AUC). The probabilities that a lesion belongs to a particular class were extracted from the best performing models for each phase. These probabilities were included as new features in a final model, which was also evaluated using the above metrics.

The best results were obtained with the binary classifier. The f1-score for predicting the desmoplastic pattern was 0.84 for the training data and 0.80 for the test data. The test data also showed an AUC of 0.83. The accuracy was 0.85 for the training data and 0.82 for the test data, showing that some models were efficient in classifying the desmoplastic histological growth pattern.

Keywords: colorectal cancer, histological growth patterns, imaging biomarkers, magnetic resonance imaging, radiomics, machine learning, multiclass, binary class

Acknowledgments

First of all, I would like to thank my supervisors Dr. Henrique Alexandrino and Dr. João Duarte for trusting me and giving me the opportunity to participate in this incredible project, for always being available to answer my questions, for their suggestions and revisions of this thesis, and for their constant support and motivation during the development of this work.

I would also like to thank Dr. Miguel Castelo-Branco for his suggestions and corrections. I thank him, my supervisors and the Portuguese Foundation for Science and Technology (FCT) for the financial support through the UID/4959/2020 grant, which enabled me to continue my studies. I also thank the FCTUC for providing the merit scholarship and award for international students at the University of Coimbra.

I would like to thank Dr. Carolina Terra, Dr. Rui Caetano and Dr. Filipe Caseiro Alves for taking the time to answer my questions about magnetic resonance imaging and the identification of histological growth patterns, and for their advice.

I am grateful to professors João Campos Gil, Maria Carmen Alpoim and Liliana Ferreira for their support and motivation during my first years at this institution.

I am very grateful to my friends for the good company during these years, especially Rita Monteiro for her great support, friendship and partnership in my work and studies.

I would like to thank Luis Felipe Coelho for his constant support, his trust in me, his love, his companionship, for being by my side in all good and bad times, and for being a great inspiration.

Above all, I would like to thank my parents Rosângela Aparecida and José Tomaz, and my sister Júlia Barbosa. Without you, without your unconditional love and support, I would not have come this far.

This research work was funded by the Portuguese Foundation for Science and Technology (FCT) : UID/4959/2020

Contents

List of Tables

List of Figures

1

Introduction

Throughout the world, cancer is an alarming disease due to its high mortality rate. This disease results from genetic mutations that cause changes in the signalling pathways that control cell cycles, allowing uncontrolled cell division and growth. The main cause for the high mortality rate in cancer is the invasive behaviour of cancer cells, which allows the disease to progress and metastasise. Nearly 90% of cancer deaths are due to metastasis. Malignant cells, derived from the primary tumour, infiltrate into the surrounding parenchyma and, through the intravasation of blood vessels, reach the circulation [1–5].

Approximately 10% of all cancers diagnosed worldwide and cancer-related deaths are due to colorectal cancer (CRC). Each year, 1.2 million new cases of CRC are discovered worldwide, and about half of these cases will have metastases [6, 7].

The liver is frequently affected by metastases. In Europe, secondary liver tumours are much more common than primary ones. The venous drainage of the colon and rectum via the portal vein to the liver may be one reason why this organ is the main region for the occurrence of metastases in case of CRC [8, 9].

Liver resection is currently the most effective treatment for patients with colorectal cancer liver metastases (CRCLM). However, not all patients can undergo this procedure, and liver recurrence in CRCLM is extremely common in the first two years after surgery, so not all patients have long-term survival. In addition, this approach is associated with a risk of perioperative morbidity and mortality, which can significantly affect the patients' quality of life. Another strategy that has been widely accepted is the combination of surgery and neoajuvant chemotherapy (NAC). It is important to emphasise that these treatments are not only aggressive, but also require a clear, personalised medical approach, with the treatment of CRCLM being carried out by a multidisciplinary team [7, 10, 11].

1.1 Contextualisation

A thorough understanding of tumour biology and clinical biomarkers is extremely important for the multidisciplinary team to choose an appropriate strategy for action. Histological growth patterns (HGPs) were shown to be a practical prognostic factor with predictive value. These patterns involve inter-and intra-lesion heterogeneities of epigenetic, genetic, morphologic, and phenotypic properties, conducting to variances in overall survival of patients with CRLM in relation to the treatments performed. These growth patterns may be of desmoplastic or replacement type, which are the most common, or the rarer types such as pushing, sinusoidal and portal [8, 10, 12–14].

The prognostic role of HGPs has been observed in several studies in which the desmoplastic pattern, characterised by the presence of a thick band of stroma, rich in blood vessels and lymphocytes, between the non-tumour and the tumour parenchyma, is a predictor of favourable overall survival. On the other hand, the replacement growth pattern, in which the tumour component spreads in a poorly defined manner and invades the non-tumour liver cells, is associated with poorer overall survival, which also occurs in others non-desmoplastic patterns [11].

Histological growth patterns also play an important role in more targeted treatment selection. An example of this is the identification of patients who will benefit from a liver transplant. In this procedure, one of the necessary prerequisites is the liver-only disease, a feature more commonly present in patients with desmoplastic patterns. Immunomodulatory therapies are also recommended for patients with this type of pattern, while more aggressive perioperative chemotherapy may be indicated for non-desmoplastic patterns. In addition, HGPs have also been shown to play a prognostic role in non-colorectal cancers such as breast cancer, gastric cancer and uveal melanoma, helping to clarify the failure of certain treatments. All of this underscores the importance of predicting histopathological patterns in order to spare patients unnecessary suffering that does not lead to positive outcomes [11, 15].

Nowadays, these pathological details of liver metastases are only known after resection, which is an invasive and risky procedure for the patient. Fortunately, however, there is a very attractive approach that allows to extract biological information from images: radiomics. Radiomics is a powerful technique that allows the analysis of textural features in medical images and enables effective classification of tumours at the morphological and molecular levels. It has the advantage of being non-invasive and is also time and cost effective [10, 11, 13, 16].

1.2 Motivation

The introduction of a technique that allows prognostic biomarkers to be predicted in a non-invasive way from images obtained from patients using routine methods, such as magnetic resonance imaging, will save patients inconvenience and possible complications. Moreover, the fact that these biomarkers have the potential to help making decisions about the most effective and personalised treatments for each patient is of paramount importance in a disease where mortality is high and time is often short.

1.3 Goals

This project aims to combine the texture analysis technique, radiomics, with machine learning algorithms to predict the histological growth pattern of liver metastases from images acquired using magnetic resonance imaging technology and establish a biomarker of prognostic with clinical value. The specific goals of this work are *i)* to determine which textural and non-textural features are most prevalent in this type of study using radiomics, as many are not visible to the naked eye and *ii)* to determine which of these features better predict the nature of the histological growth pattern using machine learning algorithms.

1.4 Structure

Chapter 2 of this thesis will cover the topics that will support this work. An overview of the biology of colorectal cancer and liver metastases is given, followed by the prevalence in the world, as well as risk factors, diagnosis and treatment. Radiomics and machine learning methods are also presented in this chapter. Finally, the last section of this chapter presents an overview of the application of radiomics and machine learning in the context of predicting HGPs in colorectal cancer liver metastases.

Chapter 3 is devoted to explaining the methods used to attempt to solve the proposed problem. The chapter covers the criteria for the selection of patients for the compilation of the database, the pathological characterisation, the acquisition of images of the liver metastases as well as the selection of the regions of interest (ROI), the description of the processing of these images, the feature selection methods and the model generation.

In Chapter 4, a multiclass classification problem was considered in which the model was trained and tested to predict the classes for the desmoplastic pattern,

pushing and replacement. The results obtained were presented and discussed in sections according to the methods used. The conclusions on this classifier can also be found in this chapter.

In Chapter 5 the results for a binary classifier were presented and discussed, where only two classes were considered, the desmoplastic and the non-desmoplastic class. The conclusions for the results obtained with this classifier were also included.

Finally, in chapter 6, the final conclusions are drawn and perspectives for future work are presented .

2

Background Concepts And State Of The Art

This chapter contains the information that supports the work developed in this thesis. It covers concepts about colorectal cancer as well as liver metastases, histological growth patterns and their types, the prevalence of the disease, risk factors, diagnosis and current treatments, the importance of growth patterns in selecting treatments, concepts related to the radiomics approach and a summary of articles on the use of this technique in CRCLM.

In the gastrointestinal system, the colon and the rectum form the large intestine. The colon consists of four sections (ascending colon in the proximal part, transverse colon, descending colon and sigmoid colon, the last two in the distal part) and is mainly responsible for the re-absorption of water from the intestinal contents after they have passed through the jejunum and ileum. The final wastes eventually enters the rectum, which serves as a faecal reservoir, and are finally excreted through the anus [17]. Of all CRCs, 41% occur in the proximal part, with 22% involving the distal part and 28% the rectum [18].

Generally, the progress of CRC follows characteristic patterns. It begins with microscopic lesions at the crypt (crypt lesions), which are epithelial cell-based and become small polyps over time (a neoplastic precursor lesion). As the polyps grow, the epithelial cells that make them up have a growing number of mutations associated with cancer genes and gradually show a dysplastic phenotype [6, 17, 19].

A focus of carcinoma in situ formed by some malignant cells confined to the polyp epithelium has the potential to invade and metastasise [17, 19]. Malignant cells cross the basement membrane, invade the intestinal wall, then the lymphatics, and finally enter the bloodstream through vessels and metastasise to various organs such as the liver [17].

2.1 Prevalence

CRC is the third most common type of cancer diagnosed worldwide, in both women and men. This type of gastrointestinal cancer is a major health concern as it is the fourth leading cause of cancer death in the world, causing about 900 000 deaths per year [6, 17, 20].

Incidence and mortality rates are about 25% lower in women than in men. Developed countries have been shown to be the regions where these rates are higher. More than two thirds of all cases of CRC and about 60% of all deaths related to this cancer occur in countries with a high Human Development Index (HDI) [6, 21].

Of all newly diagnosed cancers in Europe, about 12.9% are CRC. Between 2000 and 2005, the number of deaths increased by 3% if only Portugal is considered [22]. By 2030, the incidence of this cancer is expected to increase by 60% to 2.2 million new cases and 1.1 million deaths worldwide [21].

Although tracking and health care programs have recently evolved, most patients already have metastatic disease at the time of diagnosis. Approximately 14-18% of patients with CRC have metastases at the first clinical presentation, and approximately 10-25% have metastases when the primary CRC is resected. Within 5 years of diagnosis, 20% to 50% of patients with CRC, usually die from metastatic disease. If there is no treatment, patients with CRCLM have an average survival of only 5 to 20 months [8, 23, 24].

2.2 Risk factors

The probability of developing CRC is equal to 4%-5% and risk factors include lifestyle, environmental factors, heredity, and age [25].

Age is the main risk factor and one of the factors that cannot be changed. The risk of developing CRC increases significantly after the age of 50 [25].

Another unmodifiable factor is genetic predisposition. Polyposis and non-polyposis syndromes (Lynch syndrome) are related to CRC by heredity [6]. This factor accounts for about 35% of the risk factors. Families with a positive history account for 10-20% of the total number of patients with CRC [6, 20, 25].

In patients with inflammatory diseases associated with the large intestine, the risk of developing CRC increases by 2.5% in people with Crohn's disease, and 3.7% in ulcerative colitis [25].

Smoking, poor diet, lack of exercise and alcoholism are some of the risk factors

related to lifestyle, but unlike the above factors, these can be changed.

In people who have a sedentary lifestyle with unhealthy eating habits, there is an association with obesity, which is a very relevant risk factor. Due to the increase in levels of visceral adipose tissue (VAT), there is an active hormonal component of total body fat that promotes the release of pro-inflammatory cytokines, which in turn can contribute to the occurrence of CRC. In this context, the probability of developing CRC increases to 70% [25].

In the case of alcoholism, depending on the polymorphism of the enzymes that metabolise alcohol, the acetaldehyde contained in the drink is considered carcinogenic, which increases the likelihood of developing CRC. As for smoking, due to the high content of carcinogens in its components, the risk of developing CRC increases by up to 10.8%, as these components can easily enter the intestine and cause polyps [25].

The authors of a study conducted in 1,099 patients with CRC over a 7-year period suggest that in univariate analyses, characteristics such as age, depth of tumour invasion, perineural and vascular invasion, and lymph node metastasis are among the possible risk factors for the development of metachronous liver metastases (MLM) [26].

2.3 Diagnosis

Although it is a slowly progressive cancer (it takes about 10 to 15 years for the precursor lesion to develop into CRC) [6, 19], patients may present with various symptoms such as anemia, abdominal pain, and rectal bleeding. However, this type of cancer is usually not noticeable until it reaches an advanced stage [19].

The method of choice for diagnosis of CRC is colonoscopy, but as described above, many patients with CRC develop liver metastases. Although computed tomography (CT) is widely used for the detection of liver metastases, magnetic resonance imaging (MRI) is the gold standard for diagnosis [6, 27].

MRI allows radiologists to noninvasively diagnose focal liver lesions, hepatic vascular disease, biliary and diffuse liver disease by careful selection of MRI parameters. The human body has an abundance of tissues rich in hydrogen (H1), such as water, fat, and proteins, which is why proton or H1 magnetic resonance imaging is mainly used in clinical settings. The patient is positioned inside a scanner with a strong magnetic field B_0, expressed in Tesla (T), generally in the cranial-caudal direction. The protons then align their magnetic field either parallel or antiparallel to it. The preferred state for alignment is the one that requires the least energy,

so that more protons align in the low-energy state parallel to B_0, while a minority align in a higher energy state, which is antiparallel to B_0. Moreover, protons also have a precession motion, which is like spinning around its own axis. This motion is quite fast, and the number of precession spins of the proton per second is called the "precession frequency". This value depends on the strength of the applied magnetic field B_0 and can be calculated by the Larmor equation [28, 29].

To obtain an image, it is necessary to apply a radiofrequency (RF) pulse at the Larmor frequency perpendicular to B_0 in the region to be evaluated. When the RF is absorbed by the tissue, some protons will change their energy state and the number of protons in the antiparallel state with higher energy increases, causing a net longitudinal magnetisation further away from the B_0 direction. At the same time, the protons will precess in phase, causing a new horizontal magnetisation. After RF is turned off, the protons return to their normal low energy state, i.e., they return to the original longitudinal magnetisation direction align with the B_0 field, and the interactions between protons and the surrounding tissue will also cause a decay of the horizontal magnetisation. The energy released by the protons during this process is electronically detected, allowing the formation of the image. The RF signals, emitted and received, provide the information needed to build the image and depend on two tissue properties, called T1 and T2 relaxation times. The T1 characterises the relaxation as the longitudinal magnetisation of protons realigns with B_0. This signal is tissue specific and depends on the surrounding structure or lattice. The T2 component corresponds to the decay signal of the horizontal magnetisation, since phase coherence is lost after RF excitation. Both T1 and T2 are significantly impaired under pathological conditions [28].

The main parameters that allow the images to be generated when they are changed are the echo time (TE), the repetition time (TR) and the flip angle. By choosing these parameters correctly, as well as the RF pulses, it is possible to modify the resulting signal and obtain weighted images, i.e. T1-weighted (T1W) images, where the T1 signal is emphasised, or T2-weighted (T2W) images, where the T2 signal is emphasised. These images are part of the series that make up liver MRI protocols [28].

Compared to other cross-sectional imaging techniques, MRI provides much better resolution for soft tissue. Furthermore, the excellence of this technique is demonstrated when it comes to detecting small-sized metastases, with the sensitivity of contrast-enhanced MRI (CE-MRI) being approximately 91%-97% [27].

Hepatobiliary contrast-enhanced magnetic resonance imaging agents such as gadolinium ethoxybenzyl diethylenetriamine pentaacetic acid (Gd- EOB-DTPA,

Primovist® in Europe or Eovist® in the United States) has been shown to be a safe and sensitive method for detecting liver metastases. This is due to the superiority of the contrast between the lesion and the liver created by the uptake of gadolinium at the base of the liver parenchyma. Accurate mapping of the number of colorectal cancer metastases and their location is of paramount importance for preoperative planning by surgeons and enables these professionals to more appropriately counsel patients regarding the required surgery. In addition, adequate information about metastases can also prevent patients from undergoing unnecessary surgery [30].

2.4 Colorectal liver metastasis

The liver is one of the organs where colorectal cancer metastases most commonly occur, with 80% of all CRC metastases occurring in this organ [15]. This is because the liver's blood supply originates from the junction of the gastrointestinal tract blood vessels via the hepatic portal vein, and cancer cells migrate into the liver parenchyma in this bloodstream to form the liver metastasis [31].

Some studies have shown that about 25% to 30% of patients with CRC develop liver metastases during the progression of the disease [32]. Metastases are still the most common cause of death in patients with solid tumours [15]. Patients with CRC have a worse prognosis if metastases are present in the liver [33].

Vidal-Vanaclocha proposed in 2011 that the progression of liver metastases is composed of four phases. The first is called the microvascular phase. It occurs in sinusoidal vessels as soon as the circulating tumour cells arrive. At this point, the cancer cells must be able to attach to the endothelial cell layer and subsequently transmigrate through the vascular endothelium (extravasation). The second phase comprises the pre-angiogenic or intra-lobular micrometastatic phase and is characterised by the recruitment of immune cells and stromal cells from the liver. At this stage, the micrometastases are not yet vascularised. In the third phase, called the angiogenic phase or pan-lobular phase, the recruitment of endothelial cells and the formation of blood vessels is induced by the hypoxic microenvironment. In the last phase, called the lobar growth phase, the new tumour can already be clinically detected. The interactions between the hepatic microenvironment and the malignant cells are responsible for the subsequent growth of the metastases as well as for the progression of the four phases mentioned [16].

2.5 Histological Growth Patterns

Histopathologic growth patterns (HGPs) represent the various interactions between micro-environmental cells and metastatic tumour cells, and provide a unique interface between the surrounding liver and the tumour [15].

Considering the differences in the interface between the adjacent liver parenchyma and tumour cells, the CRLM exhibits several types of patterns due to the multiple biological processes that occur in this region, such as angiogenesis, interaction with the cells' immune system, paracrine and autocrine effects of growth factors, which may ultimately have implications in the choice of the most appropriate surgical therapy. These patterns may be pushing patterns (pHGP), sinusoidal and portal patterns, or the most common and main types: Desmoplastic (dHGP) and the Replacement (rHGP) [11, 13, 14].

2.5.1 Desmoplastic histologic growth pattern (dHGP)

In the desmoplastic pattern (also named encapsulated), in addition to a dense lymphocytic infiltration, there is an arc of fibrous tissue isolating the metastases of the liver parenchyma (Figures 2.1 and 2.2). This type of pattern shows a more favourable diagnosis because the thick wall of collagen-enriched stroma provides a barrier that prevents the tumour from spreading. Sprouting angiogenesis form new blood vessels, and a reaction that is comparable to the healing of a wound is caused by the cancer cells, inflammation, and the new blood vessels produce scar tissue [8, 14, 15].

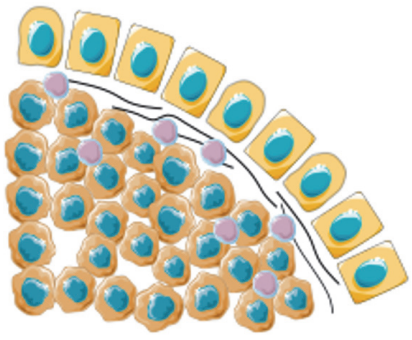


Figure 2.1: "Schematic representation of the desmoplastic pattern, the tumour is separated from the liver parenchyma by a band of fibrous tissue, which contains tumour infiltrating lymphocytes". Adapted from [8].

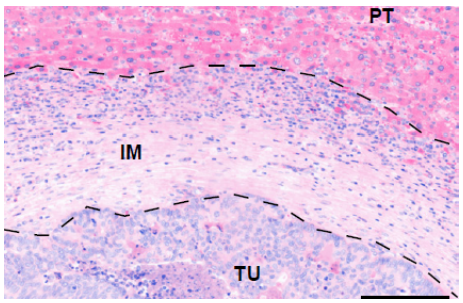


Figure 2.2: "Representative image of desmoplastic HGP in colorectal liver metastases identified on H& E-stained tissue sections. **PT**: peritumor, **TU**: tumour regions, **IM**: invasive margin. It is possible visualise a rim of fibrotic tissue that encapsulates the metastasis (IM region)". Adapted from [14].

The lymphocytic infiltration along with integrin blockade and the increase in collagen type IV allow for a decreased infiltration of non-tumour parenchyma, leading to a more favourable prognosis. Furthermore, in a recent study, a higher R0 resection rate (when no tumour remains after surgical treatment) was observed in CRCLMs with dHGP, suggesting that there may be protection against positive marginal resection in this pattern due to the fibrous stroma being rich in inflammatory cells, which is not seen in the replacement or pushing patterns [8, 11, 34].

It has also been found that patients who have mixed patterns in certain regions (dHGP in combination with non-dHGP) have a less favourable prognosis than those who have only a dHGP pattern at the interface between liver and tumour[11, 16].

2.5.2 Pushing histologic growth pattern (pHGP)

The pushing pattern is present when the surrounding liver tissue is compressed by cancer cells, but without imitating the architecture of the healthy organ (Figures 2.3 and 2.4). This pattern is described as angiogenic type and is characterised by a hypoxic environment, an aggressive factor, and resistance to treatments [8, 14].

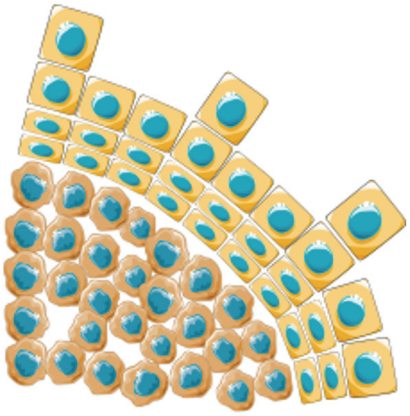


Figure 2.3: "Schematic representation of the pushing pattern, the tumour the tumour expands and compresses the surrounding hepatocytes." Adapted from [8].

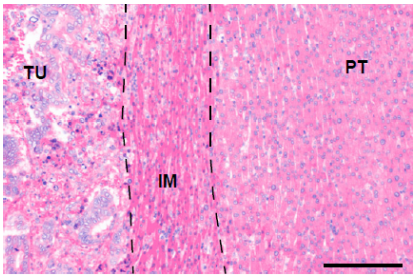


Figure 2.4: "Representative image of pushing HGP in colorectal liver metastases identified on H& E-stained tissue sections. **PT**: peritumor, **TU**: tumour regions, **IM**: invasive margin. It is possible visualise the liver tissue compressed and pushed away by the tumour". Adapted from [14]

2.5.3 Replacement histologic growth pattern (rHGP)

In the rHGP, the cancer cells replace the hepatocytes and co-opt the sinusoidal blood vessels at the tumour-liver interface (because plates formed by the cancer cells are in continuity with the liver cell plates), without disturbing the stroma architecture of the liver or inducing sprouting angiogenesis [15], as shown in Figures 2.5 and 2.6 .

Since the rHGP liver metastases do not present secondary structures at the tumour-liver interface, such as glandular structures, they are poorly differentiated. An immune desert is represented by this type of metastasis, because of the low infiltration of inflammatory and/or immune cell types. There is no fibrous stroma in this type of pattern, as observed in dHGP, so the patient is at increased risk of hepatic recurrence because of the risk of marginal resection [11, 15].

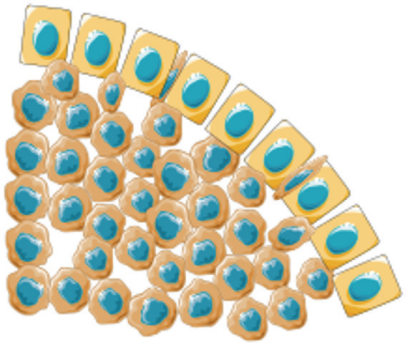


Figure 2.5: "Schematic representation of the replacement pattern, the tumour permeates between the liver hepatocytes, without disruption of the normal architecture;". Adapted from [8].

