UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Prostate MRI Radiomics for Prediction of Gleason Score

Ana Carolina Vitorino Rodrigues

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Nickolas Papanikolaou

Professor Doutor Francisco José Moreira Couto

2021

Acknowledgements

Firstly, I would like to thank the co-supervisor of this project Doctor Nickolas Papanikolaou for accepting me into this team, but especially for his invaluable input of knowledge, patience and confidence in my work throughout the entire duration of this project. Co-supervisor Professor Francisco Couto must also be thanked for his unwavering attention and continuous follow-up on the project.

Secondly, I would like to thank all the members of CCIG that, despite having no obligation to help, still managed to do so in any way they could. A special thanks to João Santinha and Bernardo Galvão for always being available to answer any burning questions.

I would also like to thank my family for the emotional and financial support they provided.

Finally, a special thanks to my dance family for keeping me sane.

Resumo

O cancro da próstata é um dos cancros mais prevalentes em Portugal, estando entre as 4 principais causas de morte por neoplasias em 2018, com uma taxa bruta de mortalidade de 38.23 mortes por 100 000 homens.

O atual diagnóstico e classificação do cancro da próstata não é ideal, baseando-se em medidas pouco específicas como os níveis de PSA e DRE, seguidos de biópsia, onde é atribuído um nível de agressividade sob a forma da classificação de Gleason. Foi demonstrado no passado que o exame de ressonância magnética multiparamétrica é útil na deteção de lesões de cancro da próstata. No entanto, a interpretação deste exame, sendo um processo subjetivo, está inevitavelmente afetada por uma elevada taxa de variabilidade entre observadores. Foi demonstrado também que a classificação de Gleason atribuída a uma lesão aquando da biópsia, irá provavelmente ser corrigida após prostatectomia radical. Portanto, um método confiável e de preferência não invasivo para classificação do cancro da próstata é necessário. Com este objetivo, esforços têm sido feitos no passado para usar radiómica e aprendizagem automática para prever a classificação de Gleason a partir de imagens clínicas, apresentando resultados promissores. Radiómica é a transformação de imagens médicas em dados quantitativos de alta dimensão. Assim, com base na hipótese de que as características do tumor que são causa ou consequência da classificação de Gleason estão refletidas nas variáveis radiómicas extraídas da imagem de ressonância magnética, estas podem ser usadas para construir modelos de aprendizagem automática capazes de avaliar este parâmetro. Dito isso, o objetivo principal deste trabalho foi desenvolver modelos de aprendizagem automática explorando variáveis radiómicas extraídas de exames de ressonância magnética para prever a agressividade biológica na forma de classificação de Gleason.

Neste trabalho, 288 modelos foram desenvolvidos, correspondendo a diferentes combinações de aspetos de uma *pipeline* típica, mais especificamente, origem dos dados de treino, estratégia de pre-processamento dos dados, método de seleção de variáveis e algoritmo de aprendizagem automática. Num conjunto de 281 lesões (210 para treino, 71 para validação) e 183 pacientes (137 para treino, 46 para validação), verificou-se que as variáveis radiómicas extraídas do VOI da glândula inteira produziram modelos extremamente mais confiáveis do que as variáveis radiómicas extraídas dos VOIs das lesões. Sugerindo que as áreas em volta das lesões tumorais oferecem informações relevantes sobre a classificação de Gleason que é atribuída a essa lesão. Além de sugerir que o trabalho monótono de segmentação das lesões realizado pelo radiologista pode não ser necessário ou mesmo prejudicar a assinatura radiómica.

Palavras Chave: Radiómica, Aprendizagem automática, Cancro da Próstata.

Abstract

Prostate cancer is one of the most prevalent cancers in Portugal, being among the top 4 malignant neoplasm causes of death in 2018, with a crude mortality rate of 38.23 deaths per 100 000 males.

Prostate cancer diagnosis and classification is not ideal, relying on unspecific measures such as PSA levels and DRE, followed by biopsy, where an aggressiveness level is attributed in the form of Gleason score. Multiparametric MRI has proven to be useful in the detection of prostate cancer. However, it is unavoidably affected by a high rate of inter-reader variability. It has also been shown that the Gleason score attributed to a lesion after biopsy is likely to change after radical prostatectomy.