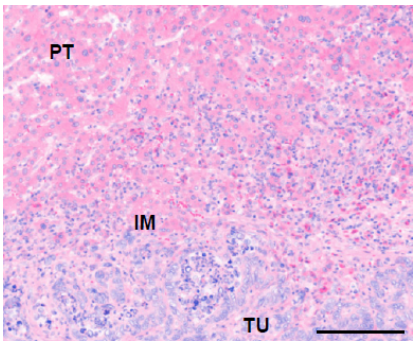


Figure 2.6: "Representative image of replacement HGP in colorectal liver metastases identified on H& E-stained tissue sections. **PT**: peritumor, **TU**: tumour regions, **IM**: invasive margin. It is possible visualise the tumour cells infiltrating the surrounding liver parenchyma (IM region)". Adapted from [14].

In a recent report, rHGP was found to be more prevalent in the new-onset liver lesions during systemic treatment and to have a worse prognosis, having poorer progression-free survival (PFS) and overall survival [15].

2.5.4 Sinusoidal and Portal Patterns

The presence of growing cancer cells within the septa of the liver or within the connective tissue space of the liver capsule or portal tract is responsible for defining the portal type. In the sinusoidal pattern, the cells grow around peri-sinusoidal spaces or sinusoids [14].

2.6 Treatment

Currently available treatments for CRC include endoscopic or surgical resection, immunotherapy, preoperative downstaging radiotherapy, and systemic therapy. For liver metastases, successful treatment requires a multidisciplinary team composed

of oncologists, radiation oncologists, colorectal surgeons, hepatobiliary surgeons, among others, as the choice of treatment depends on the presentation, amount, and location of these metastases, as well as the possibility of surgical resection, which is the best option as it offers a survival rate of 38% to 51% in 5 years [6, 35, 36].

The technique of liver resection is quite risky and must be performed by experienced surgical teams at high volume centers. In this treatment approach, it is extremely important to determine the amount of parenchyma to be removed. Preserving as much of the liver as possible is of utmost importance, as many patients undergo chemotherapy, which carries a potential for hepatotoxicity, and are therefore at high risk of liver failure. Although considered the "gold standard" for CRCLM, only 10%-15% of patients can undergo partial hepatectomy. This is due to the presence of lesions with an unfavourable anatomic location, insufficient future residual liver volume to resect all lesions, poor general health, and the presence of significant extrahepatic disease [37, 38].

Although the overall survival rate in 5 years is 47-60% after hepatectomy for CRCLM, recurrence occurs in approximately 40-75% of patients, and of these, 50% have recurrent liver involvement [36]. Compared with liver resection, ablative therapy is considered minimally invasive and relatively safe in all its approaches: either percutaneous, which is the least invasive and usually does not require deep general anesthesia; during laparoscopic surgery, which requires experienced surgeons and general anesthesia; or during open surgery, usually combined with resection and which is the most invasive. In the literature on ablative therapy, the morbidity rate is reported to be between 4% and 9%, while the mortality rate is between 0% and 2%. However, the most common complications associated with this technique are postoperative bleeding, infectious complications such as liver abscesses, liver failure, and cardiopulmonary complications, among other [39].

2.6.1 The importance of HGPs in selecting treatments

As we can see in the previous section (2.6), the treatments for CRCLM are quite aggressive, and therefore individualised treatment tailored to the patient would be ideal. Histopathologic growth patterns have already been shown to be relevant prognostic factors (see subsections 2.5.1, 2.5.2 and 2.5.3).

The properties of HGPs also offer great potential for the selection of personalised treatments, taking their different immune phenotypes as an example. Tumours with low T-cell infiltration have generally been shown to have greater resistance to immune checkpoint inhibitors. The desmoplastic growth pattern, which shows dense infiltration of lymphocytes as well as high expression of genes related to immunity, is

considered a valuable biomarker for the use of immunomodulatory therapy [11, 15].

In cases where partial hepatectomy is not possible as treatment for CRCLM and liver transplantation is being considered, patients have a satisfactory response to systemic chemotherapy, the primary tumour must be removed with liver-only disease, the latter being a common feature in patients with desmoplastic-type HGP. Even in nanotherapeutics, HGPs may have an impact, considering the development of a method with great potential of nanocarriers that allow the drug to remain at the desired tumour site and mitigate the side effects of conventional agents. Patients who have a dHGP pattern will have a barrier to the administration of these drugs, again showing the importance of knowing HGPs before treatment selection [11].

Predicting HGP before treatment may also improve the risk profile for recurrence. The desmoplastic pattern is associated with improved recurrence free survival, when compared to other types. Thus, patients with non-desmoplastic patterns may benefit from thermal ablation techniques as they have an increased risk of aggressive and early recurrence, sparing them to futile and potentially risky surgical procedures. In patients with this pattern, more aggressive perioperative chemotherapy may also lead to positive outcomes [11].

HGPs also play a prognostic role in liver metastases from non-colorectal cancers, such as gastric cancer, breast cancer and uveal melanoma. In a study of patients with uveal melanoma liver metastases, Barnhill et al. [40] showed that - similar to colorectal cancer and breast cancer - patients with a replacement pattern had a lower survival rate than patients with a desmoplastic pattern. In breast cancer, anti-angiogenic therapy failed in clinical trials because the predominant pattern of HGP was found to be of the replacement type, and as observed in the CRCLM data, this type of lesion does not respond effectively to this therapy. Therefore, the importance of prior prediction of histopathological patterns is reiterated [11, 15].

However, the HGPs have the disadvantage of being known only after surgical resection of the liver metastasis. Since this tumour is very heterogeneous, liver biopsy, which also carries a risk of complications, is not a viable option [15].

Jones et al.[41] have demonstrated that preoperative biopsies have a negative impact on survival after liver resection and that it may not be justifiable in patients with potentially resectable disease. In New Zealand, a study of 43 patients who underwent preoperative biopsy of CRCLM found that seven of them had extrahepatic dissemination. The researchers associated this situation with a image-guided biopsy or biopsy at the time of laparotomy for bowel resection [41, 42]. Therefore, a non-invasive method for detecting HGPs is needed.

2.7 Radiomics Approach

The explosion of "Big Data" ushered in the new era of artificial intelligence (AI) algorithms - intelligence proven by machines - which is being widely applied in a variety of fields, including medicine, especially radiology. However, AI is unlikely to replace human experience in this field, as converting human experience into consistent and adequate computer models is a major challenge, especially when human knowledge is not complete. Nevertheless, there has recently been a great deal of interest in an approach known as the "science of discovery", which focuses on the exploration of large amounts of data to discover patterns that enable the formulation of new hypotheses. Such approaches are purely data-driven. One of the most popular is called radiomics or texture analysis [43].

Radiomics is a technique for extracting high-dimensional data from medical radiographic images. Medical images contain a large amount of data (in the form of grey level patterns) with information about disease-specific processes that cannot be detected by the human eye and conventional visual inspection. It is very promising and is widely used in the field of oncology thanks to initial support from the *Quantitative Imaging Biomarker Alliance* and *National Cancer Institute*. In 1973, some researchers attempted to classify images based on texture features. Decades later, in 1995, there were signs that it would be possible to train computer algorithms to identify medical images when researchers used a convolutional neural network to identify lung nodules. Then, in the late 2000s, researchers tried to determine the relationship between tumour pictures and their genome types. However, these studies were conducted with a small dataset, meaning that the radiomic models created at the time could not be validated by external organisations, as they were based on small datasets from individual organisations. However, with the improvements and innovations in the field of medical imaging, radiomics in oncology has rapidly evolved and was first proposed by Philippe Lambin in 2012. This technique allows the identification of quantitative imaging biomarkers (features) to determine, for example, prognostic evaluation, response to therapy and survival in several tumours [10, 44–48].

Radiomics goes beyond the visual interpretation of images, as it allows us to detect variations in texture, shape, or intensity. Moreover, the goal of this technique is to convert images into data with high throughput and accuracy. Radiomics can be performed with different modalities of medical imaging, such as magnetic resonance imaging (MRI), positron emission tomography (PET) and computed tomography (CT) [46, 49]. In 2014, radiomics was used in the field of oncology to study features from CT images for diagnostic and predictive purposes. In 2016, it was found for the

first time that the radiomics technique applied to CT images would allow the prediction of lymph node metastases in patients with CRC. Used in MRI, the radiomics technique allows, for example, the detection of micro-structural changes in the liver parenchyma. In a study with animals suffering from CRLM, micrometastases could be detected with this technique even before histopathological evidence [8, 44, 50].

In summary, this method quantifies the image's texture information by mathematically extracting the spatial distribution of signal intensities and the relationships between pixels using AI analysis techniques. Examples of the steps that can be followed in the radiomics process are: Image acquisition, segmentation, image processing, feature extraction, feature selection and finally model building [46, 48, 49]. This workflow can be seen in Figure 2.7.

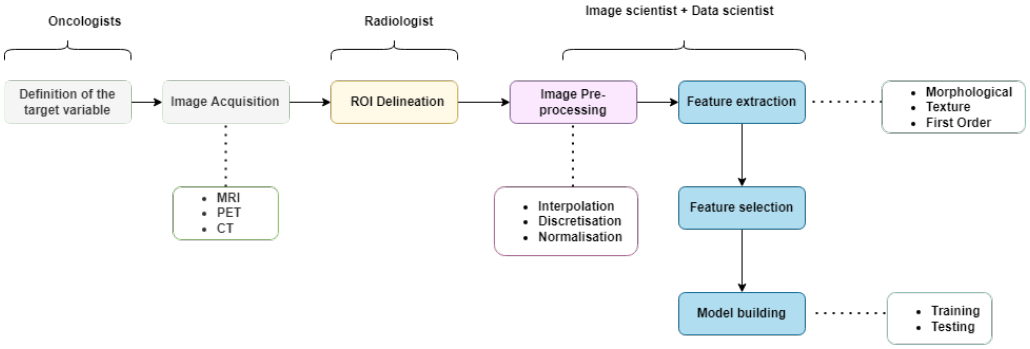


Figure 2.7: Example of a radiomics workflow for an oncological problem.

2.7.1 Radiomics framework

As mentioned earlier, radiomics technique can be used in various types of medical imaging such as PET, CT and MRI. However, since this project is based on images acquired using MRI, the description of the technique will focus on this imaging modality.

2.7.1.1 Image acquisition

Soft tissues can be characterised thanks to the functional information and high structural contrast provided by MRI technology. Tissue properties, such as acquisition parameters and relaxation times, form a complex interplay that produces the signal intensities in this type of images [49, 51].

In MRI examination, both common 1.5T and 3T scanners are very suitable for imaging the liver, although the examination may take some time. The exam consists of pre-contrast and post-contrast sequences. T1-weighted (T1W) images, which are part of the pre-contrast phase, refer to image series that have a low signal for

water molecules, i.e. dark areas. Images of a normal liver should show a consistent T1 signal, or isointense with the paraspinal muscles and slightly hyperintense in relation to the spleen. On T1W images, evidence of abnormal signal from the liver parenchyma is an indicator of pathology. Liver masses (whether malignant or not) are often seen as a low-intensity signal on these images. T1 images can also be in phase if the signals from fat and water add up, or out of phase if the signals from these substances go in opposite directions and cancel each other out [28, 29].

Even in the pre-contrast phase, T2-weighted (T2W) images are very useful to detect lesions because of the high contrast and low dynamic range. Liver cysts, biliary hamartomas, abscesses, and hemangiomas are hyperintense. Hepatic solid masses are typically hyperintense in these organs, but less intense than cysts [28, 29]. Infiltrating tumor cells and vasogenic edema are detected by the hyperintensity of the T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) sequence [51]. Diffusion-weighted imaging (DWI) and dynamic contrast-enhanced MRI (DCE) are also part of the protocol for magnetic resonance imaging of the liver. Based on local differences in the movement of water molecules, DWI allows the characterisation of possible changes in the cellular structure of the tissue [28, 29, 49, 51]. It has been hypothesised that this technique, namely by providing apparent water diffusion coefficient (ADC) maps, may indicate cell death following therapy [51]. Solid masses and tumors have low ADC values, while tumor necrosis due to treatments has high ADC values [28].

In DCE imaging, the gadolinium contrast agent is administered intravenously, and its passage through the vessels and tissues is observed over time. This allows visualisation of the distinction between normal tissue and lesions as capillary permeability and vascularization increase. Therefore, this technique allows the extraction of volume fraction, permeability and vascular flow. The likelihood that lesions are malignant can also be determined by the kinetic study of enhancement, which assesses the time course of signal intensity within the lesion [29, 49, 51].

The phases of a dynamic image consist of [28, 29]:

(I) The pre-contrast phase, which helps detect hemorrhage.

(II) The late arterial phase (AP), which occurs 15-30 seconds after injection, which is useful for detecting hypervascular lesions.

(III) The portal venous phase (PVP), which occurs between 45 and 75 seconds after contrast injection; at this time, the portal veins are completely opacified and the liver parenchyma enhances homogeneously from the hepatic arteries.

(IV) The delayed phase, which occurs 2 to 5 minutes after contrast injection and is slowly excreted by the kidneys, causing the liver to begin to show reduced

enhancement.

2.7.1.2 Segmentation of regions of interest

The segmentation of regions of interest is a crucial step in the radiomics approach. In this step, the region in which the features are to be computed must be defined. In two dimensions, the region of interest (ROI) is delineated, and in three dimensions, the volume of interest (VOI) is delineated. Image segmentation can be performed manually, semi-automatically or fully automatically, with the first two methods being the most widely used [46].

Manual segmentation must be performed by an experienced radiologist, or radiation oncologist, and has the disadvantage of being time consuming and varying depending on the number of datasets and images to be segmented. Semi-automatic segmentation can be performed using standard image segmentation algorithms and is usually performed with manual correction. Both segmentations involve significant observer bias, and according to some studies, several radiomic features are not free from inter- and intra-observer variations in the considered delineations [46, 51].

Fully automated segmentation can be performed using Deep Learning algorithms. This is the best option for image segmentation as it avoids variability in inter- and intra- observations. Currently, several algorithms have been trained for segmenting different organs, and platforms such as 3D Slicer and MITK have several integration options for these algorithms. However, the generalizability of the trained algorithms is still quite limited and their application to a diverse dataset leads to failure. Therefore, further studies are needed to improve these algorithms [46].

Intensities in images are arranged uniformly in intervals or spaces. These regular positions are called pixels in 2D images and voxels in 3D images. Segmentation promotes the creation of a ROI mask, assigning a value of 1 to each voxel belonging to the ROI, while assigning a value of 0 outside the ROI [52].

2.7.1.3 Image processing

Image processing is the step aimed at improving the quality and homogeneity of the image from which the radiomic features are extracted. Images acquired by MRI, for example, have Gaussian and Rician noise. Moreover, the intensities are not homogeneously distributed over the image, so corrections are needed [52, 53].

Most textural features require interpolation for isotropic voxel spacing to allow comparisons between different samples, batches or cohorts, and image data. They also need interpolation to become rotationally invariant. After applying the interpolation algorithms, the ROIs or VOIs masks must also be interpolated. Since MRIs

might not provide isotropic data, different interpolation approaches are required [52, 53].

Pixels/voxels that fall outside a certain range of grey levels can be removed by outlier filtering and resegmentation. In the case of MRI, it is necessary to use the first method because this type of data has units of arbitrary intensity, which makes a range resegmentation impossible. The most commonly used method is to calculate the average (μ) and standard deviation (σ) of the grey levels that are inside the ROI and exclude the grey levels that are outside the range given by: $\mu \pm 3\sigma$ [46].

Finally, it is necessary to discretise the range of image intensities that lies within the ROI/VOI. This step is important to reduce the noise. It also makes the textural features tractable and improves the computational efficiency. Discretisation essentially has the function of grouping the voxel intensities into uniformly distributed intervals or bins, which is very similar to creating a histogram [46, 51, 52]. Discretisation can be performed in two ways:

1. Fixing the bin size in units of voxel intensity [51].
2. Fixing the number of bins, e.g., 32 or 64. Some authors recommend fixing 64 bins to achieve a high degree of robustness and reproducibility [51].

2.7.1.4 Feature extraction

The final processing step is called feature extraction. In this step, the radiomic features that describe the tumour's texture patterns (e.g. heterogeneous or not), its shape, its relationship with the surrounding tissue, its location, and the properties of the tumour intensity histogram (e.g. high or low contrast) can be calculated [46, 49].

According to some authors, features are classified as being of first-order, second-order or higher-order. First-order features (or global statistics) are considered as features of the histogram-based type. They represent the distribution of voxel intensities that lie within ROIs, but without taking into account the spatial interactions between them. Second-order (or textural) features represent the spatial distribution of intensity levels in the voxels. The grey level scale allows visualisation of the spatial variation in the intensity (texture) levels. These features can be used to determine intratumoral heterogeneity. Higher order features aim to emphasise a particular feature by applying filters to the image to obtain patterns that may or may not be repeated [51, 54].

Since there are several formulas for calculating these features [46], the guidelines of the Image Biomarker Standardisation Initiative (IBSI), which allows to standardise these calculations [52], were followed in this work. According to the IBSI, the

classification of the features is divided into six categories:

1. **Morphological** - Features that describe the geometric aspects of the region of interest (example: area, volume, etc.) and the shape properties (example: elongation, sphericity, etc.) [52].
2. **Local intensity** - The intensity of local features can be calculated using the voxel intensities that lie within a defined neighbourhood around a central voxel. The local intensity peak and the global intensity peak belongs to this class of features [52].
3. **Intensity-based statistical** - The distribution of intensities within the ROI can be described by these types of features. These features include mean intensity, intensity variance, intensity skewness, 10^{th} intensity percentile, and so on. These features do not need to be discretised [52].
4. **Intensity histogram-based** - The discretisation of the original intensity distribution into intensity bins makes it possible to construct an intensity histogram. Features such as mean discretised intensity, median discretised intensity, discretised intensity skewness, among others, belongs to this category [52].
5. **Intensity volume histogram-based** - The relation between the discretised intensities i and the volume fraction containing the last intensity i can be determined by this category of features, such as the features intensity at volume fraction, volume fraction difference between intensity fractions, and so on [52].
6. **Texture matrix-based** - Although textural features were developed for evaluating surface texture in two-dimensional images, their analysis can be performed on images for both 2D slices and 3D objects. Features in this category are divided into:
    - **Grey level co-occurrence matrix (GLCM)** - This matrix describes how the combinations of grey levels (discretised intensities) of both adjacent pixels and voxels, in the case of a 3D volume, are distributed around one of the directions of the image [52].
    - **Grey level run length matrix (GLRLM)** - The length of a sequence of pixels/voxels having the same grey level around a m direction is defined by this matrix [52].
    - **Grey level size zone matrix (GLSZM)** - This matrix allows counting the number of zones/groups of linked voxels. Voxels are linked only if the adjacent voxel has the same discretised grey level [52].
    - **Neighbourhood grey tone difference matrix (NGTDM)** - Alternative to GLCM defined by Amadasun and King [55].

- **Neighbouring grey level dependence matrix (NGLDM)** - Alternative to the GLCM defined by Sun and Wee [56].
- **Grey level dependence matrix (GLDM)** -This matrix quantifies the dependencies of the grey levels (number of connected voxels within a distance δ that are dependent on the central voxel) in an image [57].

2.7.1.5 Feature selection

It is possible to extract a large number of features from a single image, and depending on the number of filters used in the process, this number may be unlimited. However, such a large number of features in a model with a small number of cases (for example) can lead to overfitting. Overfitting occurs when the model performs very well on the training data but not on the new unseen data, so that generalisation is not possible. It is extremely important to reduce the overfitting to allow the construction of a robust radiomic signature that is generalisable and, furthermore, robust to detect variations between new patients that were not considered in the training model. It should also be noted that not all features are useful, as many are redundant or highly correlated with each other, or are not yet strongly associated with the given classification task, so their removal is a crucial step at this stage, which is called feature selection [46, 51, 53, 58].

Recursive feature elimination (RFE) and minimum redundancy maximum relevance (mRMR) are some of the different feature selection algorithms and have been used in this project (see Section 3). mRMR is an algorithm developed for selecting features from microarray data (which can be used to monitor gene expression of thousands of genes) [59]. It simultaneously selects the features with the highest correlation to a class, which is called relevance, and the features with the lowest correlation to each other, which is called redundancy [60]. For the calculation of relevance, the F-statistic can be used, which is the ratio between two variances (it measures the dispersion of a data set around the mean), and for redundancy, the correlation can be used [61].

RFE is a method that recursively selects features considering an ever decreasing set of features. It uses an estimator that is trained on the initial set of features and determines the importance of each feature. The less important features are then removed from this set. Then this process is repeated, and the feature set is reduced until reaching the minimum number of features that yield the best performance [62].

2.7.1.6 Model building

After selecting the most relevant, stable, and non-redundant features in the feature selection step (2.7.1.5), a model can be developed to solve the clinical problem [53].

Ideally, two different image sets should be used to estimate the performance of a machine learning , with the largest used for training and fine-tuning the model and the smallest (ideally from another institution) used to validate the model and allow for external validation, resulting in more realistic estimates of the model's performance and ensuring the development of radiomic signatures that can be applied in the clinical setting [53].

Most published models are based on retrospective cohorts of patients from a single institution. In this case, internal validation is used. In internal validation, the study sample is divided into a training subset, which is used to develop the training model, and a test subset, which is used to evaluate and validate the model. In cases where the dataset is very small (approximately 50 to 100 patients), internal validation carries a high risk of bias, as a single test set with only a few data instances (20-30 patients) will lead to increased optimistic or pessimistic estimates regarding model performance. The cross-validation approach - in more detail in the section 2 - may be one approach to address this issue, where a small cohort can be split into multiple training and test sets [53].

In radiomics, there are two main learning schemes: supervised learning, where the performance of the model is assessed relative to the ground truth (output data), which is known, and unsupervised learning, where the model uses the input data to be trained to uncover possible correlations or associations, without known output objectives [53]. These concepts are explained in more detail in the next section 2.8.

2.8 Machine learning Approaches

Machine learning is a programmable computational approach that is capable of learning from data (experience). Typically, a model is created that is intended to predict a specific outcome based on a set of features. This outcome (output) can be quantitative, such as the amount of leukocytes in the blood, or qualitative, such as the presence or absence of a particular disease. In radiomics, identifying the best-fitting machine learning models is crucial for stable and clinically relevant radiomic biomarkers [45, 63, 64].

2.8.1 Types of machine learning

It is possible to categorise machine learning systems as 'supervised', 'unsupervised', 'semi-supervised ' and 'reinforcement' depending on the type of supervision they receive during the data training (training phase) [65].

2.8.1.1 Supervised learning

In supervised learning, the algorithm is fed labelled training data, i.e. in addition to the input values of the individual variables (features), the desired solutions (labels) are also inserted. In this way, the algorithm attempts to model the relationships between the variables and the labels [65].

In the supervised learning category, *"Regression"* and *"Classification"* tasks are typical of this type of system.

1. *Regression* - In the Regression task, the aim is to predict a continuous target value, e.g. the amount of glucose in the body [65].
2. *Classification* - The classification task is about learning how to classify new input data into a certain class label. While regression predicts a continuous quantity, classification predicts a nominal output value. To assess how good a classification algorithm is, one can use metrics extracted from the so-called *confusion matrix* [65].

   This matrix is a two-dimensional matrix that summarises the performance of the classifier with respect to a given test dataset [66]. The table 2.1 is a typical representation of a confusion matrix. The rows represent the actual (true) class of an object, while the columns represent the prediction made by the model. In this case there are two classes, the positive represented by the number 1 and the negative represented by the number 0. In a model to be developed, the aim is to obtain of the best possible class assignment to each in the data, depending on the threshold value reached (see Figure 2.8).

Table 2.1: Confusion matrix example.

	Confusion matrix	
	Predicted 1	Predicted 0
Actual 1	TP	FN
Actual 0	FP	TN

An example of a confusion matrix for a three-case multiclass problem is given in table 2.2:

Table 2.2: Example of a confusion matrix for a three-case multiclass problem. On this matrix, the rate of true positives for class 2 is calculated.

	Confusion matrix		
	Predicted 2	Predicted 1	Predicted 0
Actual 2	TP	FN	FN
Actual 1	FP	TN	FN
Actual 0	FP	FN	TN

The confusion matrix parameters are [67]:

- **True positives (TPs)** - True positives occur when the model correctly predicts the positive class, i.e. whether a person has a disease or not, the model says that the disease test is positive, and in reality, the patient has the disease.
- **True negatives (TNs)** - True negatives is when the model correctly predicts the negative class, i.e. in the same case of the portability of a disease, the model says that the person does not have it, and in fact the person does not have it.
- **False positives (FPs)** - In this case, the model incorrectly predicts the positive class. The model says that the person has the disease when in fact the person does not have it.
- **False negatives (FNs)** - False negatives occur when the model incorrectly predicts the negative class, i.e. when the model classifies a person as not having the disease, when indeed the person has it.

A lot of information can be gained with the confusion matrix. One of the most important metrics is *recall*, also known as true-positive rate (TPR) or sensitivity, and is given by the formula:

$$TPR = \frac{TP}{TP + FN} \tag{2.1}$$

This formula gives us the rate of positive cases correctly detected by the classifier [65].

Another important metric is *Specificity (TNR)*, which measures the proportion of people who do not have the disease and have received a negative result. It is given by equation [66]:

$$TNR = \frac{TN}{TN + FP} \tag{2.2}$$

There are other metrics, such as the false positive rate (FPR), i.e. the probability that the classifier gives a positive answer when in fact it is negative, given by the formula:

$$FPR = \frac{FP}{TN + FP} \tag{2.3}$$

It can also be calculated as FPR = 1-Specificity. Picture 2.8 represents the output probability produced by a model that classifies an event into a certain class if the probability is greater than a threshold value. If we shift the threshold to the left, we get more negative values, so we have a lower sensitivity and a higher specificity. If we do the opposite, i.e. if we lower the threshold, we get more values that are classified as positive, so that the sensitivity increases and the specificity decreases. If we increase the sensitivity (TPR), since the FPR is 1-specificity, the number of false positives also increases. There is a tradeoff between these quantities and a good compromise between them is required.

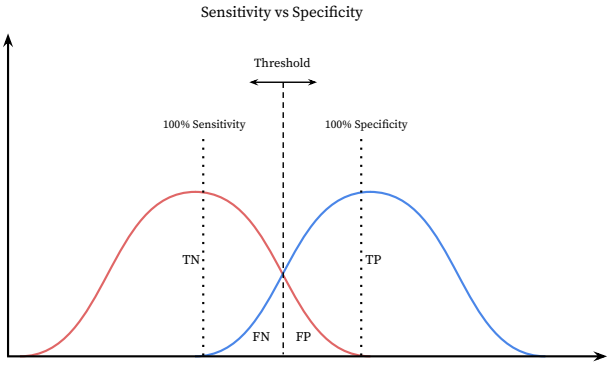


Figure 2.8: Scheme of the decision threshold. If the threshold is shifted to the left, the sensitivity is higher and also the number of false positives. If the threshold is shifted to the right, the specificity is higher and the number of false negatives results also increases.

The false negative rate (FNR), on the other hand, indicates the probability of the classifier giving a negative answer when in fact it is positive. It is described by the following equation:

$$FNR = \frac{FN}{FN + TP} \tag{2.4}$$

The proportion of correctly obtained classifications (true positives and true negatives) can be determined with the expression:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

It is also possible to calculate the precision, which is defined as the ratio between the positive real cases and the total number of positive cases predicted by the model [66]. The formula is as follows:

$$Precision = \frac{TP}{FP + TP} \tag{2.6}$$

In general, precision is used in conjunction with the TPR metric because it is possible for the classifier to achieve trivially perfect precision by ignoring all instances except one (precision=1/1=100%) [65].

The ratio between the actual negative instances and the total number of instances predicted by the model is called the Negative Predictive Value (NPV) and can be calculated as follows:

$$NPV = \frac{TN}{TN + FN} \tag{2.7}$$

The harmonic mean between precision and recall is called the F1-score and is calculated with the following formula:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.8}$$

The receiver operating characteristic curve, or ROC curve, is a plot of the true positive rate versus the false negative rate for different thresholds. The figure 2.9 shows a diagonal line representing the ROC curve of a random classifier. For a classifier to be considered good, it must be as far as possible from this line, i.e. it must be in the upper left corner. One way to check the performance of the classifier, i.e. its ability to distinguish between classes, is to measure the area under the curve (AUC). The higher the AUC, the better the model is at predicting values that belong to the negative class (predicting zeros as zeros) and predicting values that belong to the positive class (predicting ones as ones). Models whose AUC corresponds to the value 1 are the perfect models. Models that have an AUC value of 0.5, on the other hand, are completely random models [66, 68].

An example of a ROC curve in a three-class classifier is shown in picture 2.10. In the picture, class 1 is classified better than class 0, which in turn is better than class 2.

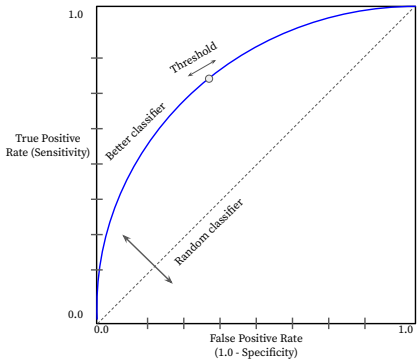


Figure 2.9: Example of a ROC curve. In the figure, the dashed line represents a random model, while the blue line represents a better classifier as it is closer to 1. The ROC curve is the representation of the results obtained by performing a scan on the thresholds values.

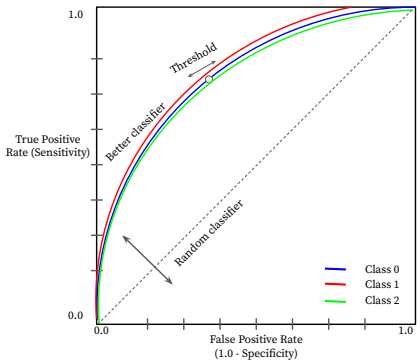


Figure 2.10: Example of a ROC curve for multiclass problem.

An important factor in a predictor is that it must be able to perform well not only for existing data but also for new data, i.e. we want to generalise its performance. Therefore, it is common to divide the database into two sets. One is used for training to calculate the loss function (which is nothing more than the representation of the error between the obtained and expected values), minimise it and optimise the parameters. The other set is used in the testing step to evaluate the performance of the model with new data [65, 69].

However, typically the amount of data may be limited, particularly in clinical settings. Therefore, when using a large amount of data available to train the model, the validation set will be too small, resulting in a very noisy estimate of forecast performance (with high variance). The cross-validation approach may be one approach to address this issue, where a small dataset can be split into multiple training and test sets. *Cross-validation* is a procedure in which the data is divided into K partitions. The model is trained with k-

1 of these partitions and the last partition is used to test the model. We can do the same process for all k possible combinations. A scheme of cross-validation can be seen in the figure 2.11. Then it is possible to calculate the average performance of our model for all repetitions. This procedure makes it possible to use the maximum available data to train the model and estimate the generalised accuracy [69].

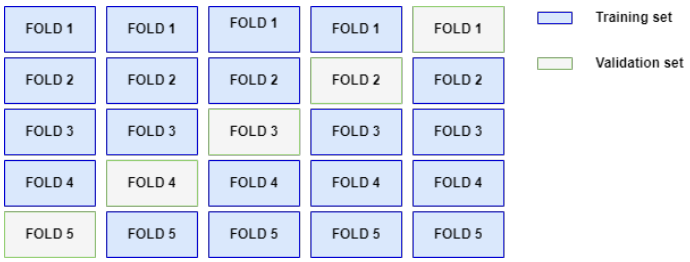


Figure 2.11: Example of a five fold cross-validation. The data is divided into five folds, four for training (blue) and one for validation (green).

2.8.1.2 Unsupervised, Semisupervised and Reinforcement learning

In unsupervised learning, the system attempts to learn without a label. The training data is unlabelled and the algorithm tries to identify groups in the database. The algorithms *"clustering"*, *"visualisation and dimensionality reduction"* and *"association rule learning"* are part of unsupervised learning [65].

The semi-supervised learning algorithm works with a large part of the training data without labelling and a very small part with labelling [65].

The reinforcement learning algorithm is based on a reward and punishment system. The *"agent"* can observe the environment it is in and select/execute actions to perform in that environment, being punished or rewarded depending on the action selected [65].

2.8.2 Machine learning classifiers

2.8.2.1 Support vector machine

A very powerful classification algorithm capable of detecting extremely subtle patterns in complex databases with high accuracy and widely used in bioinformatics projects is the Support Vector Machine (SVM). Its purpose is to generate a decision boundary (hyperplane) between two classes that allows prediction of the labels of one or more feature vectors. On the figure 2.12, two dotted lines represent the classes' margins and a solid line represents a hyperplane that not only separates the

two classes but is also as far away as possible from the closest training instances of each class. If we compare this illustration to a road, the hyperplane will not be affected if we add more training instances outside the road, as the instances at the corners of the road will support it. These instances are called support vectors [65, 70, 71].

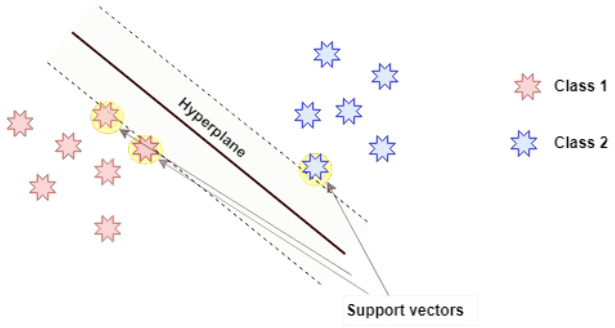


Figure 2.12: Example of a support vector machine classifier.

Originally, Vladmir Vapnik proposed the SVM algorithm in 1963 to create a linear classifier [71]. The linear SVM classification model allows, given a new case $\mathbf{x}$, to predict to which class this case belongs using the decision function given by $\hat{y}= \mathbf{w}^T.\mathbf{x}+b = w_1.x_1 + ... + w_n.x_n + b$, where $\mathbf{w}$ is the feature weights vector, $\mathbf{x}$ is the input feature vector and b is the bias. If the result obtained gives a positive value, it means that class $\hat{y}$ is a positive class (1), if it is negative, it means that the class is negative (-1), as can be seen from the following equation [65, 71].

$$\hat{y} = \begin{cases} -1 & \text{if } \mathbf{w}^T.\mathbf{x}+\text{b} < \text{-1}, \\ 1 & \text{if } \mathbf{w}^T.\mathbf{x}+\text{b} \geq 1 \end{cases}$$

Finding a value for w and b that allows the margin to be as wide as possible to avoid (hard margin) or restrict (soft margin) margin violations means training the linear SVM classifier [65].

2.8.2.2 Naive Bayes

The Naive Bayes classifier is a supervised learning algorithm based on the general assumption that all features are independent of each other given the value of the class variable. This algorithm is quite simple to create, with seemingly simplifying assumptions. Nevertheless, this classifier works very well in complex real-world applications such as medical diagnosis [72, 73].

This algorithm requires a small amount of training data to estimate the necessary parameters and can be extremely fast compared to other, more sophisticated

methods. It is based on Bayes' theorem and helps to determine the conditional probability of the occurrence of two events based on the probability of the occurrence of each event [62, 74].

Two very popular algorithms of this classifier are the Naive Bayes Bernouilli (NBB), where there can be multiple features but each one is considered a binary variable, and the Naive Bayes Gaussian (NBG), which is based on the assumption of a normal distribution of probabilities [62, 74].

2.8.2.3 Decision Trees

Decision trees are a model that integrates a set of basic tests in a coherent and efficient manner. This model allows the comparison of a numerical characteristic with a threshold value for each test. In this model, decision making is hierarchical and follows a path based on the measurement of parameters and subsequent tests. Each tree consists of branches and nodes. In a category to be classified, the features in each node are represented, while the value that the node should take is defined by each subset. This type of algorithm can be used to solve both classification problems and regression problems [75, 76].

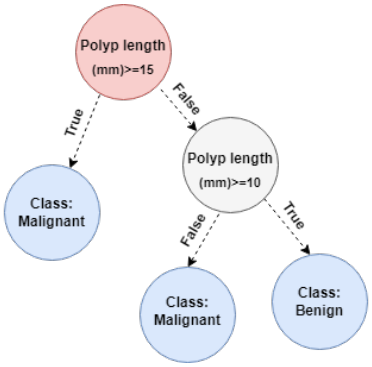


Figure 2.13: Example of a decision tree in the hypothetical classification of a polyp.

An example of how decision trees work can be the following: in a case where a polyp is potentially malignant, we start the model at the root node or at depth 0. At this point, for example, we ask whether or not the polyp has a size greater than 15 mm. If the answer is yes, the next node is called the child node or depth 1, which corresponds to the left side of the image 2.13. If there are no other child node, then we have a leaf node where no further questions are asked. So just look at the class predicted by the model in this node, which in this example is malignant. Now if the polyp is not larger than 15 mm, then the right child node is created where it is asked whether this polyp is larger than 10 mm or not. If yes, then the polyp is probably malignant (depth 2, left), if no, then it is probably benign (depth 2, right).

At the end, the output of the decision tree is given by an intuitive set of rules that follow a certain path along the decision tree.

2.8.2.4 Random forest

The Random Forest Classifier (RFC) was first proposed by Leo Breiman of the University of California in 2009 and is an algorithm that has been successfully used to identify diseases in disease diagnosis. It consists of several decision trees (basic classifier) that are completely independent of each other [77, 78].

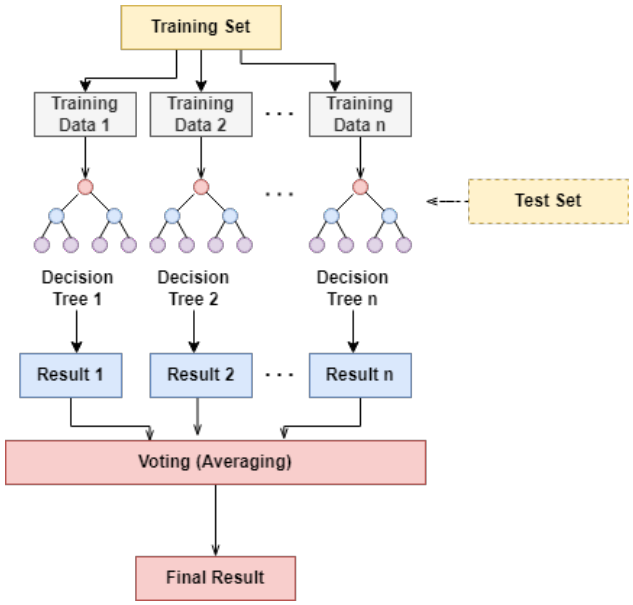


Figure 2.14: Illustration of a Random forest classifier.

An example of this classifier can be found in Figure 2.14. In this figure, when the test data is inserted into the classifier, its label is determined based on the voting of the results of each classifier (each decision tree).

2.8.2.5 Logistic regression

The logistic regression classifier (LR) is a statistical model in which a logistic-like function is fitted to the dataset, modeling the probability of occurrence of a class. These models are widely used in statistics and have proven useful in many real-world problems [79].

This model calculates a weighted sum of the input and output characteristics of logistics for this outcome. This logistic (or logit) is a sigmoid function that outputs a number between 0 and 1. The equation 2.9 gives the Logistic Regression model

estimated probability [65].

$$\hat{p} = h_\theta(X) = \sigma(\theta^T.X) \tag{2.9}$$

From the equation 2.9, θ is the parameter vector. The logit is the σ and is calculated as:

$$\sigma(t) = \frac{1}{1 + exp(-t)} \tag{2.10}$$

Once the model estimates the probability that an instance (X) belongs to a positive class, a prediction $\hat{y}$ is easy to make [65].

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5, \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases}$$

2.8.2.6 Multi-layer perceptron

The multilayer perceptron (MLP) is a type of neural network algorithm consisting of three types of layers: the input layer, the output layer and the hidden layer. The input layer enables the reception of the signals to be processed, while the output layer is responsible for the necessary tasks such as prediction and classification. Between the input and output layers are the hidden layers, which can be any number and are responsible for the actual computational engine of this algorithm. In MLP, the data is received in the input layer and the desired task (e.g. a classification task) is executed in the output layer [80]. An example of this algorithm can be seen in 2.15.

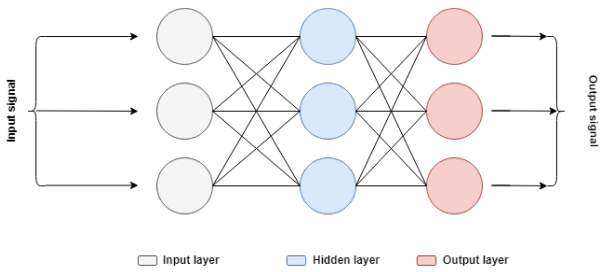


Figure 2.15: Illustration of a multilayer perceptron classifier with a single hidden layer.

2.9 Radiomics Applications - CRCLM

In a study of 194 patients with rectal cancer, Shu et al. [81] used radiomics to predict the risk of synchronous liver metastases (SLM) in these patients. They succeeded in extracting about 328 radiomic features from T2W images acquired with 3T MR. To reduce the feature set and create the radiomic signature, they turned to LASSO. They then used principal component analysis (PCA) to select the remaining features. The predictive model was then built using linear regression and decision curve analysis to test the advantages of LASSO and PCA. They also used two independent cohorts, one for training and one for validation, and found that the model using LASSO had the greatest benefit, with an AUC of 0.857 and a confidence interval of [0.787-0.912] in the training set and an AUC of 0.834 [0.714 -0.918] in the validation set. They created a nomogram using the multivariate logistic regression model, which was combined with clinical risk factors and the features obtained using LASSO. With this nomogram, they obtained very positive results, with a predictive performance in the training set of 0.921 [0.862-0.961] and in the validation set of 0.912 [0.809-0.97]. Finally, they concluded that the use of the radiomic technique in primary rectal cancer has great potential to become a non-invasive clinical tool for predicting the risks associated with SLM.

In a study with mice, Becker et al.[82] investigated whether they could correlate texture features with the growth of an intrahepatic tumour before metastases became visible to the naked eye. To perform this textural analysis procedure, they applied the Radiomics technique to MRI-extracted images of the livers of eight male C57BL6 rats (8-10 weeks) injected with syngeneic MC -38 colon cancer cells and two rats injected with phosphate-buffered saline as a control group. They used MATLAB routine for the textural analysis, which allowed them to extract four first-order features and twenty-eight higher-order features. The R software was used for the statistical analysis. Since it was a very small sample, it was not possible to use machine learning algorithms to test the usefulness of a large number of features. In this study, they found that three features, Energy, SRE (GLRLM) and GLN (GLSZM), were independent and had a linear correlation before metastases became visible, and they also checked several other co-dependent features. Surprisingly, the features obtained from the gray-level GLCM, GLRLM and GLSZM matrices were found to be influenced by the growth of metastases, which could be an indicator of tumour neovascularisation or destruction of the hepatic acini (smallest functional unit of the liver). Although these authors conclude that MRI textural analyses have the potential to detect liver metastases so early that they are not yet visible to the naked eye, further studies are needed and should be performed in human patients if

possible.

2.9.1 Radiomics to predict HGPs

2.9.1.1 CT images

Cheng et al [33] used the radiomics approach to predict HGPs in CRLMs from images of patients with chemotherapy naive CRLM, undergoing to an abdominal contrast-enhanced multidetector CT (MDCT). A total of 126 CRLMs with histopathological confirmation were evaluated. Image segmentation was performed in the pre- and post-contrast phases (AP and PVP) and resulted in a total of 540 radiomic features extracted from the tumour-liver interface (TLI). The lesions were divided into training and validation for model construction and external validation. Minimum redundancy and maximum relevance were used to select the HGP-related features. To distinguish between desmoplastic and replacement, they used decision trees as classifiers. Finally, to assess whether qualitative imaging and clinical factors have a potential power to predict HPs, they combined selected clinical factors with the three radiomic signatures obtained in the previous step. In addition, the authors constructed a nomogram based on the extent to which the radiomic signature as well as clinical factors contribute to the differentiation of HGPs. As a result, they found that fusion of the three fused radiomics signature phases resulted in better predictive performance, with an AUC of 0.926 for the training cohort and 0.939 for the external validation cohorts. Finally, they concluded that a radiomic model applied to MDCT images has the potential to predict the HGPs of CRLMs non-invasively.

2.9.1.2 MR images

Han et al. [13] performed a study with a sample of 182 resected and histopathologically proven patients from two institutions, with exclusion criteria being inadequate quality of images from MR for analysis, patients who had received preoperative systemic and/or regional treatments, and inadequacy of haematoxylin and eosin-stained sections of the tumour-liver interfaces of the resected CRLM specimen. To obtain images of the lesions, they used magnetic resonance imaging. Subsequently, the ROIS were manually delineated using the ITK-SNAP software, in 5 image sequences, T1W, T2W, AP, PVP and ADC. Finally, about 74 textural features and 18 first-order features were identified. For feature selection, the intra/interclass correlation coefficient was used to select the most stable features and the robust feature selection (RFS) [83] method was used to select the radiomic features. To build the model, they used the decision tree to evaluate the ability of each sequence to predict the HGPs. They obtained five signatures after applying this algorithm to the

five sequences for the tumour region and for the tumour-liver interface. Then they used the forward stepwise regression method to select the desired sequence from the five signatures and finally generated the final signature. They also performed cross-validation at both institutions to avoid the effect of a training/validation cohort split.

The final result of this study showed that the TLI radiomics model performed better than the tumour zone radiomics, with an AUC of 0.912 versus 0.879 for internal validation. The combination of models proved to discriminate well, with an AUC of the training cohort nomogram of 0.971, the internal validation of 0.909 and the external validation cohort of 0.905, so at the end of the study they concluded that MRI-based radiomics is a possible method with great potential for predicting the predominant HGPs in CRCLM.

3

Methods

This chapter presents the methods used to carry out this work. It is divided into six sections describing every step from the acquisition and compilation of the dataset to the final model for the classification task.

3.1 Patient population

The dataset is composed by MR images of the liver of patients who underwent hepatectomy for CRCLM in the surgical department of the Coimbra University Hospital Center between 2013 and 2020. The exclusion criteria for patient selection were: rehepatectomy; incomplete clinical records; suboptimal imaging files; non-assessable pathological material; and no preoperative hepatospecific contrast-enhanced liver MR was performed more than 40 days before surgery. A total of 37 patients were included.

The patient records contain information such as age, sex, interval between MRI and surgery, date of surgery, MRI equipment, clinical presentation (whether the lesion is metachronous or synchronous), location of the lesions and size of the lesions. This information corresponds to the categorical features. Each individual lesion was considered as one instance of the dataset in this project, and the lesions should have a size equal or greater than 10 mm to be included. Each patient may or not present more than one lesion. Therefore, a total of 82 lesions were evaluated.

The present study was authorised by the Institutional Review Board of the Coimbra University Hospital Center (number CHUC-127-19).

3.2 Pathological characterisation

The Hematoxylin and Eosin (H&E) Staining was used to carry out the pathological characterisation of each lesion. The H&E technique is the modern basis for cancer diagnosis. This technique is routinely used to identify different tissue types

and morphological changes. The contrast between acidophilic eosin and basophilic haematoxylin makes it possible to observe the different parts of the cells and their types. Haematoxylin has a dark blue colour. It can bind to structures that have an acidic character, such as the nucleus and the rough endoplasmic reticulum, giving them a blue-violet hue. Eosin, on the other hand, has a pink colour and stains proteins. So, in a typical healthy tissue, we see the nucleus in blue and the cytoplasm (as well as the extracellular matrix) in different shades of pink [84, 85].

Histopathological growth patterns can be quickly identified by the pathologist and therefore do not require time-consuming or expensive investigations, which is an advantage for low-resource settings. To correctly analyse and classify CRCLM HGPs, a detailed macroscopic examination of a specimen deemed appropriate is required with at least one specimen per tumour centimetre [8].

The pathologist can recognise the desmoplastic pattern by a fibrous arc, amorphous nuclei and the angiogenic character. In metastases with the replacement pattern, the infiltrative character is observed, while in pushing, there is a compression of the hepatocytes by the tumour cells. An example of a hepatologist's vision for the desmoplastic pattern can be seen in Figure 2.2, for the pushing pattern in Figure 2.4 and for the replacement in Figure 2.6.

In mixed patterns, the pathologist observes more than one of these patterns, each corresponding to up to 25% of the tumour surface area of the liver parenchyma [7]. For a pattern to be correctly classified, it must be present on approximately 50-75% of the peace.

In this project, all lesions were assessed by an experienced pathologist using the H&E technique. After evaluation, a total of 41 lesions were identified as desmoplastic, 24 as replacement, 10 as pushing, 6 as pushing and replacement (mixed patterns) and 1 as pushing and desmoplastic (mixed patterns), thus forming the "ground truth" for our machine learning approach.

3.3 Image acquisition

Magnetic resonance images of the liver of the patients were acquired using 1.5T and 3T machines (see sections 2.3 and 2.7.1.1). In this project, images from both devices were used to avoid reducing the sample size. The images assessed were extracted the T1W Dual Echo Sequence (with in-phase and out-phase), T2W FS (fat-suppressed to suppress the signal from the adipose tissue) and T1 FS Portal phases (contrast study), with the Portal phase being the best phase to observe the metastases at the radiological level.

The parameters used for the image acquisition are listed in Tables 3.1 and 3.2. The image files of each patient were retrieved from the image archiving and communication system (PACS) of the hospital. The corresponding slices from each patient were stored in Digital Imaging and Communications in Medicine (DICOM) format, the international standard for medical imaging.

Table 3.1: Parameters used for image extraction with a 1.5T machine.

Phase	Parameter	Value
Portal	*Slice*	3 *mm*
	TR	4.88 *ms*
	TE	2.31 *ms*
$T1W$	*Slice*	8 *mm*
	TR	101 *ms*
	TE	2.27 *ms*
$T2W$	*Slice*	5 *mm*
	TR	900 *ms*
	TE	77 *ms*

Table 3.2: Parameters used for image extraction with a 3T machine.

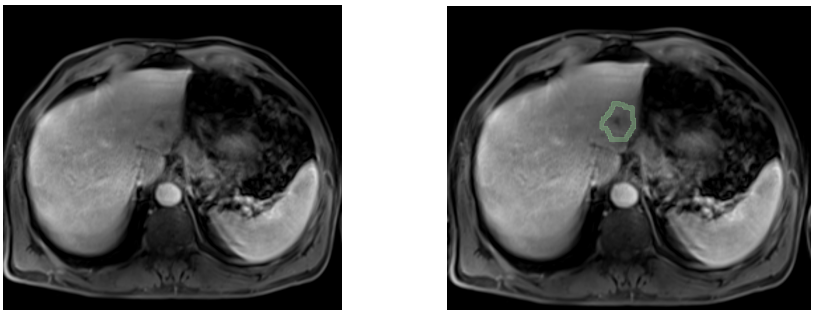
Phase	Parameter	Value
Portal	*Slice*	2.5 *mm*
	TR	4 *ms*
	TE	1.31 *ms*
$T1W$	*Slice*	2.5 *mm*
	TR	4 *ms*
	TE	1.31 *ms*
$T2W$	*Slice*	4 *mm*
	TR	2000 *ms*
	TE	92 *ms*

3.4 ROI Selection

The ROIs were delineated, for each lesion, by an experienced radiologist and manually drawn using the free open-source software package 3D Slicer. In this project, the ROIs were drawn on the raw image.

A 4 mm painting tool was used to delineate the ROIs, which were drawn to include the liver-lesion interface, up to 2 mm to the inside of the lesion and 2 mm to

the outside, to identify features between the tumour and the interface of the liver. Figure 3.1 is an example of a ROI, drawn in one liver metastasis from CRC.



(a) Example of a MR image.

(b) Example of a ROI in a MR image.

Figure 3.1: The first image is a picture of the liver with a metastasis, taken in the Portal phase. The second image shows the ROI, which is located above this metastasis.

Some lesions had vessels in their vicinity, these vessels were excluded during delineation. For each of the 82 lesions, one ROI was extracted in each image acquisition phase (T1W, T2W, and Portal), resulting in 246 ROIS, and each one was saved as an image file in *nrrd* format.

3.5 Image processing and feature extraction

For image processing as well as feature extraction, the `pyRadiomics` package was used [57]. The `pyRadiomics` is an open-source Python package that follows the IBSI [52] for radiomics feature extraction from medical imaging.

First, the images of each phase were introduced together with the ROI masks, then the image processing was performed, consisting of normalisation, interpolation, discretisation and selection of the image type (original and filtered with LoG and wavelet filters), and finally the features were extracted in different files for each phase. These process follows the workflow shown in the figure 3.2 and was carried out using the Python programming language version 3.7.

3.5.1 Image processing

Normalisation was applied to the images of MR because the intensity of the image is usually relative, and it is not possible to compare two images directly [86].

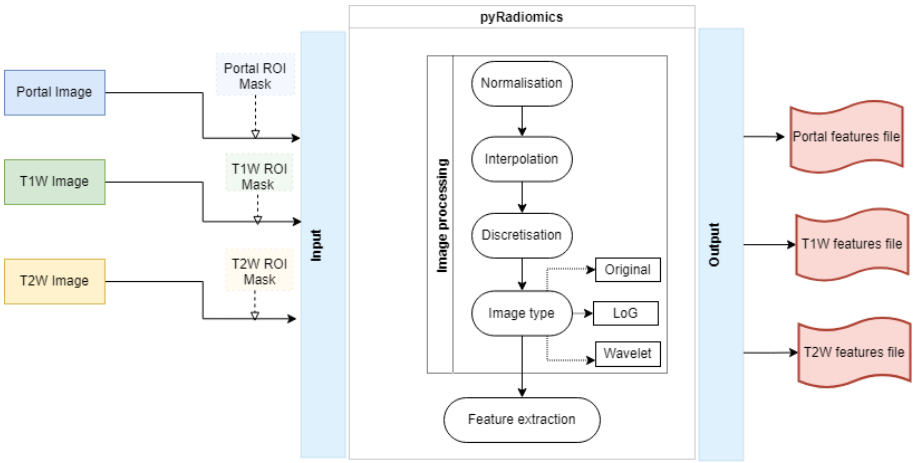


Figure 3.2: Workflow of image processing and feature extraction with the open source Python package `pyRadiomics`. In this package, the original images and their respective ROI masks are used as input values. Then the processing procedures and finally the feature extraction are applied. The values obtained are then stored in files corresponding to the individual phases.

The calculation used for normalisation is described by `pyRadiomics` as follows:

$$f_x = \frac{s(x - \mu_x)}{\sigma_x} \tag{3.1}$$

Where x is the original intensity and f_x is the normalised image. The average of the image intensity values is given by μ_x and σ_x is the standard deviation. The s value is a scaling factor and has the value 1 in this calculation.

3.5.1.1 Interpolation

As described in 2.7.1.3, interpolation is a relevant process in image processing. It is important for reproducibility to have consistent spacing of isotropic voxels across different devices and measurements [52]. Therefore, the images were interpolated with a resample pixel spacing of $2\text{mm} \times 2\text{mm} \times 2\text{mm}$, using the B-spline interpolator.

3.5.1.2 Discretisation

According to Tixier *et al* [87], to calculate the texture features it is necessary to discretise the voxel values in an interval chosen as a power of 2, and therefore the total number of bins should be approximately between 8 and 128 [86, 87]. For this dataset, the majority of gray values in ROIs are in the range between 34 and 300, thus the chosen binWidth value is equal to 3, to include values within the bin range when dividing the smallest range by the chosen binWidth and the same for the largest range.

3.5.1.3 Image type

The wavelet transform and the Laplacian of Gaussian (LoG) filter were applied to the information of the input image (signals). The wavelet transform applies a high-pass filter (H), which allows the passage of signals with high frequencies, and a low-pass filter filter (L), which allows the passage of signals with low frequencies, in the three dimensions of the image, and decomposes it into eight parts. An image labelled as Wavelet-LLL means that the L filter has been applied in the x-direction as well as in the, y- and z- directions. The possibilities of the eight parts are given by the combination of filters in each of the three dimensions: Wavelet-LLL, Wavelet-LLH, Wavelet-LHL, Wavelet-LHH, Wavelet-HLL, Wavelet-HLH, Wavelet-HHL and Wavelet-HHH [13, 57]. An example of a wavelet transformation is shown in the picture 3.3.

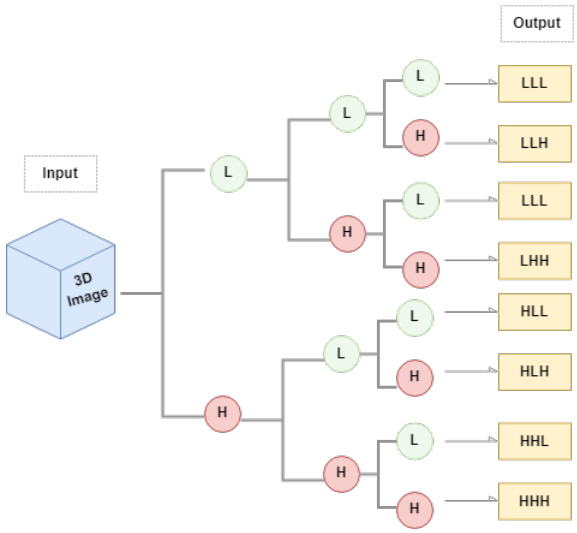


Figure 3.3: Example of a wavelet transformation applied to a three-dimensional image (input). In the output you can see the eight parts of the decomposition.

The LoG filter combines the Gaussian distribution (smooths the image according to the σ value of the filter) with the Laplacian (a differential operator to detect intensity changes in the smoothed image) [88]. In this project, the values of σ equal to 2.0, 3.0, 4.0 and 5.0 were used.

3.5.2 Feature extraction

A total of 1222 features, summarised in Table 3.3, were extracted from each of the ROIs, in each image type.

For the categorical features, only features *Size*, *Lesion type* and *Lesion size* were considered in this study, as these are the only clinically relevant features. The feature

Table 3.3: Radiomic features: First order; Grey level co-occurrence matrix (GLCM); Grey level dependence matrix (GLDM); Grey level run length matrix (GLRLM); Grey level size zone matrix (GLSZM); Neighbourhood grey tone difference matrix (NGTDM), Wavelet and Laplacian of Gaussian were extracted from each ROI using the `pyRadiomics` package.

Radiomic Features	Total
First order	19
GLCM	24
GLDM	14
GLRLM	16
GLSZM	16
NGTDM	5
Wavelet	752
LoG	376
Total	**1222**

Size corresponds to the size of the lesion at the histological level, while the feature *Lesion size* corresponds to the size of the lesion seen on MR images. Therefore, only this information can be compared. A total of 1224 features were considered. The *Lesion type* corresponds to the ground truth.

3.6 Feature selection and model building

After the extraction of the features, the selection of the features and the creation of the classification model was done considering two classification types: Multiclass and Binary classification. For these purposes, the Python language library `scikit-learn` [62], which is geared towards the practical application of machine learning, and the open-source tool `pandas` [89, 90] for data analysis and manipulation, were used.

All the processes described in this chapter were applied to each image acquisition phases (Portal, T1W and T2W). A general overview of the workflow used to create the final model can be seen in Figure 3.4. First, the data were divided into a training set and a test set, then the features were standardised, and then the number of features was reduced (feature selection). The model was then trained and evaluated. In the end, the prediction probabilities were extracted for each phase of the image acquisition (more on this in section 3.6.2). Finally, the probabilities were used as input data for a final decision model. In the final model, feature (probabilities)

selection and performance evaluation were performed.

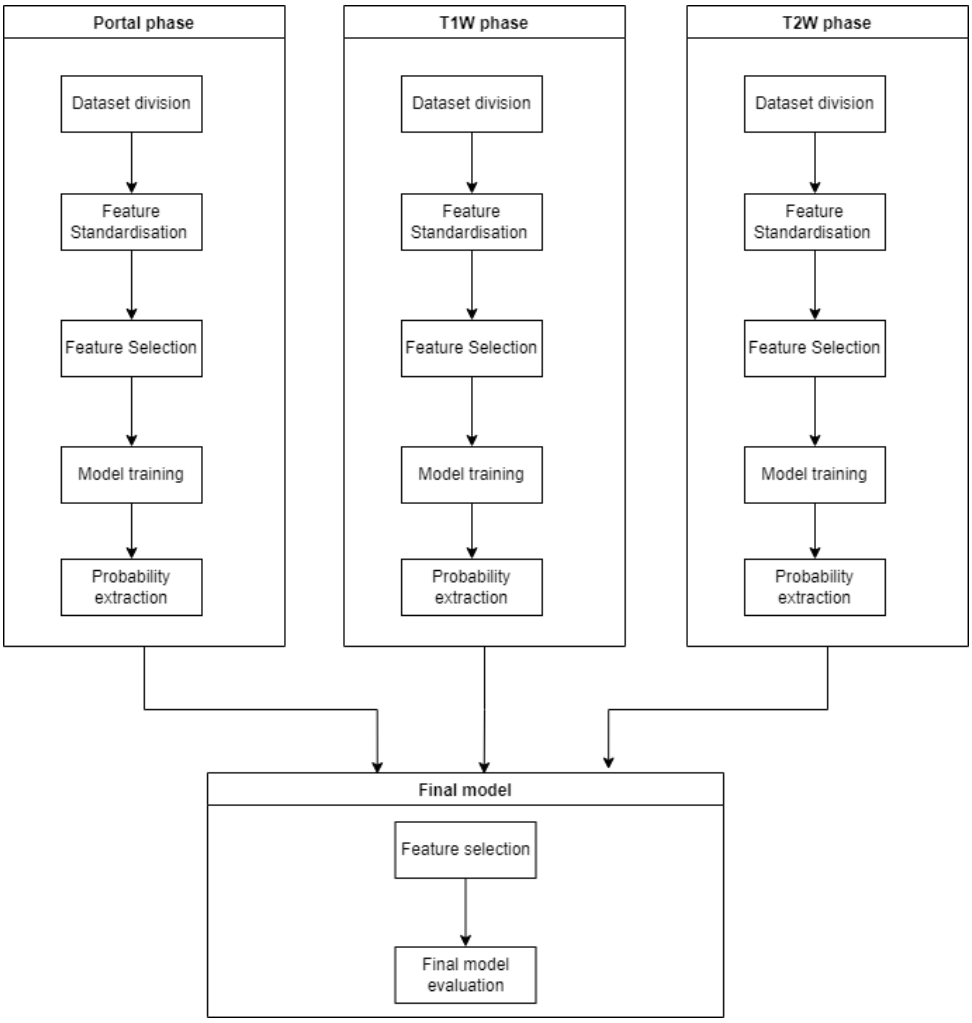


Figure 3.4: Scheme of the workflow adopted. For each of the phases (Portal, T1W and T2W), the data were split, the features were standardised, the features were selected, the model was trained with the training data, the model was evaluated for training and testing data, and finally the probabilities were extracted. The extracted probabilities were then used as input data for the final model combining the three phases. For this model, feature selection and model evaluation were performed again.

Multiclass classification did not include patients with mixed patterns, leaving only desmoplastic (D), pushing (P) and replacement (R) patterns. The data were then split into 90% for training and 10% for testing, randomly, as the number of pushing and replacement cases is quite small.

Clinically, identifying the desmoplastic pattern from an image would already help physicians to choose a more appropriate treatment because, as mentioned in sections 2.5.1 and 2.6.1, this pattern responds better to several less aggressive therapies than the other patterns. For this reason, in the binary classifier, lesions were

divided in two categories, desmoplastic and non-desmoplastic, and there was no exclusion of lesions. The split of the dataset for the binary classifier was 80% for training and 20% for testing, randomly.

When dividing the data set into a training set and a test set, the number of examples of each class in each set were kept proportional. After splitting the dataset, the non-categorical features were standardised to be centred on zero. The calculation performed for standardisation is described in equation 3.2, where x is the standard score of a sample, u is the average of the training samples and s is the standard deviation of the training samples.

$$z = \frac{(x - u)}{s} \tag{3.2}$$

To avoid data leakage between sets, centralisation and dimensioning is done for each feature by calculating the relevant statistics of on the training set only, as the test set is considered future information and if it would be introduced could influence the prediction accuracy and lead to overestimated results. Therefore, after normalising the training group, the mean and variance were stored and used to normalise the test group as well. Figures 3.5 and 3.6 show an example of a feature before and after standardisation respectively.

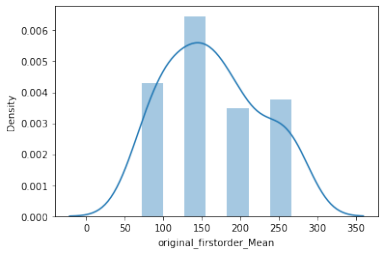


Figure 3.5: Example of a non-standardised first-order feature that represents the average gray level intensity within the ROI.

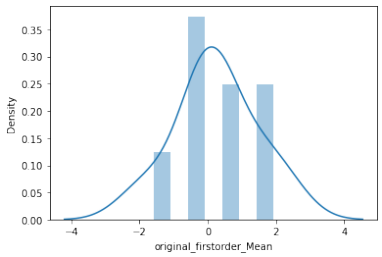


Figure 3.6: Example of a first-order feature representing the average gray level intensity within the ROI after standardisation.

3.6.1 Feature selection

Although the number of features is quite high, many of them may contain redundant or unimportant information, which may affect the performance of the model. Therefore, two different approaches to reduce the features' space were tested.

3.6.1.1 First method

In this first approach (shown schematically in Figure 3.7), which was only applied in the multiclass case approach, features with the module of the Pearson correlation above a threshold of 0.8 were first removed. The Pearson method assigns values in a range between -1 and 1, where -1 means a completely negative correlation, 0 means no correlation and 1 means a completely positive correlation. For example, a positive correlation between two variables means that when one variable increases, the other also increases, while with a negative correlation when one variable increases, the other decreases [91].

Then, agglomerative hierarchical clustering was applied to agglomerate similar features in the same cluster (a new feature is made of an agglomeration of features) to further reduce their dimension. The parameters affinity and linkage were set. Affinity is the method used to calculate the linkage. In this case, the Euclidean metric, which indicates the shortest distance between two points, was used.

The linkage is a criterion that determines the distance between features, and features are grouped to minimise this criterion [62]. In this case, the average of the distances of each observation of two groups was used. A total of 50 clusters were formed for each phase.

The categorical features *Size* and *Lesion Size* had their importance evaluated by applying the tree-based estimator and ranking their importance relative to the non-categorical features. They seemed to have some importance for the three phases, so both were introduced. Finally, the recursive feature elimination (RFE) algorithm with cross-validation was applied.

The RFE is a wrapper-type method that allows you to select a set of features that are most appropriate for a given method. The RFE algorithm recursively considers smaller and smaller groups of features. The importance of each feature is calculated and the less important features are removed from the group [62, 92]. To find the optimal number of features, a cross-validation loop (RFECV) was used.

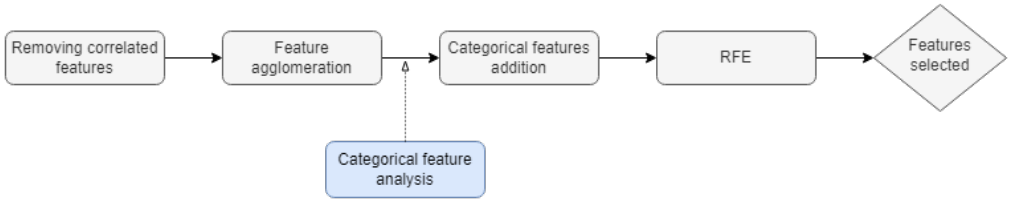


Figure 3.7: Workflow of feature selection according to the first feature selection method.

3.6.1.2 Second method

Considering the possibility that the features with high correlation contain information relevant to the model but were likely to be discarded in the selection method described above, a different approach to feature selection was tested.

This second method was implemented to improve the predictability of the multiclass model, and only this method was used in the binary classification.

First, mRMR (described in 2.7.1.5) was applied, which selects the features with high correlation with the output (class) and low correlation between themselves. Then, the RFE was applied to select the optimal number of features. However, as it can exclude the categorical features and thus prevents the evaluation of their impact on the predictive power of the model, the model was evaluated without the categorical features (see Figure 3.8) and with their introduction after the RFE was applied (see Figure 3.9).

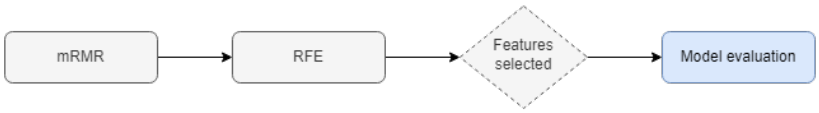


Figure 3.8: Workflow of feature selection according to the second method of feature selection without categorical features.

Figure 3.9: Workflow of feature selection according to the second method of feature selection with categorical features.

3.6.2 Classification

3.6.2.1 Individual models

After feature selection, the training group was evaluated using stratified 8-fold cross-validation for the multiclass classifier and 7-fold cross-validation for the binary classifier, preserving the percentage of samples for each class. The models trained and evaluated were: Logistic Regression, Random Forest, Multi-layer Perceptron, Naive Bayes Bernoulli, Naive Bayes Gaussian and Support Vector Machine. For each model, a grid search was applied, which exhaustively searches for the best parameters for the model. The list of parameters optimised for each model can be found in Table 3.4.

Table 3.4: Models and their respective optimised parameters, the description of which was adapted from [62].

Model	Parameters
LR	*'inverse of regularisation strength','penalty','solver'*
RF	*'bootstrap','maximum depth of the tree'*
	'number of features to consider when looking for the best split'
	'minimum number of samples required to be at a leaf node'
	'minimum number of samples to split an internal node'
	'number of trees in the forest','criterion'
SVM	*'regularisation parameter','kernel coefficient','kernel type'*
MLP	*'hidden layer sizes','learning rate',' alpha','solver'*
	'maximum number of iterations','activation function'
NBB	*'smoothing parameter','class prior','fit prior'*
	'threshold for binarizing ofsample features'
NBG	*'variance smoothing'*

To assess the ability of the model to predict with the training data, the metrics described in 2.8.1.1 obtained with the confusion matrix were used: Precision, Recall, Accuracy and f1-score. The ROC curve was plotted along with the AUC. The One vs. Rest Classifier was used to calculate the ROC and the AUC curve in multiclass models. For the multiclass models, which have an unbalanced amount of data for each class, Precision versus Recall was used as another evaluation metric.

Test data was then fed to the models to evaluate the generalisation performance. The probability of an event belonging to a particular class was extracted for the model with the best performance for the Portal, T1W and T2W phases and these probabilities were used as input variables for the final model for training and testing.

3.6.2.2 Final model

The probabilities from the previous step became the new features, resulting in 9 features for multiclass and 6 for binary. At this stage, the new dataset is already divided into training and testing. RFE was applied again to select the features. The classifier used in this step was SVM. The final model was then evaluated using the metrics mentioned earlier: Precision, Recall, Accuracy, f1- Score, ROC curve and a Precision versus Recall curve.

3.6.2.2.1 Evaluating the redundancy of the phases

The possibility of redundant information from the phases was also tested. Therefore, three additional final models were created, considering only two phases instead of three, i.e. Portal with T1W, T1W with T2W and Portal with T2W.

4

Multiclass classifier

This chapter presents and discusses the results of the multiclass approach for predicting classes desmoplastic (D), pushing (P) and replacement (R), using the procedures described in chapter 3.

After splitting the dataset, the number of cases in the training data was 37 cases of class D, 8 cases of class P and 22 cases of class R. For the test data, the split was 4 cases of class D, 2 cases of class P and 2 cases of class R.

4.1 Feature selection with the first method

After the clustering process, a total of 50 clusters were formed for each phase (Portal, T1W and T2W). The categorical features were assessed for relevance in relation to the other clusters and individually (described in 3.6.1.1). Both categorical features showed relevance in the data set and were therefore retained.

These features were added to the RFE along with the clustered features, which resulted in a total of 37 optimal features for the Portal phase, 13 for the T1W phase and 37 for the T2W phase. The model was trained and evaluated considering these selected features.

4.1.1 Model evaluation

The metrics used to evaluate the models (precision, recall, f1-score and accuracy), obtained through the confusion matrix (described in section 2 of topic 2.8.1.1), were described in chapter 3.

The f1-score, which combines the precision and recall metrics, provides information about how accurate and robust the classifier is. It was therefore used as the main metric for selecting the best model based on cross validation on the training data.

The ROC curve was also used, but this metric may be considered insufficient

with unbalanced data, as a small number of correct or incorrect predictions may cause a sudden change in the curve and thus in the AUC value. Therefore, the precision vs. recall curve, which expresses the relationship between precision and recall for different probability thresholds, was plotted to facilitate the interpretation of the results and the selection of the model.

The plot of precision vs. recall is interpreted similarly to the ROC curve. The closer to 1, the better the classifier; at 50%, the classifier is considered random and below that, poor.

4.1.1.1 Portal phase

The best results for the Portal phase using the training data were obtained with the Naive Bayes Gaussian algorithm previously described in 2.8.2.2.

For this phase, a precision of 63% for class D and 50% for class R was obtained. Using the recall metric, it was verified that the lesions can be correctly identified as D in 89% of the cases, compared to 32% for R. The model has a classification capacity (f1-score) of 0.74 for class D, while for class R it is only 0.39.

The accuracy for the training data for this phase was 60%. However, this metric is not appropriate for imbalanced data, as it is possible to obtain good results even with a poor classifier by ignoring minority classes such as class P. Therefore, this metric was not considered in the selection of the model.

For the testing data, the classifier of the Portal phase achieved a classification capacity (f1-score) of 67% for class D and 80% for class R. This indicates that this model is a good classifier for class R. However, looking at the results of the training data, which includes many more cases, the classifier for class D seems to be more reliable.

In the case of class P, due to the small amount of data for both training and testing, it was expected that the model would not be able to classify the cases correctly and this was observed as most of the metrics gave a zero result.

4.1.1.2 T1W phase

For the T1W phase, the support vector machine described in 2.8.2.1 proved to be the best classifier. For the training data, the model was able to correctly classify all cases belonging to class D and therefore had a recall of 1.0 or 100%.

For class R, only two cases were correctly classified and verified a precision of 1.0. However, at the time the model was selected, this was not considered a reliable metric because, according to the equation described in 2.6, TP and FP are taken into

account, but there was no FP in the confusion matrix, so only TP was considered, resulting in a precision of 100% (or 1.0).

Thus, the f1-score continued to be the most considered metric. It was 0.73 for class D and only 0.17 for class R. The Precision versus Recall plot was consistent with these results.

The accuracy was 0.58 in the training data and 0.62 in the test data. In the test data, the recall of the class D was 100%, the precision was 57% and the f1-score was equivalent to 0.73.

4.1.1.3 T2W phase

In the T2W phase, the random forest classifier described in 2.8.2.4 was considered the one with the best performance. The accuracy was 0.58 for the training data and 0.5 for the test data.The choice of this classifier was mainly based on the results obtained with the f1-score. The f1-score in the training data was 0.70 for class D, indicating that this model might be suitable for classifying this class. Its performance was then evaluated for the test data.

The confusion matrix obtained with the test data showed that all D cases were classified correctly, but no R and P cases were classified appropriately, resulting in an f1-score of 0.73 for class D and zero for the others.

4.1.2 Final model

As described in the chapter 3, the probabilities extracted from each phase were introduced as new features in a final model. In this step, the features did not need to be standardised as they are already on the same scale. After applying the RFE, a total of 6 features (out of 9) were identified.

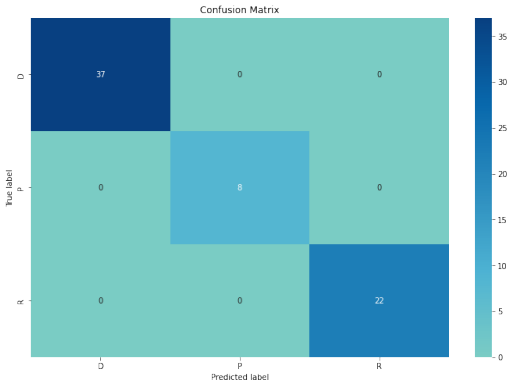


Figure 4.1: Confusion matrix for the training data for the final model.

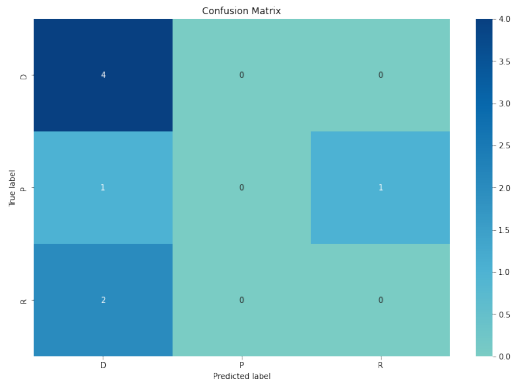


Figure 4.2: Confusion matrix for the test data for the final model.

The SVM was then applied. Figure 4.1 and 4.2 shows the results of the confusion matrix for the training data and for the test data, respectively. In figures 4.3 and 4.4 it is possible to observe the ROC curve for the training and test data, respectively.

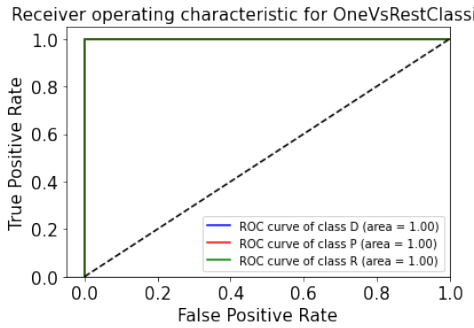


Figure 4.3: ROC curve and AUC for the training data in the final model.

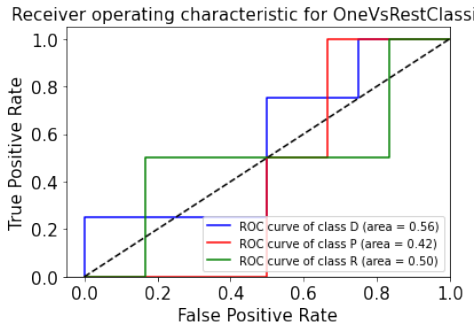


Figure 4.4: ROC curve and AUC for the testing data in the final model.

Table 4.1: Results of the evaluation metrics for the final model with the SVM classifier for the training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	1.00	1.0	1.00	*D*	0.57	1.00	0.73
P	1.00	1.00	1.00	*P*	0.00	0.00	0.00
R	1.00	1.00	1.00	*R*	0.00	0.00	0.00

From figures 4.1 and4.3 and table 4.1, it appears that the model is perfect for the training data. This could be possible considering that the information entered into this final model is information retrieved from previously trained individual phase models.

In the test data, it can be observed that the classifier seems to classify only class D, which is the majority class and therefore not a suitable classifier.

4.1.2.1 Phases redundancy

Considering that some phases may introduce redundant information and that one classifier is not as good as the others due to this effect, the final model was tested by introducing information from only two phases: Portal with T1W, Portal with T2W and T1W with T2W.

It was found that the TIW with T2W model have better performance than the others for the test data, with an f1-score of 0.73 for class D, while the others had a

score of 0.67.

The accuracy was also calculated for each of the models and all had a score of 0.5 for the test data. The f1-score showed no significant improvement compared to the final model using the three phases in the test data, wich was 0.73 (Table 4.1).

4.1.3 Conclusions

Some results suggest that this model has the potential to classify class D. However, perhaps clustering is not the best approach to feature selection for this study, motivating the use of mRMR as presented in the next section.

4.2 Feature selection with the second method

In an attempt to improve on the previous results, the second approach described in 3.6.1.2 was tested.

4.2.1 Disregarding categorical features

The workflow shown in Figure 3.8 has been used, where categorical features are not considered. After applying mRMR, the best features were extracted and then introduced into the RFE. In total, 11 features were obtained for the Portal phase, 13 for the T1W phase and 10 for the T2W phase.

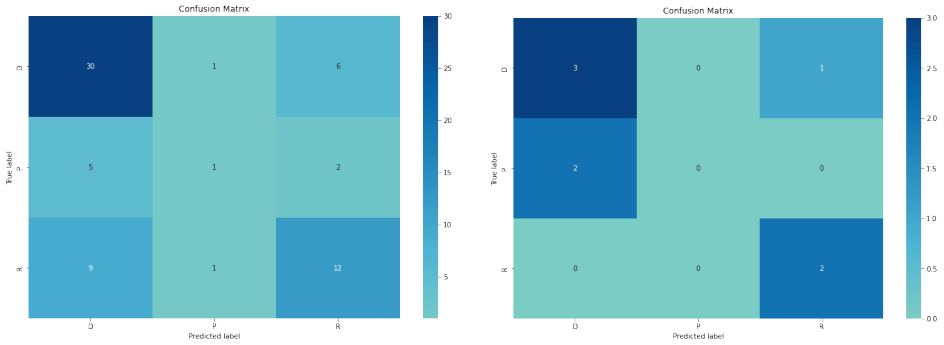
4.2.1.1 Model evaluation

The results of applying the metrics already described to this new approach are shown below.

4.2.1.1.1 Portal phase

In the Portal phase, the best results were obtained with the Naive Bayes Bernoulli classifier described in 2.8.2.2.

In figure 4.5 it is possible to verify the results of the confusion matrix acquired for the training data and for the test data. For the training data, the model was able to predict 30 D cases correctly, 12 R cases correctly and only 1 P case. As for the test data, the model correctly predicted 3 D cases, zero P cases and all R cases.



(a) Train set **(b)** Test set

Figure 4.5: Confusion matrix for the training and test data of the Portal phase with the classifier NBB.

The results of the evaluation metrics for the training and test data are described in Table 4.2.

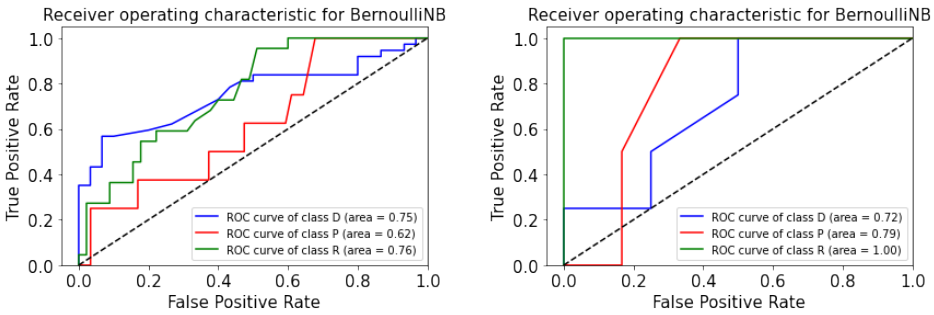
Table 4.2: Results of the evaluation metrics for the Portal phase model for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.68	0.81	0.74	*D*	0.60	0.75	0.67
P	0.33	0.12	0.18	*P*	0.00	0.00	0.00
R	0.60	0.55	0.57	*R*	0.67	1.00	0.80

The table shows that the model has an acceptable performance in predicting class D in both the training and test data. The f1-score for the test data for class R is quite high as the model was able to predict all cases correctly. However, it should be noted that the sample space is quite small (only two cases).

In 4.6 it is possible to verify the results for the ROC curve and the AUC for the training and test sets, and in figure 4.7 the results for the precision versus recall curve are presented.

It is possible to observe that the ROC curve for the training data matches the precision versus recall curve for classes D and R, as both indicate that this is a good classifier for these classes. However, there is a discrepancy in the evaluation of class P, where the ROC curve indicates that it is a relatively suitable classifier for this class, with an AUC of 0.62, but if we look at the table 4.2 and the Precision versus Recall curve, we can see that it is a weak classifier for this class.



(a) Train set **(b)** Test set

Figure 4.6: ROC Curve and the respective AUC obtained for the training and test data of the Portal phase with the NBB classifier.

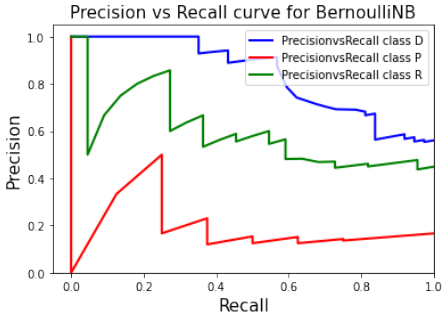


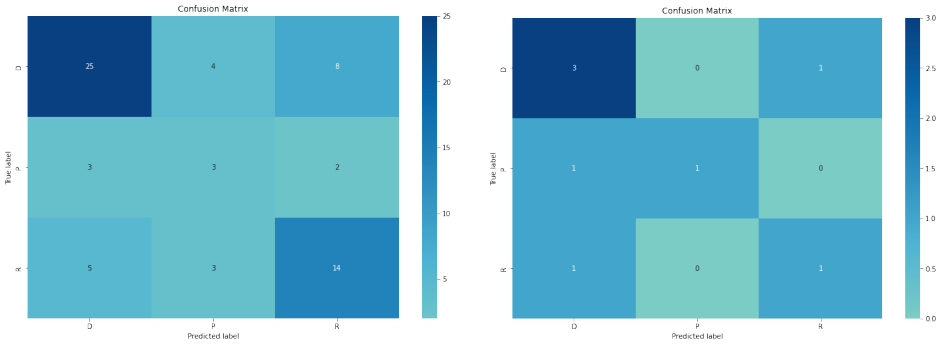
Figure 4.7: Precision vs recall curve obtained for the training data of the Portal phase with the NBB classifier.

For this phase, the accuracy is 0.64 for the training data and 0.62 for the test data.

4.2.1.1.2 T1W phase

The best model for classification in the T1W phase was the NBB, as in the Portal phase. The results of the confusion matrix obtained for the training data and for the test data can be seen in figure 4.8. For the training data, the model correctly predicted 25 D cases, 14 R cases and 3 P cases. In the test, the model correctly predicted 3 out of 4 D cases, 1 out of 2 P cases and 1 out of 2 R cases.

Table 4.3 shows the results of the evaluation metrics for the training and test data. For class D, it is observed that f1-score is suitable for both the training and test data, while for class P, f1-score is suitable for the test data, but looking at the training data, it is seen that the classifier for this class is weak. Figure 4.9 shows the ROC and the AUC curves for the training and test data, and Figure 4.10 shows the precision versus recall for the training data.

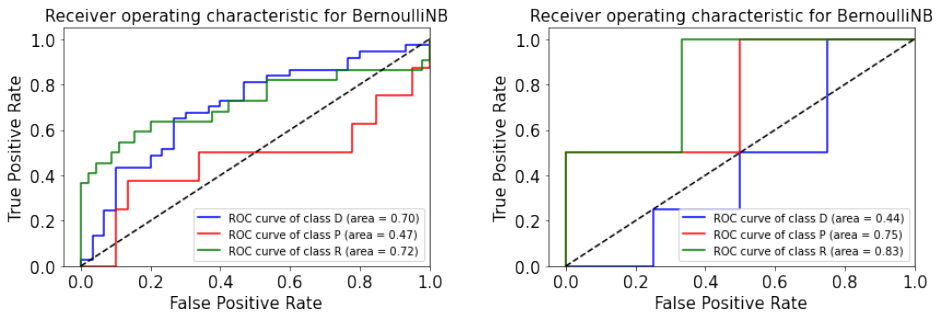


(a) Train set **(b)** Test set

Figure 4.8: Confusion matrix for the training and test data of the T1W phase with the classifier NBB.

Table 4.3: Results of the evaluation metrics for the T1W phase model for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.76	0.68	0.71	*D*	0.60	0.75	0.67
P	0.30	0.38	0.33	*P*	1.00	0.50	0.67
R	0.58	0.64	0.61	*R*	0.50	0.50	0.50



(a) Train set **(b)** Test set

Figure 4.9: ROC Curve and the respective AUC obtained for the training and test data of the T1W phase with the NBB classifier.

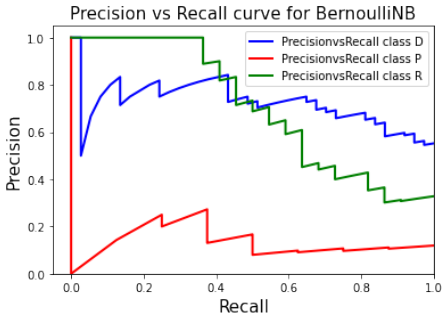
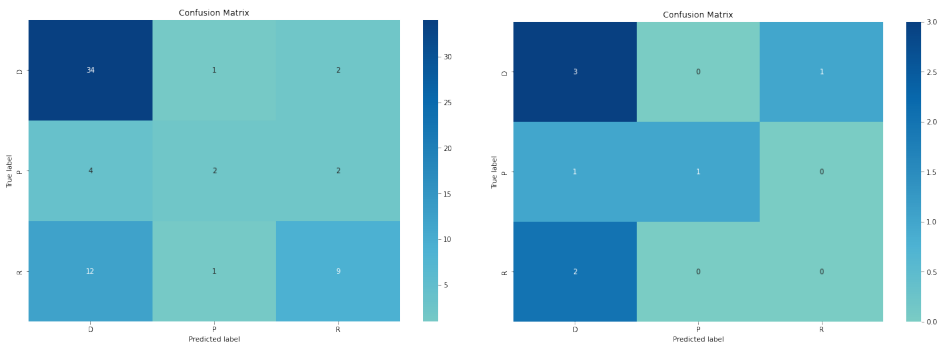


Figure 4.10: Precision vs recall curve obtained for the training data of the T1W phase with the NBB classifier

From Figures 4.9 and 4.10, it can be seen that for class P in the training data, the ROC curve and the precision vs recall curve are consistent with each other and also consistent with the results described in the table, which means that this classifier is a weak classifier for class P. For class D data, the model is a suitable classifier for both training and testing data, and the same is true for class R. The accuracy obtained for this phase is 0.63 for the training data and 0.62 for the testing data.

4.2.1.1.3 T2W phase

For the T2W phase, the best results were obtained with the Logistic regression classifier described in 2.8.2.5. From the confusion matrices obtained for the T2W phase, for the training data almost all cases from class D were correctly predicted (34 out of 37) and for the test data the performance was similar (3 out of 4).



(a) Train set **(b)** Test set

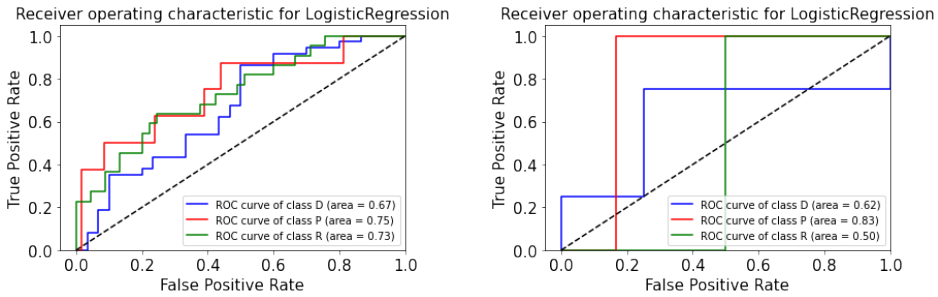
Figure 4.11: Confusion matrix for the training and test data of the T2W phase with the classifier LR.

From the ROC curves of the training and test data, it can be seen that the

classifier is appropriate for class D, with an AUC of 0.67 for training and 0.62 for test, which is consistent with the graph of Figure 4.13 and with the results in Table 4.4. The accuracy of the model for the T2W phase is 0.67 for the training data and 0.5 for the test data.

Table 4.4: Results of the evaluation metrics for the T2W phase model for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.68	0.92	0.78	*D*	0.50	0.75	0.60
P	0.50	0.250	0.33	*P*	1.00	0.50	0.67
R	0.69	0.41	0.51	*R*	0.00	0.00	0.00



(a) Train **(b)** Test

Figure 4.12: ROC Curve and the respective AUC obtained for the training and test data of the T2W phase with the LR classifier.

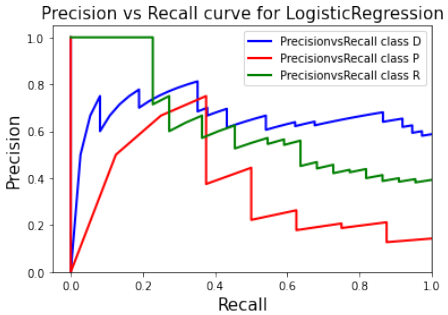
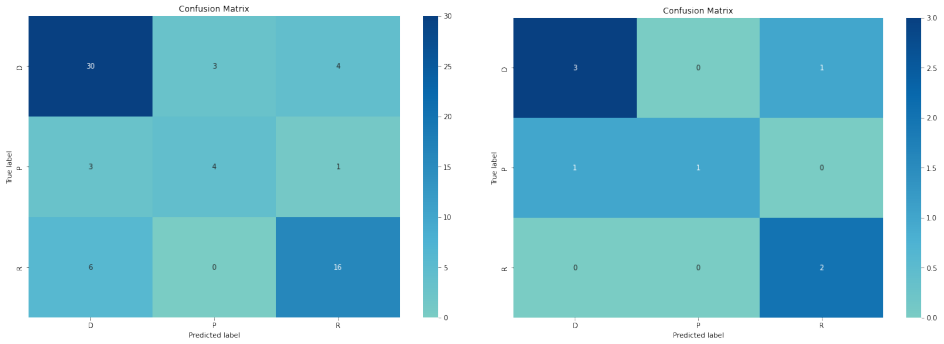


Figure 4.13: Precision vs recall curve obtained for the training data of the T2W phase with the LR classifier

4.2.1.2 Final model

After applying the RFE, a total of 8 features were obtained and used in the SVM. The results of the confusion matrix for the training and test data can be seen in figure 4.14. From the confusion matrices, the results of the evaluation metrics are described in table 4.5, where we can observe a very satisfactory f1-score for classes D and R for the training data, and the same is also observed with the test data.

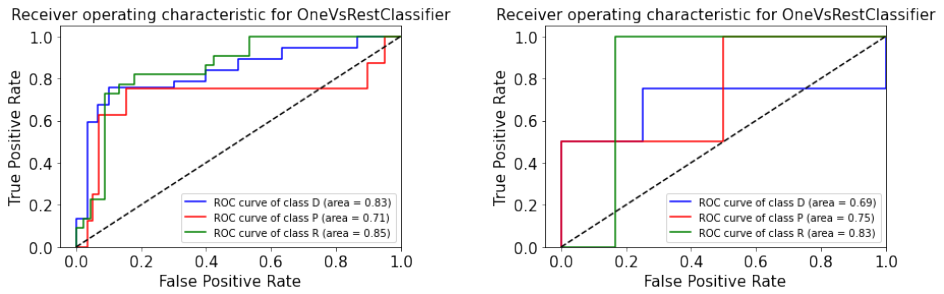


(a) Train set **(b)** Test set

Figure 4.14: Confusion matrix for the train and test data for the final model.

Table 4.5: results of the evaluation metrics for the final model with the SVM classifier for the training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.77	0.81	0.79	*D*	0.75	0.75	0.75
P	0.57	0.50	0.53	*P*	1.00	0.50	0.67
R	0.76	0.73	0.74	*R*	0.67	1.00	0.80



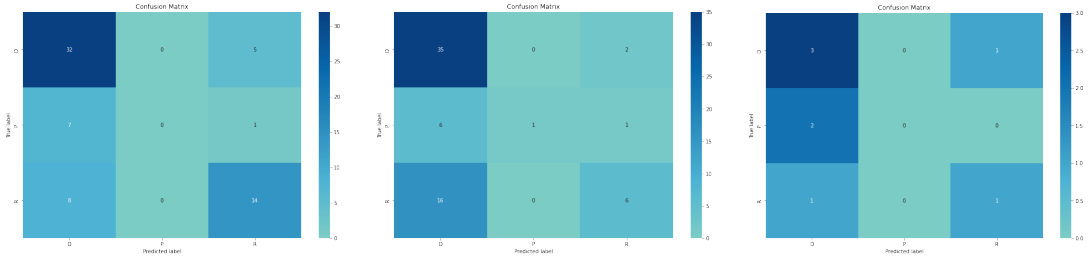
(a) Train set **(b)** Test set

Figure 4.15: ROC curve and AUC for the final model.

In Figure 4.15, it can be seen that this model has an AUC of 0.83 for the training data of class D, and an AUC of 0.69 for the test data of the same class. For class R, the AUC is 0.85 for the training data and 0.83 for the test data. Finally, using the accuracy, a score of 0.76 was obtained for the training data and 0.75 for the test data.

4.2.1.2.1 Phases redundancy

Finally, it was checked whether the performance of the model improved when the information from one of the phases was excluded. From the confusion matrices in figure 4.16, the results for the metrics shown in Table 4.6 were obtained.

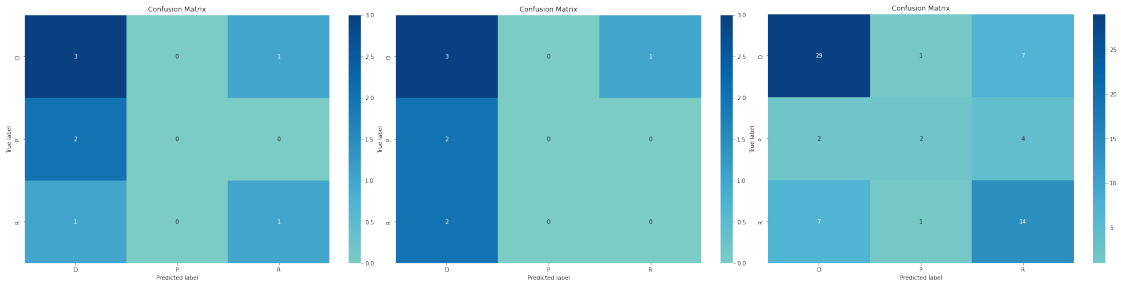


(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 4.16: Confusion matrix for the training data with the SVM classifier.

Table 4.6: Results of the evaluation metrics for the final models with the SVM classifier for the training data.

PortalwithT1W				PortalwithT2W				T1WwithT2W			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.68	0.86	0.76	*D*	0.61	0.95	0.74	*D*	0.76	0.78	0.77
P	0.00	0.00	0.00	*P*	1.00	0.12	0.22	*P*	0.50	0.25	0.33
R	0.70	0.64	0.67	*R*	0.67	0.27	0.39	*R*	0.56	0.64	0.60



(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

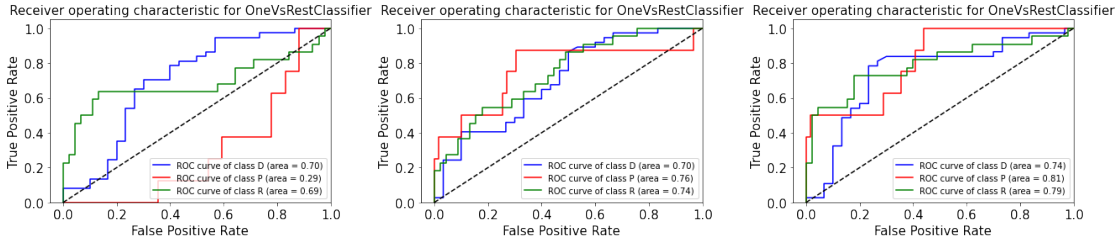
Figure 4.17: Confusion matrix for the test data with the SVM classifier.

From the confusion matrices for the test data, presented in figure 4.17, the results of the evaluation metrics for the test data shown in Table 4.7 were obtained. The ROC curve together with the AUC for the training and test data can be seen in figures 4.18 and 4.19.

The accuracy for the Portal with T1W model was 0.69 for the training data and 0.50 for the test data, while for the Portal with T2W model the accuracy was 0.63 for the training and only 0.38 for the test, while the accuracy of the T1W with T2W phase was 0.67 for the training and 0.50 for the test.

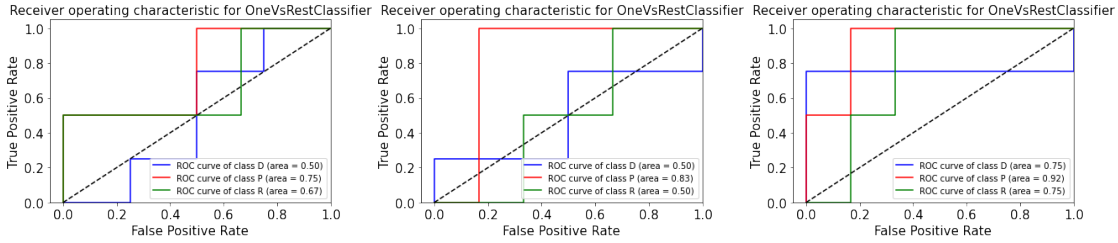
Table 4.7: Results of the evaluation metrics for the final models with the SVM classifier for the test data.

PortalwithT1W				PortalwithT2W				T1WwithT2W			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.50	0.75	0.60	*D*	0.43	0.75	0.55	*D*	0.50	0.75	0.60
P	0.00	0.0	0.00	*P*	0.00	0.00	0.00	*P*	0.00	0.00	0.00
R	0.50	0.50	0.50	*R*	0.00	0.00	0.00	*R*	0.50	0.50	0.50



(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 4.18: ROC Curves and respective AUC for training data.



(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 4.19: ROC curves and respective AUC for test data.

4.2.1.3 Conclusions

From the results, it can be concluded that the model using the information from the three phases has a very satisfactory classification capacity for classes D

and R. When one of the phases is removed, the classification capacity of the model decreases, indicating that the information from each phase is relevant to the model.

4.2.2 Introduction of categorical features

According to the workflow in Figure 3.9, the categorical features were introduced after applying the RFE.

4.2.2.1 Model evaluation

In the Portal phase, the model performed well for the training and test data using the random forest algorithm. The f1-score for the training data was 0.76 for class D and 0.67 for the test data. Class R had an f1-score of 0.56 for the training data, and this result was equal to 0.80 for the test data. Class P had no predicted cases.

The T1W phase showed the best results with the NBB algorithm, with an f1-score of 0.71 for class D, 0.33 for class P and 0.61 for class R on the training data. For the test data, this was 0.67 for classes D and P and only 0.5 for class R.

In the T2W phase, the best results were obtained with the NBG. However, only class D was classified in the test data, with an f1-score of 0.67 for this class.

4.2.2.2 Final model

Using the information from the three phases and applying the SVM, the model obtained an f1-score of 0.57 for the test data in classes D and R (0 for class P), but for the training data, the score was 0.95 for classes D and R, and 0.86 for class P.

4.2.2.2.1 Phases redundancy

In an attempt to improve the model, one of the phases was removed. The results for the Portal with T1W and Portal with T2W models were exactly the same as those obtained with the three phases. This is because RFE selected the same features and included much of the information from the Portal phase. However, when this phase is removed, the model performs even worse on the training data. As mentioned earlier, when using the three phases or the models with the Portal phase, the f1-score was 0.95 for classes D and R. However, when this phase is removed, the f1-score decrease to 0.76 for class D and 0.62 for class R.

The accuracy was also lower for this model, while for the others it was 0.94 for the training data, in T1W with T2W it was only 0.67. In the case of the test data,

all models had a score of 0.5.

4.2.2.3 Conclusions

Given the introduction of the categorical features, the model shows a worse result when the phases are combined. However, when the phases are used individually, the Portal phase was the one that showed the best performance in classifying the class D.

4.3 General conclusions

The model with the second approach, without categorical features and taking into account the information from the three phases, was the one that could classify classes D and R best. Unfortunately, class P could not be reliably classified by the model because the sample was very small.

It has also been found that the performance of the final model for classification decreases when the categorical features are added. However, when evaluating each phase individually, the results for class D are acceptable.

The model was found to have the potential to classify class D. As more data is collected, it is possible that the performance for this class can be improved. For this reason, a binary model was created to try to improve the model's ability to classify this class (against any other class).

5

Binary Classifier

This chapter will present and discuss the results obtained considering a binary classification model for the desmoplastic (D) and non-desmoplastic (ND) classes, using the second method described in 3.6.1.2.

After splitting the dataset, the number of distributed cases in the training data was 32 cases belonging to class D and 33 cases belonging to class ND. In the test data, the split was 9 cases for class D and 8 for class ND.

5.1 Model without categorical features

Following the workflow shown in Figure 3.8 in Chapter 3, a model without the categorical features *Size* and *Lesion size* was first tested.

5.1.1 Feature selection

After applying the mRMR algorithm, the best features were extracted, which were then used in the RFE. For the Portal phase, an optimal number of 9 features was obtained, while for the T1W phase this value corresponded to 10 and for the T2W phase this value was equal to 8.

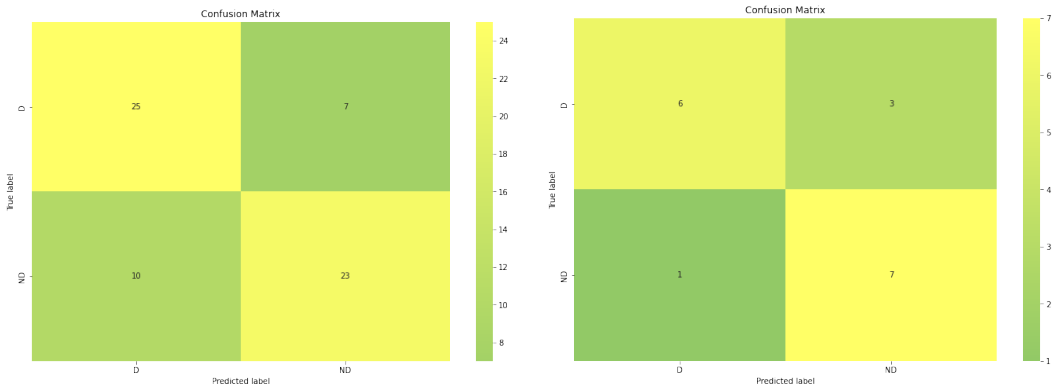
5.1.2 Model evaluation

The metrics used to assess the classification capacity of the model were mentioned in section 3.6.2 of chapter 3 (precision, recall, accuracy, f1-score, ROC curve) and explained in detail in 2.8.1.1 of chapter 2. As the data are balanced between classes, the accuracy metrics and the ROC curve itself are considered.

5.1.2.1 Portal phase

In the Portal phase, the best results were obtained with the linear classifier *logistic regression*. The confusion matrices for the training and test data are presented

in Figure 5.1.



(a) Train set results. **(b)** Test set results.

Figure 5.1: Confusion matrix for the training and test data of the Portal phase with the classifier LR.

From the matrices, the results of the model evaluation metrics described in Table 5.1 were obtained.

Table 5.1: Results of the evaluation metrics for the Portal phase model for training and test data.

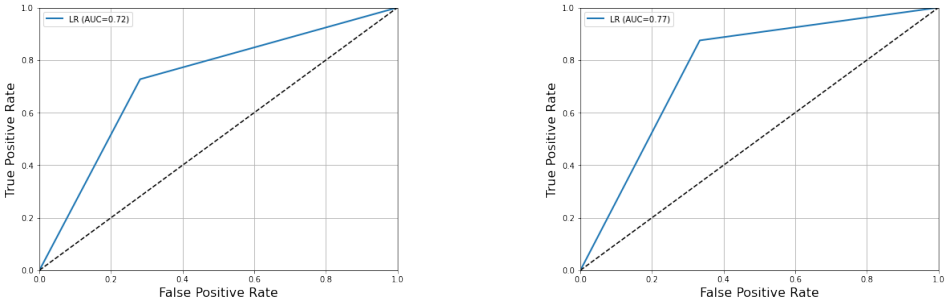
Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.71	0.78	0.75	D	0.86	0.67	0.75
ND	0.77	0.70	0.73	ND	0.70	0.88	0.78

The table shows that in this model, the rate of cases correctly classified as belonging to class D (true-positive cases) relative to all cases classified as belonging to class D (true-positive cases added to false-positive cases) is 71% for the training data and 86% for the test data.

The rate of the cases correctly classified as class D (true-positive cases) in relation to all cases classified as class D (true-positive cases plus false-negative cases) is equal to 78% for the training data and 67% for the test data. Based on the f1-score, it can thus be stated that this classifier is around 75% accurate and robust for both training and test data.

The accuracy for the training data is equivalent to 0.74 or 74%, while for the test data it was 0.76 or 76%. The ROC curves, along with the AUC, of the training and test data can be seen in figures 5.2a and 5.2b.

From the curves it can be seen that the model has a very satisfactory classification performance for the training and test data, with an AUC of 0.72 and 0.77 respectively.

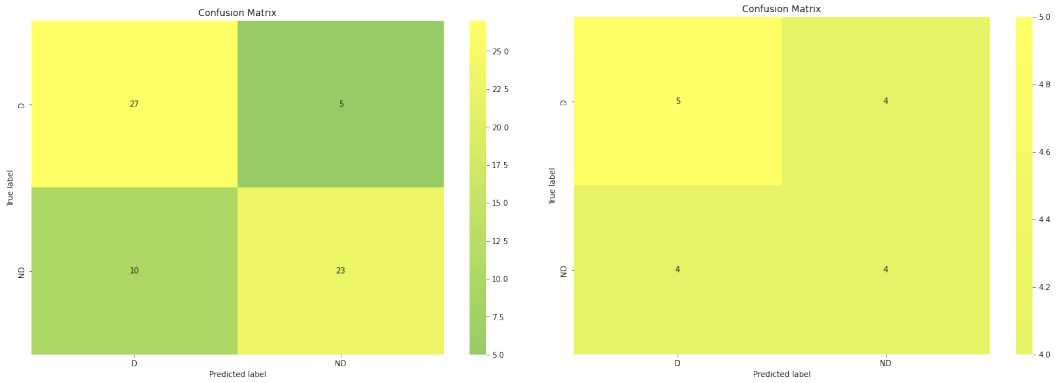


(a) ROC curve and AUC for train set. **(b)** ROC curve and AUC for test set.

Figure 5.2: ROC Curve and the respective AUC obtained for the training and test data of the Portal phase with the LR classifier.

5.1.2.2 T1W phase

For the T1W phase, the best results were obtained with the multilayer perceptron classifier (MLP) described in 2.8.2.6.



(a) Train **(b)** Test

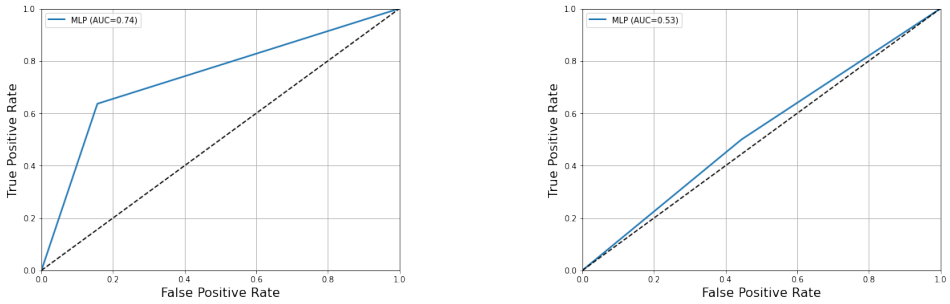
Figure 5.3: Confusion matrix for the training and test data of the T1W phase with the classifier MLP.

From the table 5.1.2.2, the f1-score of only 56% shows that this is probably a random classifier for the test data. The accuracy for the training data was 0.77, while for the test data it was only 0.53.

Figure 5.4 shows that the model for the training data has an AUC of 0.74, which is considered satisfactory, while the classifier for the test has an AUC of only 0.53. Therefore, the model is considered too complex for this phase.

Table 5.2: Results of the evaluation metrics for the T1W phase model for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.73	0.84	0.78	D	0.56	0.56	0.56
ND	0.82	0.70	0.75	ND	0.50	0.50	0.50



(a) ROC curve and AUC for train set. **(b)** ROC curve and AUC for test set.

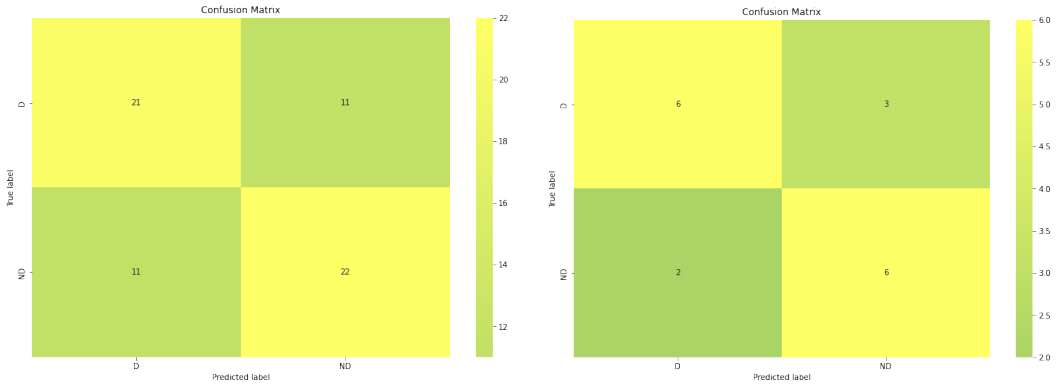
Figure 5.4: ROC Curve and the respective AUC obtained for the training and test data of the T1W phase with the MLP classifier.

5.1.2.3 T2W phase

In the T2W phase, the best results were obtained with the NBB classifier. The results of the confusion matrices and the results of the evaluation metrics for the training and test data can be seen in Figure 5.5 and Table 5.3 respectively.

The table shows that the f1-score of this classifier for the training data is 66% for class D and 67% for class ND. For the test data, this value increases to 71% for class D and 75% for class ND. The accuracy of this model is 0.66 for the training data and 0.71 for the test data.

Figure 5.6 shows that the AUC is 0.66 for the training data and 0.71 for the test data, and the model of this phase generally gives adequate results for classification.

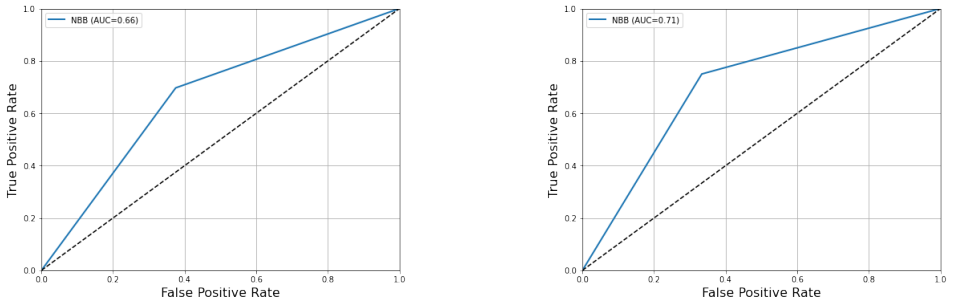


(a) Train **(b)** Test

Figure 5.5: Confusion matrix for the training and test data of the T2W phase with the classifier NBB.

Table 5.3: Results of the evaluation metrics for the T2W phase model for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.66	0.66	0.66	D	0.75	0.67	0.71
ND	0.67	0.67	0.67	ND	0.67	0.75	0.71



(a) ROC curve and AUC for train set. **(b)** ROC curve and AUC for test set.

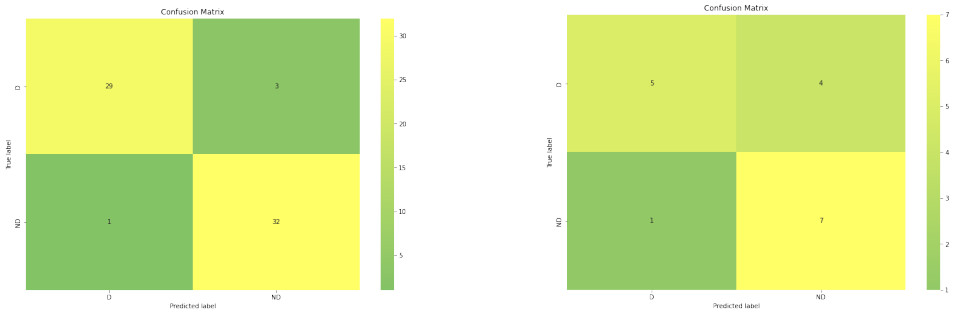
Figure 5.6: ROC Curve and the respective AUC obtained for the training and test data of the T2W phase with the NBB classifier.

5.1.3 Final model

In order to test the impact of a phase containing information from a model that does not perform as well, it was decided to keep the T1W phase in the final model. Therefore, in the final model, the information from the three phases was included. Then the RFE was applied. A total of 5 features were kept. The selected

features were Portal D, Portal ND, T1D, T1ND and T2ND, where Portal D and Portal ND are the probabilities obtained for the Portal phase for the desmoplastic and non-desmoplastic classes respectively, T1 D and T1 ND are the probabilities obtained for the desmoplastic and non-desmoplastic classes of T1W phase and T2 ND is the probability obtained for the non-desmoplastic class of T2W phase.

Confusion matrices for training and test data using the SVM classifier can be visualised in figure 5.7. The results of the assessment metrics tabulated in 5.4 were obtained from the matrices.



(a) Train **(b)** Test

Figure 5.7: Confusion matrix for the training and test data of the final model.

Table 5.4: Results of the evaluation metrics for the final model with the SVM classifier for the training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.97	0.91	0.94	D	0.83	0.56	0.67
ND	0.91	0.97	0.94	ND	0.64	0.88	0.74

The accuracy of this classifier was 94% for the training data and 71% for the test data.

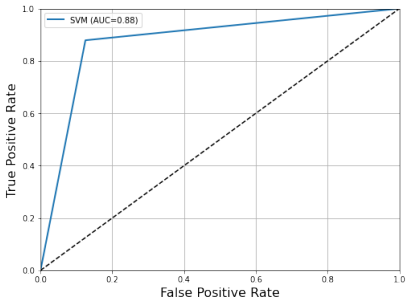


Figure 5.8: ROC Curve and AUC for the training set in the final model.

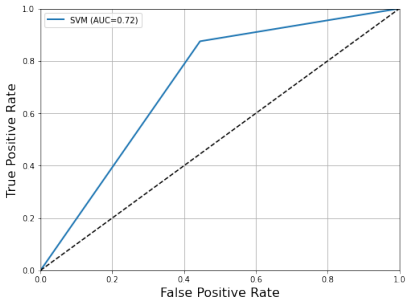


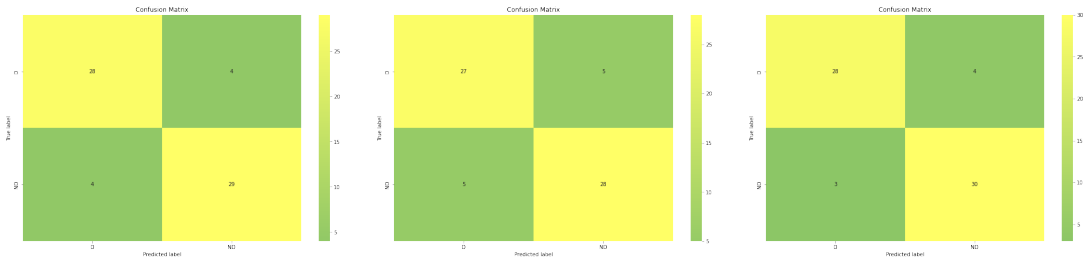
Figure 5.9: ROC Curve and AUC for the test set in the final model.

Figure 5.8 shows an AUC of 0.88 for the training data and Figure 5.9 shows an AUC of 0.72 for the test data.

5.1.3.1 Phases redundancy

The impact of the information from the phases on the final model and possible redundancies were then tested. However, it is noteworthy that while the T1W phase had the worst classification performance, the RFE algorithm did not exclude the information from this phase neither in the Portal with T1W model nor the T1W with T2W model, thus having an impact on these classifiers.

In all models, RFE selected an optimal number of 3 features, and for the Portal with T1W model, the selected features were Portal D, T1 D and T1 ND. For the Portal with T2W model, the selected features were Portal D, Portal ND, T2 ND and finally, for the T1W with T2W model, the selected features were T1 D, T1 ND and T2 ND.



(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 5.10: Confusion matrix for the training data with the SVM classifier.

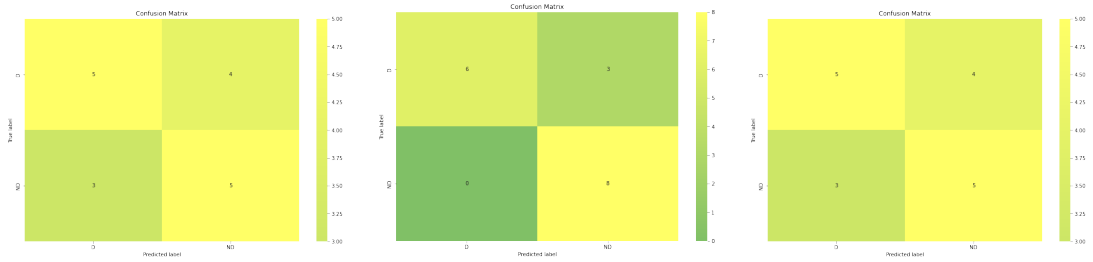
From the matrices shown in image 5.10, the results for the metrics listed in Table 5.6 were obtained.

The accuracy considering the training data for the Portal with T1W model is

Table 5.5: Results of the evaluation metrics for the final models with the SVM classifier for the training data.

PortalwithT1W				PortalwithT2W				T1WwithT2W			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.88	0.88	0.88	D	0.84	0.84	0.84	D	0.90	0.88	0.89
ND	0.88	0.88	0.88	ND	0.85	0.85	0.85	ND	0.88	0.91	0.90

88%, for the Portal with T2W model it is 85% and for the T1W with T2W model it is 89%.



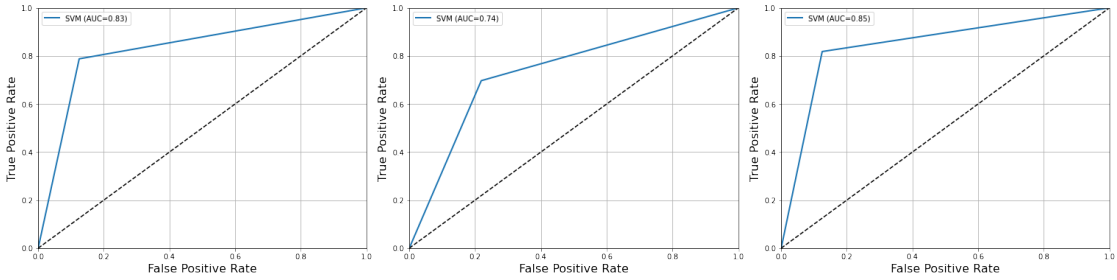
(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 5.11: Confusion matrix for the test data with the SVM classifier.

Table 5.6: Results of the evaluation metrics for the final models with the SVM classifier for the test data.

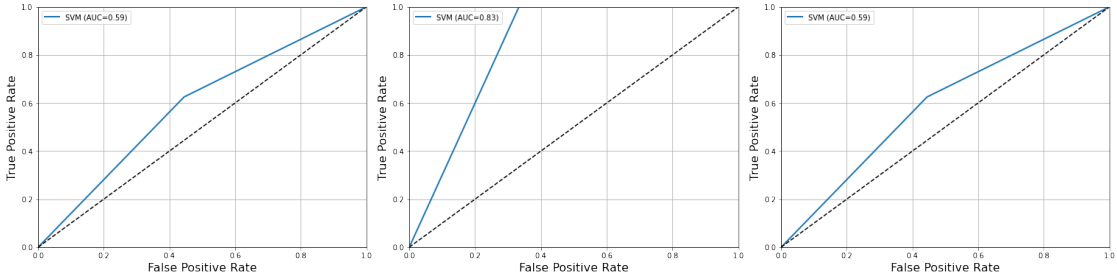
PortalwithT1W				PortalwithT2W				T1WwithT2W			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.62	0.56	0.59	D	1.00	0.67	0.80	D	0.62	0.56	0.59
ND	0.60	0.75	0.67	ND	0.73	1.00	0.84	ND	0.56	0.62	0.59

The accuracy considering the test data for the Portal with T1W model and T1W with T2W model is 59%, for the Portal with T2W model it is 82%. Figure 5.13 shows that the Portal with T2W model has the best ROC curve with an AUC of 0.83 for the test data.



(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 5.12: ROC Curves and respective AUC for training data.



(a) Portal with T1W. **(b)** Portal with T2W. **(c)** T1W with T2W.

Figure 5.13: ROC curves and respective AUC for test data.

From the results, the Portal with T2W model exhibits an excellent classification capacity.

5.1.4 Conclusions

Of all the phases presented, the Portal phase was the one that scored highest in the classification of classes, individually. The T1W phase was the one that performed worst in the classification. However, it is important to emphasise that only a small amount of data was considered and that the model could perform better if more data were added.

If the T1W phase is not included in the final model, i.e. a model that only includes information from the Portal and T2W phases, one obtains a model with a very satisfactory classification capacity that gives much better results. In summary, in these data, considering only two classes and ignoring the categorical features, the best final model for classification is the one considering information from images of phases Portal and T2W.

5.2 Model with categorical features

In this section, the workflow shown in Figure 3.9 from Chapter 3 was used. After applying mRMR and RFE, the two categorical features *size* and *lesion size* were added to check whether their information improves or degrades the model.

5.2.1 Model evaluation

For the models of each phase, it was found that the results were generally considered reasonable. For the Portal phase, an f1-score of 0.73 for the training data and 0.75 for the test data was obtained for class D using SVM. The AUC of the ROC curve was 0.71 for the training data and 0.77 for the test data. The accuracy was approximately 0.72 for the training data and 0.76 for the test data.

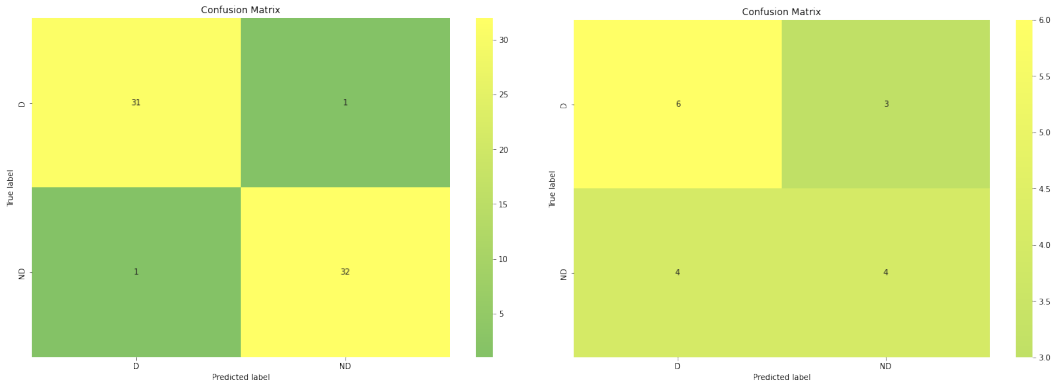
For the T1W phase, the model had results close to those of a random model. The f1-score was 0.81 for the class D training data when the Random Forest was evaluated, but when the test data was applied this value drops to 0.59. The AUC of the ROC curve also shows similar behaviour, with an AUC of 0.71 obtained for the training data and an AUC of 0.59 for the test data. The accuracy for the training data was 0.78, while it was only 0.59 for the test data.

For the T2W phase model, the best results were obtained with the NBB algorithm. The f1-score for class D was 0.66 for the training data and 0.71 for the test data. The AUC for the training data was 0.66 and for the test data it was 0.71. Finally, the accuracy was approximately 0.66 for the training data and 0.71 for the test data.

5.2.2 Final model

Although the individual phases with categorical features show relatively better results than the individual phases when the categorical features are disregarded, the evaluation of the final model yielded a considerably worse result. Confusion matrices for training and test data using the SVM classifier can be visualised in 5.14.

The accuracy for the training data was 0.97 and for the test data 0.59. From the results obtained, it is clear that the model for the test data has much lower performance when compared to the training data.



(a) Train **(b)** Test

Figure 5.14: Confusion matrix for the training and test data of the final model.

Table 5.7: Results of the evaluation metrics for the final model with the SVM classifier for the training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.97	0.97	0.97	D	0.60	0.67	0.63
ND	0.97	0.97	0.97	ND	0.57	0.50	0.53

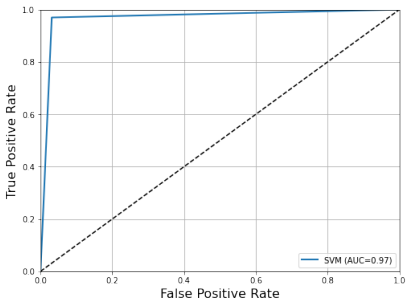


Figure 5.15: ROC Curve and AUC for the training set in the final model.

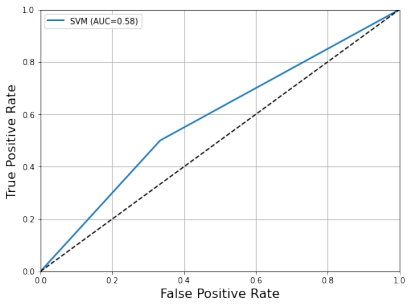
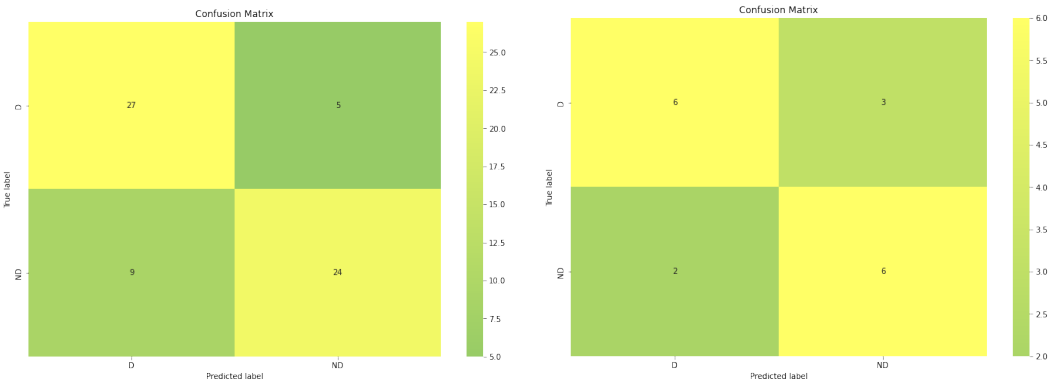


Figure 5.16: ROC Curve and AUC for the test set in the final model.

5.2.2.1 Phases redundancy

The weight of each phase in the final model was reviewed. As with the model without categorical features, the best results were obtained when the information from the T1W phase was disregarded. The results for the model containing only information from the Portal phase with the T2W phase can be seen in Figures 5.17 and 5.19 and in the table 5.8. The accuracy for this model for the training data was 0.78 and for the test data 0.71.



(a) Train **(b)** Test

Figure 5.17: Confusion matrix for the training and test data of the Portal with T2W model.

Table 5.8: Results of the evaluation metrics for the Portal with T2W model with the SVM classifier for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
D	0.75	0.84	0.79	D	0.75	0.67	0.71
ND	0.83	0.73	0.77	ND	0.67	0.75	0.71

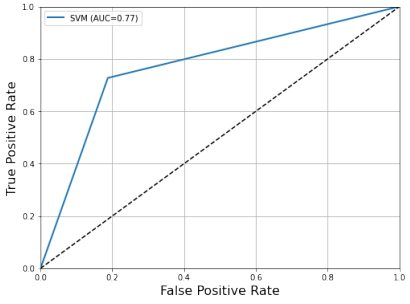


Figure 5.18: ROC curve for the training data of the Portal with T2W model.

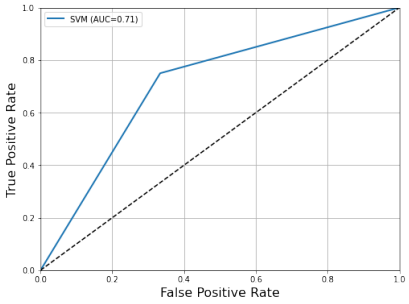


Figure 5.19: ROC curve for the test data of the Portal with T2W model.

5.2.3 Conclusions

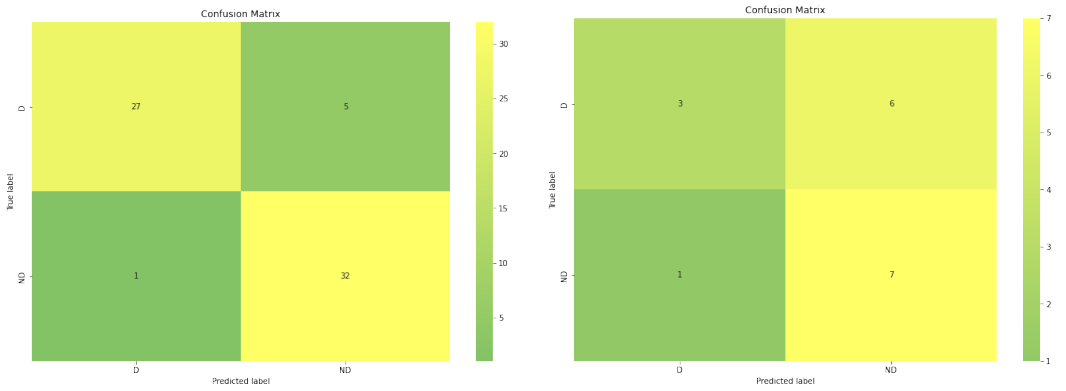
The results indicate that the final model has an acceptable classification capacity for class D, if the T1W phase is disregarded.

5.3 Leave-One-Out Cross-Validation

In cross validation, there is a special type called Leave One Out Cross Validation (LOOCV) where the number of iterations is exactly equal to the number of instances in the dataset [93]. In each iteration, the test set is only one data instance, and the training set includes all other instances. In this case, the same processes as described in Chapter 3 were carried out, using LOOCV instead of 7-fold cross validation.

Initially, the models were evaluated without considering the categorical features. Based on the individual results, it was found that the Portal phase with the SVM classifier produced the best results, with an f1-score of 0.75 for the training data and of 0.67 for the test data for class D. In addition, the AUC of the ROC curve was 0.77 for the training data and 0.65 for the test data. The T1W phase produced the worst results. In this phase, LR was used and it resulted in an f1-score of 0.81 for the training data, but for the test data it was only 0.56 for class D. The AUC of the training data was 0.80 and for the test data it was only 0.53.

The final model that accounted for the T1W phase showed an f1-score of only 0.38 for class D with the test data and an AUC of only 0.42. Removing the T1W phase improved the f1-score for class D and the AUC of the test data slightly to 0.46 and 0.60, respectively.



(a) Train **(b)** Test

Figure 5.20: Confusion matrix for the training and test data of the Portal with T2W model without categorical features.

The introduction of categorical features was then tested and was found to have a positive effect on the models. Again, the best results for each phase were obtained with the Portal phase, with an f1-score of 0.75 for the test data in class D using the SVM algorithm and an AUC of 0.77. The worst results were also obtained with the T1W phase, with an f1-score of 0.47 for class D with the test data using the LR and an AUC of only 0.47.

Table 5.9: Results of the evaluation metrics for the Portal with T2W model with the SVM classifier for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-score	Class	Precision	Recall	F1-score
D	0.96	0.84	0.90	D	0.75	0.33	0.46
ND	0.86	0.97	0.91	ND	0.54	0.88	0.67

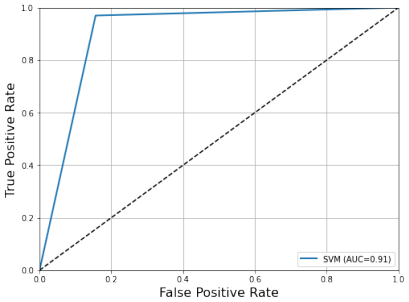


Figure 5.21: ROC curve for the training data of the Portal with T2W model without categorical features.

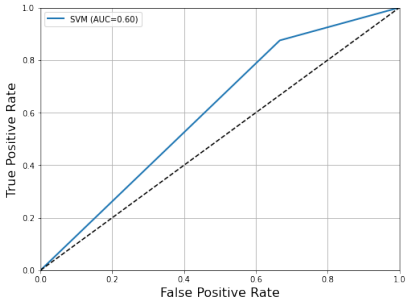


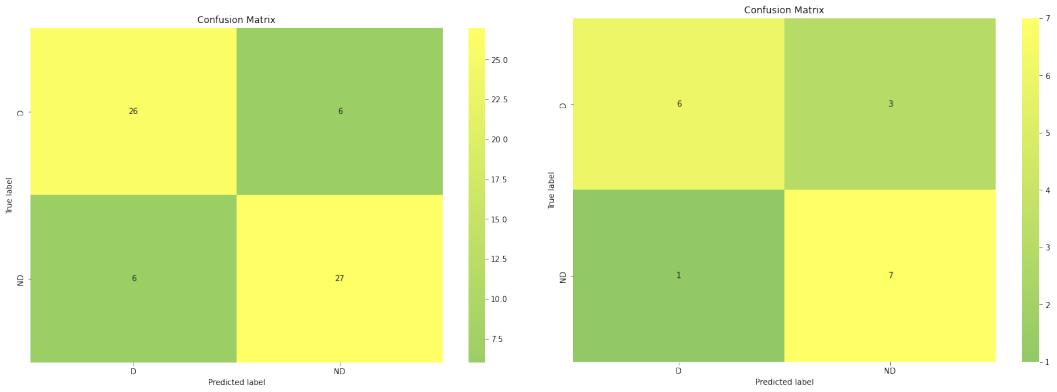
Figure 5.22: ROC curve for the test data of the Portal with T2W model without categorical features.

The final model considering the three phases showed an f1-score of 0.47 for class D and an AUC of 0.47 with the test data. However, when the T1W phase is removed, the results are much better, with an f1-score of 0.75 for class D with the test data and an AUC of 0.77.

Table 5.10: Results of the evaluation metrics for the Portal with T2W model with the SVM classifier for training and test data.

Train set				Test set			
Class	Precision	Recall	F1-score	Class	Precision	Recall	F1-score
D	0.81	0.81	0.81	D	0.86	0.67	0.75
ND	0.82	0.82	0.82	ND	0.70	0.88	0.78

The accuracy for the Portal with T2W model was 0.82 for the training and 0.76 for the test, while this value for the Portal with T2W model without taking the categorical features into account was 0.91 for the training but only 0.59 for the test data.



(a) Train **(b)** Test

Figure 5.23: Confusion matrix for the training and test data of the Portal with T2W model with categorical features.

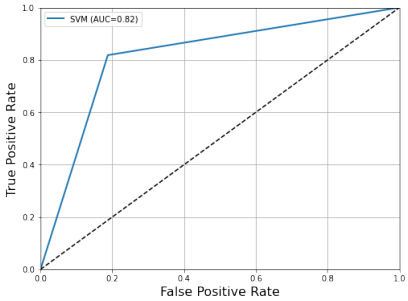


Figure 5.24: ROC curve for the training data of the Portal with T2W model with categorical features.

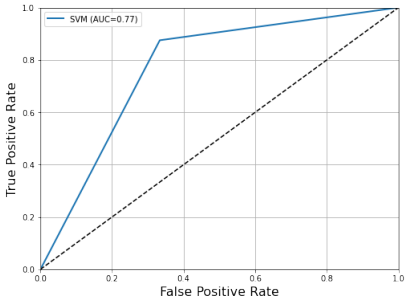


Figure 5.25: ROC curve for the test data of the Portal with T2W model with categorical features.

5.4 General conclusions

From the results presented in this chapter, it appears that it is possible to create a model that can satisfactorily predict the desmoplastic growth pattern from an image. The model disregarding the categorical features and using information from only Portal and T2W phases as input yielded an f1-score of 0.84 for the training data and 0.80 for the test data. In addition, the training data had an AUC of 0.83. However, for the model to adapt better or as satisfactorily to new data, more data is required.

Acquiring more data is also interesting for building models that can reliably predict the R or P class, since using only the available data will obtain a model with very high but unreliable results.

In the case of a classifier of class P, for example, out of 82 lesions only 10 belong to this class. Therefore, the classifier will have an excellent ability to predict the

non P class (NP), resulting in extremely high values for precision, recall, etc. for this class, but will probably be close to zero for the P class.

In the Portal, T1W and T2W phases, which were used in the model with the best results, it was found that after applying the RFE algorithm, most of the selected features were of the texture type. In the portal phase 7 out of 9 features are texture features, in the T1W phase 7 out of 10 features are texture features and in the T2W phase 5 out of 8 features are texture features. This is to be expected as the grey levels express the different textures between healthy and tumour tissue and are therefore extremely important for identifying growth patterns. The biological interpretability of the features should as maximised as possible, to allow users (clinicians and technicians) not only to rely on these approaches but also to guide intervention strategies upon the most informative features.

6

Final conclusions and future work

The aim of this project was to develop a model that can predict the histopathological growth pattern of liver metastases in colorectal cancer from a magnetic resonance image. As the results presented in this thesis show, some models were successful in predicting the desmoplastic growth pattern, such as the model containing Portal with T2W information, without considering categorical features, which had an f1-score of 0.80 for class D, an AUC of 0.83 and an accuracy of 0.82 in the test data.

Moreover, the models built with information of the portal phase individually seemed to give reasonable results in most cases, compared to the models considering features extracted from the T1W phase, which resembled a random model in many cases. The reason that the portal phase gives better results could be the same reason why radiologists visualise metastases better in this phase, namely the contrast in the image. However, this is only a guess, because it is not possible to conclude that good results for class D classification are always obtained in this phase, as the data sample is small and therefore more studies should be performed. It was also found that the categorical features influence the model to some extent.

For future work, it would be interesting to test what weight the individual categorical features have in the final model and which of the information (histological or radiological) tends to influence the final model positively or negatively. Furthermore, an investigation of the weight of each radiomics feature in classification and its interpretability should also be addressed.

As mentioned in the methodology chapter, the database also included patients who were undergoing chemotherapy. The creation of a feature that takes this information into account and the re-testing of the model would therefore be clinically relevant and technically plausible, as chemotherapy affects the liver structures [94, 95] and may consequently affect the extracted information (feature extraction).

In order for the models presented here to better generalise, more data needs to be acquired, or in alternative these models could be developed relying upon a

database of synthetic data.

Glossary

ADC Apparent diffusion coefficient

AI Artificial intelligence

AP Arterial phase

AUC Area under the curve

CE-MRI Contrast-enhanced MRI

CRC Colorectal cancer

CRCLM Colorectal cancer liver metastases

CRLM Colorectal liver metastases

CT Computed tomography

D Desmoplastic

DCE Dynamic contrast-enhanced

dHGP Desmoplastic histological growth patterns

DICOM Digital Imaging and Communications in Medicine

DT Decision trees

DWI Diffusion-weighted imaging

FLAIR Fluid attenuated inversion recovery

FN False negative

FNR False negative rate

FP False positive

FPR False positive rate

FS Fat-suppressed

GLCM Grey level co-occurrence matrix

GLDM Grey level dependence matrix

GLRLM Grey level run length matrix

GLSZM Grey level size zone matrix

H&E Hematoxylin and Eosin

HGP Histological growth pattern

IBSI Image biomarker standardisation initiative

LoG Laplacian of Gaussian

LR Logistic regression classifier

MDCT Contrast-enhanced multidetector CT

MLM Metachronous liver metastasis

MLP Multi-layer perceptron

MR Magnetic resonance

MRI Magnetic resonance imaging

mRMR Minimum redundancy maximum relevance

NAC Neoajuvant chemotherapy

NBB Naive Bayes Bernoulli

NBG Naive Bayes Gaussian

NGLDM Neighbouring grey level dependence matrix

NGTDM Neighbourhood grey tone difference matrix

NPV Negative predictive value

P Pushing

PCA Principal component analysis

PET Positron emission tomography

PFS Progression-free survival

pHGP Push histological growth patterns

PVP Portal venous phase

R Replacement

RFC Random forest classifier

RFE Recursive feature elimination

RFP Radiofrequency pulse

rHGP Replacement histological growth patterns

ROC Receiver operating characteristic

ROI Region of interest

SLM Synchronous liver metastases

SVM Support vector machine

TE Time of echo

TLI Tumour-liver interface

TN True negative

TNR True negative rate/Specificity

TP True positive

TPR True positive rate/Recall

TR Time of repetition

T1W T1-weighted

T2W T2-weighted

VAR Variance

VAT Visceral adipose tissue

VOI Volume of interest

Bibliography

[1] F. Sanchez-Vega, M. Mina, J. Armenia, *et al.*, *"Oncogenic Signaling Pathways in The Cancer Genome Atlas."* Cell **173** (2018), 10.1016/j.cell.2018.03.035.

[2] RJ. Werner, AD. Kelly, and JJ. Issa, *"Epigenetics and Precision Oncology."* The Cancer Journal **23** (2017), 10.1097/PPO.0000000000000281.

[3] H. Khan, N.and Mukhtar, *"Cancer and metastasis: prevention and treatment by green tea."* Cancer Metastasis Rev 29 (2010), https://doi.org/10.1007/s10555-010-9236-1.

[4] Rabia Zeeshan and Zeeshan Mutahir, *"Cancer metastasis - tricks of the trade,"* Bosnian Journal of Basic Medical Sciences **17**, 172–182 (2017).

[5] AW. Lambert, DR. Pattabiraman, and RA. Weinberg, *"Emerging Biological Principles of Metastasis."* Cell **168** (2017), 10.1016/j.cell.2016.11.037.

[6] E. Dekker, PJ. Tanis, JLA. Vleugels, PM. Kasi, and MB. Wallace, *"Colorectal cancer,"* The Lancet (2019), 10.1016/S0140-6736(19)32319-0.

[7] D. Falcão, H. Alexandrino, R. Caetano Oliveira, *et al.*, *"Histopathologic patterns as markers of prognosis in patients undergoing hepatectomy for colorectal cancer liver metastases - Pushing growth as an independent risk factor for decreased survival."* European Journal of Surgical Oncology **44** (2018), 10.1016/j.ejso.2018.03.023.

[8] RC. Oliveira, H. Alexandrino, MA. Cipriano, and JG Tralhão, *"Liver Metastases and Histological Growth Patterns: Biological Behavior and Potential Clinical Implications-Another Path to Individualized Medicine?"* Journal of Oncology (2019), 10.1155/2019/6280347.

[9] M. Marra, A. Giudice, C. Arra, *et al.*, *"Target-based agents in neo-adjuvant treatment of liver metastases from colo-rectal cancer: Secret weapons in anti-cancer war?"* Cancer Biology & Therapy **8**, 1709–1718 (2009), pMID: 19729997.

[10] F. Fiz, l. vigano, N. Gennaro, *et al.*, *"Radiomics of liver metastases: A systematic review,"* Cancers (2020), 10.3390/cancers12102881.

[11] RC. Oliveira, H. Alexandrino, MA. Cipriano, FC. Alves, and JG Tralhão, *"Predicting liver metastases growth patterns: Current status and future possibilities,"* Semin Cancer Biol (2021), 10.1016/j.semcancer.2020.07.007.

[12] JB. Wu, AL. Sarmiento, PO. Fiset, *et al.*, *"Histologic features and genomic alterations of primary colorectal adenocarcinoma predict growth patterns of liver metastasis."* World J Gastroenterol. **25** (2019), 10.3748/wjg.v25.i26.3408.

[13] Y. Han, F. Chai, J. Wei, *et al.*, *"Identification of predominant histopathological growth patterns of colorectal liver metastasis by multi-habitat and multi-sequence based radiomics analysis,"* Frontiers in Oncology **10**, 1363 (2020).

[14] A. Rigamonti, F. Feuerhake, M. Donadon, M. Locati, and F. Marchesi, *"Histopathological and immune prognostic factors in colo-rectal liver metastases,"* Cancers **13** (2021), 10.3390/cancers13051075.

[15] E. Latacz, PJ. van Dam, C. Vanhove, *et al.*, *"Can medical imaging identify the histopathological growth patterns of liver metastases?"* Seminars in Cancer Biology **71**, 33–41 (2021), metastasis to the Liver: From Pathobiology through Therapies.

[16] G. Garcia-Vicién, A. Mezheyeuski, M. Bañuls, N. Ruiz-Roig, and DG Molleví, *"The Tumor Microenvironment in Liver Metastases from Colorectal Carcinoma in the Context of the Histologic Growth Patterns."* Int J Mol Sci **22** (2021), 10.3390/ijms22041544.

[17] S. Keshav, *The Gastrointestinal System, 1rd ed.*, at a Glance (Blackwell Science Ltd, 2004) p. 85.

[18] K. Thanikachalam and G Khan, *"Colorectal cancer and nutrition,"* Nutrients **11** (2019), 10.3390/nu11010164.

[19] F. Grizzi, P. Bianchi, A. Malesci, and L. Laghi, *"Prognostic value of innate and adaptive immunity in colorectal cancer."* World J Gastroenterol **19** (2013), 10.3748/wjg.v19.i2.174.

[20] JJ. GRANADOS ROMERO *et al.*, *" Colorectal cancer: a review."* International Journal of Research in Medical Sciences **5** (2017), http://dx.doi.org/10.18203/2320-6012.ijrms20174914.

[21] M. Arnold, MS. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, *"Global patterns and trends in colorectal cancer incidence and mortality."* Gut **66** (2017), 10.1136/gutjnl-2015-310912.

[22] Cotter J., *"Colorectal cancer: Portugal and the world."* Acta Med Port **26** (2013).

[23] HH. Backus, CJ. Van Groeningen, W. Vos, *et al.*, *" Differential expression of cell cycle and apoptosis related proteins in colorectal mucosa, primary colon tumours, and liver metastases."* J Clin Pathol **55** (2002), 10.1136/jcp.55.3.206.

[24] AI. Valderrama-Treviño, B. Barrera-Mera, JC. Ceballos-Villalva, and EE Montalvo-Javé, *"Hepatic metastasis from colorectal cancer."* Euroasian J Hepatogastroenterol. **7** (2017), 10.5005/jp-journals-10018-1241.

[25] I. Mármol, C. Sánchez-de Diego, A. Pradilla Dieste, E. Cerrada, and MJ. Rodriguez Yoldi, *"Colorectal carcinoma: A general overview and future perspectives in colorectal cancer,"* International Journal of Molecular Sciences **18** (2017), 10.3390/ijms18010197.

[26] SC. Chuang, YC. Su, CY. Lu, *et al.*, *"Risk factors for the development of metachronous liver metastasis in colorectal cancer patients after curative resection,"* World J Surg **35** (2011), 10.1007/s00268-010-0881-x.

[27] AD. Karaosmanoglu, MR. Onur, MN. Ozmen, D. Akata, and M. Karcaaltincaba, *"Magnetic Resonance Imaging of Liver Metastasis,"* Seminars in Ultrasound, CT and MRI **37** (2016), https://doi.org/10.1053/j.sult.2016.08.005.

[28] LN. Vu, JN. Morelli, and J. Szklaruk, *"" basic mri for the liver oncologists and surgeons.","* J Hepatocell Carcinoma. **5** (2018), 10.2147/JHC.S154321.

[29] GB. Chavhan, L. Farras Roca, and AC. Coblentz, *""liver magnetic resonance imaging: how we do it.","* Pediatr Radiol. **52** (2022), 10.1007/s00247-021-05053-4.

[30] KJ. Lafaro, P. Roumanis, AN. Demirjian, C. Lall, and DK. Imagawa, *"Gd-EOB-DTPA-Enhanced MRI for Detection of Liver Metastases from Colorectal Cancer: A Surgeon's Perspective!."* Int J Hepatol. **2013** (2013), 10.1155/2013/572307.

[31] DC. Osei-Bordom, S. Kamarajah, and N. Christou, *"Colorectal cancer, liver metastases and biotherapies,"* Biomedicines **9** (2021), 10.3390/biomedicines9080894.

[32] J. Engstrand, H. Nilsson, C. Strömberg, E. Jonas, and J. Freedman, *"Colorectal cancer liver metastases - a population-based study on incidence, management and survival."* BMC Cancer **18** (2018), 10.1186/s12885-017-3925-x.

[33] J. Cheng, J. Wei, T. Tong, *et al.*, *"Prediction of Histopathologic Growth Patterns of Colorectal Liver Metastases with a Noninvasive Imaging Method."* Journal of Oncology **26** (2019), 10.1245/s10434-019-07910-x.

[34] A. Biondi, R. Persiani, F. Cananzi, *et al.*, *"R0 resection in the treatment of gastric cancer: room for improvement,"* World J Gastroenterol **16** (2010), 10.3748/wjg.v16.i27.3358.

[35] S. Ishiguro, T. Akasu, Y. Fujimoto, *et al.*, *" Second hepatectomy for recurrent colorectal liver metastasis: analysis of preoperative prognostic factors."* Ann Surg Oncol **13** (2006), https://doi.org/10.1245/s10434-006-9067-z.

[36] AWC. Kow, *"Hepatic metastasis from colorectal cancer."* J Gastrointest Oncol. **10** (2019), 10.21037/jgo.2019.08.06.

[37] RJ. Aragon and NL. Solomon, *"Techniques of hepatic resection."* J Gastrointest Oncol. **3** (2012), 10.3978/j.issn.2078-6891.2012.006.

[38] RS. Puijk, AH. Ruarus, LGPH. Vroomen, *et al.*, *" Colorectal liver metastases: surgery versus thermal ablation (COLLISION) - a phase III single-blind prospective randomized controlled trial."* BMC Cancer. **18** (2018), 10.1186/s12885-018-4716-8.

[39] H. Takahashi and E. Berber, *"Role of thermal ablation in the management of colorectal liver metastasis."* Hepatobiliary Surg Nutr. **9** (2020), 10.21037/hbsn.2019.06.08.

[40] R. Barnhill, P. Vermeulen, S. Daelemans, *et al.*, *" Replacement and desmoplastic histopathological growth patterns: A pilot study of prediction of outcome in patients with uveal melanoma liver metastases."* J Pathol Clin **4** (2018), 10.1002/cjp2.105.

[41] OM. Jones, M. Rees, TG. John, S. Bygrave, and G. Plant, *"Biopsy of resectable colorectal liver metastases causes tumour dissemination and adversely affects survival after liver resection."* Br J Surg. **92** (2005), 10.1002/bjs.4888.

[42] MS. Rodgers, R. Collinson, S. Desai, RS. Stubbs, and JL. McCall, *" Risk of dissemination with biopsy of colorectal liver metastases."* Dis Colon Rectum. **46** (2003), 10.1007/s10350-004-6581-6.

[43] P. Savadjiev, J. Chong, A. Dohan, and others., *"Demystification of AI-driven medical image interpretation: past, present and future."* Eur Radiol. **29** (2019), 10.1007/s00330-018-5674-x.

[44] H. Ding, C. Wu, N. Liao, and others., *"Radiomics in Oncology: A 10-Year Bibliometric Analysis."* Front Oncol. (2021), 10.3389/fonc.2021.689802.

[45] S. Chen, Z. Shu, Y. Li, *et al.*, *"Machine Learning-Based Radiomics Nomogram Using Magnetic Resonance Images for Prediction of Neoadjuvant Chemotherapy Efficacy in Breast Cancer Patients."* Front Oncol **10** (2020), 10.3389/fonc.2020.01410.

[46] JE. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, *"Radiomics in medical imaging-"how-to" guide and critical reflection."* Insights Imaging **11** (2020), 10.1186/s13244-020-00887-2.

[47] James C. Korte, Carlos. Cardenas, Nicholas. Hardcastle, Tomas. Kron, Jihong. Wang, Houda. Bahig, *et al.*, *"Radiomics feature stability of open-source software evaluated on apparent diffusion coefficient maps in head and neck cancer."* Scientific Reports **11** (2021), https://doi.org/10.1038/s41598-021-96600-4.

[48] V. Granata, R. Fusco, and ML.and others Barretta, *"Radiomics in hepatic metastasis by colorectal cancer."* Infect Agent Cancer **16** (2021), 10.1186/s13027-021-00379-y.

[49] V. Kumar, Y. Gu, S. Basu, *et al.*, *"Radiomics: the process and the challenges,"* Magnetic Resonance Imaging **30**, 1234–1248 (2012), quantitative Imaging in Cancer.

[50] YQ. Huang, CH. Liang, L. He, and others., *"Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer [published correction appears in J Clin Oncol."* J Clin Oncol. **34** (2016), 10.1200/JCO.2015.65.9128.

[51] M. Avanzo, J. Stancanello, and I. El Naqa, *"Beyond imaging: The promise of radiomics,"* Phys Med **38**, 122–139 (2017).

[52] A. Zwanenburg, M. Vallières, MA. Abdalah, *et al.*, *"Image biomarker standardisation initiative - feature definitions,"* CoRR **abs/1612.07003** (2016), 1612.07003 .

[53] N. Papanikolaou, C. Matos, and DM. Koh, *"How to develop a meaningful radiomic signature for clinical use in oncologic patients."* Cancer Imaging. **20** (2020), 10.1186/s40644-020-00311-4.

[54] Y. Di Re, AM. ans Sun, P. Sundaresan, *et al.*, *"MRI radiomics in the prediction of therapeutic response to neoadjuvant therapy for locoregionally advanced rectal cancer: a systematic review,"* Expert Review of Anticancer Therapy **21**, 425–449 (2021), pMID: 33289435.

[55] M. Amadasun and R. King, *"Textural features corresponding to textural properties,"* IEEE Transactions on Systems, Man, and Cybernetics **19**, 1264–1274 (1989).

[56] C. Sun and W. G Wee, *"Neighboring gray level dependence matrix for texture classification,"* Computer Vision, Graphics, and Image Processing **23**, 341–352 (1983).

[57] JJM. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, V. Aucoin, N.and Narayan, RGH. Beets-Tan, JC. Fillon-Robin, S. Pieper, and HJWL. Aerts, *"'computational radiomics system to decode the radiographic phenotype.',"* Cancer Research **77**, e104–e107 (2017).

[58] P. Lambin, RTH. Leijenaar, and TM.and others Deist, *"Radiomics: the bridge between medical imaging and personalized medicine."* Nat Rev Clin Oncol **14** (2017), 10.1038/nrclinonc.2017.141.

[59] Hoque. N, Bhattacharyya. D, K, and Kalita. J, K, *"MIFS-ND: A mutual information-based feature selection method,"* Expert Systems with Applications **41**, 6371–6385 (2014).

[60] Radovic. M, M Ghalwash., Filipovic N, and Obradovic. Z, *"Minimum redundancy maximum relevance feature selection approach for temporal gene expression data."* BMC Bioinformatics. **18** (2017), 10.1186/s12859-016-1423-9.

[61] Zhenyu. Z, Radhika. A, and Mallory. W, in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2019) pp. 442–452.

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

[63] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and HJWL. Aerts, *"Machine Learning methods for Quantitative Radiomic Biomarkers."* Sci Rep **5** (2015), 10.1038/srep13087.

[64] T. Hastie, R. Tibshirani, and J Friedman, *The Elements of Statistical Learning, 2nd ed.* (Springer, 2017).

[65] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensor Flow, 2nd ed.* (O'Reilly Media, 2019).

[66] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining, 2nd ed.* (Springer, 2017).

[67] P. Dangeti, *Statistics for Machine Learning* (Packt Publishing, 2017).

[68] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning, 2nd ed.* (Springer).

[69] M. Deisenroth, A. Faisal, and C. Ong, *Mathematics For Machine* (Cambridge University Press, 2020).

[70] J. Zhi, J. Sun, Z. Wang, and W. Ding, *"'support vector machine classifier for prediction of the metastasis of colorectal cancer.',"* Int J Mol Med **41** (2018), 10.3892/ijmm.2018.3359.

[71] S. Huang, N. Cai, PP. Pacheco, S. Narrandes, Y. Wang, and W. Xu, *"'applications of support vector machine (svm) learning in cancer genomics. ',"* Cancer Genomics Proteomics. **15** (2018), 10.21873/cgp.20063.

[72] S. Xu, *"Bayesian naive bayes classifiers to text classification,"* Journal of Information Science **44** (2016), 10.1177/0165551516677946.

[73] K. Vembandasamy, R. Sasipriya, and E. DeepaP, *"Heart diseases detection using naive bayes algorithm,"* IJISET - International Journal of Innovative Science, Engineering Technology **2** (2015), https://ijiset.com/vol2/v2s9/IJISET$_V$2$_I$9$_5$4.*pdf*.

[74] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, *"Comparison between multinomial and bernoulli naïve bayes for text classification,"* , 593–596 (2019).

[75] P. Kulkarni, *Reinforcement and Systemic Machine Learning for Decision Making* (Institute of Electrical and Electronics Engineers, Inc., 2012).

[76] B. Jijo and A. Mohsin Abdulazeez, *"Classification Based on Decision Tree Algorithm for Machine Learning,"* Journal of Applied Science and Technology Trends **2**, 20–28 (2021).

[77] C. Archana, K. Savita, and K Raj, *"'an improved random forest classifier for multi-class classification',"* Information Processing in Agriculture **3**, 215–222 (2016).

[78] A. Parmar, R. Katariya, and V. Patel, in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, edited by J. Hemanth, X. Fernando, P. Lafata, and Z. Baig (Springer International Publishing, Cham, 2019) pp. 758–763.

[79] AC. Lorena, L. Jacintho, M. Siqueira, R. De Giovanni, L. Lohmann, A. de Carvalho, and M. Yamamoto, *"Comparing machine learning classifiers in potential distribution modelling,"* Expert Systems with Applications **38**, 5268–5275 (2011).

[80] S. Abirami and P Chitra, in *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, Advances in Computers, Vol. 117, edited by Pethuru Raj and Preetha Evangeline (Elsevier, 2020) pp. 339–368.

[81] Z. Shu, S. Fang, Z. Ding, *et al.*, *"MRI-based Radiomics nomogram to detect primary rectal cancer with synchronous liver metastases."* Sci Repl **9** (2019), 10.1038/s41598-019-39651-y.

[82] AS Becker, MA. Schneider, MC. Wurnig, M. Wagner, PA. Clavien, and A. Boss, *"Radiomics of liver MRI predict metastases in mice,"* Eur Radiol **2** (2018), 10.1186/s41747-018-0044-7.

[83] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding, in *Advances in Neural Information Processing Systems*, Vol. 23, edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Curran Associates, Inc., 2010).

[84] AA. Grosset, K. Loayza-Vega, É. Adam-Granger, and others., *"\"hematoxylin and eosin counterstaining protocol for immunohistochemistry interpretation and diagnosis.\","* Appl Immunohistochem Mol Morphol. **27** (2019), 10.1097/PAI.0000000000000626.

[85] AH. Fischer, KA. Jacobson, J. Rose, and R. Zeller, *"\"hematoxylin and eosin staining of tissue and cell sections.\","* CSH Protoc. (2008), 10.1101/pdb.prot4986.

[86] M. Schwier, J. van Griethuysen, M. Vangel, S. Pieper, S. Peled, C. Tempany, H. Aerts, R. Kikinis, F. Fennessy, and A. Fedorov, *"Repeatability of multiparametric prostate mri radiomics features,"* Scientific Reports **9** (2019), 10.1038/s41598-019-45766-z.

[87] F. Tixier, CC. Le Rest, M. Hatt, and others., *"\"reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18f-fdg pet.\","* **53**, 693–700 (2012).

[88] B. Ganeshan, KA. Miles, RC. Young, and CR. Chatwin, *"\"texture analysis in non-contrast enhanced ct: impact of malignancy on texture in apparently disease-free areas of the liver.\","* Eur J Radiol. **70** (2009), 10.1016/j.ejrad.2007.12.005.

[89] The pandas development team, "pandas-dev/pandas: Pandas," (2020).

[90] Wes McKinney *et al.*, in *Proceedings of the 9th Python in Science Conference*, edited by Jarrod Millman Stéfan van der Walt.

[91] David Nettleton, *Commercial Data Mining*, edited by David Nettleton (Morgan Kaufmann, Boston, 2014).

[92] Jason Brownlee, "Recursive feature elimination (rfe) for feature selection in python," (2020).

[93] Claude Sammut and Geoffrey I. Webb, eds., "Leave-one-out cross-validation," in *Encyclopedia of Machine Learning* (Springer US, Boston, MA, 2010) pp. 600–601.

[94] L. Calistri, V. Rastrelli, C. Nardi, *et al.*, *“"imaging of the chemotherapy-induced hepatic damage: Yellow liver, blue liver, and pseudocirrhosis.",”* World J Gastroenterol. **27** (2021), 10.3748/wjg.v27.i46.7866.

[95] D. Albano, M. Benenati, A. Bruno, *et al.*, *“"imaging side effects and complications of chemotherapy and radiation therapy: a pictorial review from head to toe.",”* Insights Imaging. **12** (2021), 10.1186/s13244-021-01017-2.