Therefore, a reliable, and preferably non-invasive, method for classification of PCa is in urgent demand. With this goal in mind, efforts have been made in the past to use computer-aided diagnosis (CAD) coupled with radiomics and machine learning to predict Gleason score from clinical images, showing promising results.

Radiomics is the transformation of medical images into high dimension mineable data. Hence, based on the hypothesis that tumour characteristics that are cause or consequence of Gleason score are reflected in the radiomic features extracted from the MRI image, these can be used to build supervised machine learning models capable of assessing this parameter. That being said, the main goal of this work was to develop supervised machine learning models exploiting radiomic features extracted from mpMRI examinations, to predict biological aggressiveness in the form of Gleason Score.

In this work, 288 classifiers were developed, corresponding to different combinations of pipeline aspects, namely, type of input data (i.e. lesion features vs whole gland features), sampling strategy, feature selection method and machine learning algorithm.

On a cohort of 281 lesions (210 for training, 71 for validation) and 183 patients (137 for training, 46 for validation), it was found that radiomic features extracted from the whole gland VOI produced extremely more reliable classifiers than radiomic features extracted from the lesions' VOIs. Suggesting that the areas surrounding the tumour lesions offer relevant information regarding the Gleason Score that is ultimately attributed to that lesion. In addition to suggesting that the monotonous lesion segmentation work performed by radiologists may not be necessary or even be harming to the radiomics signature.

Keywords: Radiomics, Machine Learning, Prostate Cancer.

Resumo Alargado

O cancro da próstata é um dos cancros mais prevalentes em Portugal, estando entre as 4 principais causas de morte por neoplasias em 2018, com uma taxa bruta de mortalidade de 38.23 mortes por 100 000 homens.

O atual diagnóstico e classificação do cancro da próstata não é ideal, baseando-se em medidas pouco específicas como os níveis de PSA (antigénio específico da próstata) e DRE (examinação retal), seguidos de biópsia guiada por ultrasom trans-rectal (TRUS), onde é atribuído um nível de agressividade sob a forma da classificação de Gleason.

Ao contrário de TRUS, imagens de ressonância magnética permitem uma visualização clara da anatomia da próstata. mpMRI (ressonância magnética multiparamétrica) corresponde a um conjunto de diferentes métodos de captação de imagem que fornecem informação de perspetivas diferentes sobre o tecido, constituindo uma ferramenta promissora para a identificação de lesões tumorais e respetiva classificação. No entanto, a interpretação deste exame, sendo um processo subjetivo, está inevitavelmente afetada por uma elevada taxa de variabilidade entre observadores. Foi demonstrado também que a classificação de Gleason atribuída a uma lesão aquando da biópsia, irá provavelmente ser corrigida após prostatectomia radical.

Portanto, um método confiável e de preferência não invasivo para classificação do cancro da próstata é necessário. Com este objetivo em mente, esforços têm sido feitos no passado para usar radiómica e aprendizagem automática para prever a classificação de Gleason a partir de imagens clínicas, apresentando resultados promissores.

Radiómica é a transformação de imagens médicas em dados quantitativos de alta dimensão. Assim, com base na hipótese de que as características do tumor que são causa ou consequência da classificação de Gleason estão refletidas nas variáveis radiómicas extraídas das imagens de ressonância magnética multiparamétrica, estas podem ser usadas para construir modelos de aprendizagem automática capazes de avaliar este parâmetro. Dito isso, o objetivo principal deste trabalho foi desenvolver modelos de aprendizagem automática explorando variáveis radiómicas extraídas de exames de ressonância magnética multiparamétrica para prever a agressividade biológica na forma de classificação de Gleason.

Neste trabalho, 321 variáveis radiómicas foram extraídas por paciente ou lesão. Estas foram utilizadas no desenvolvimento de 288 modelos, correspondendo a diferentes combinações de aspetos de uma *pipeline* típica, mais especificamente, origem dos dados de treino (por exemplo, variáveis radiómicas da lesão vs variáveis radiómicas da glandula inteira), estratégia de pre-processamento dos dados,

método de seleção de variáveis e algoritmo de aprendizagem automática. O desempenho dos vários modelos foi avaliado através das métricas F2 e Cohen's Kappa e os modelos foram comparados entre si no sentido de perceber que aspetos da *pipeline* melhor se adequavam ao contexto deste trabalho.

Num conjunto de 281 lesões (210 para treino, 71 para validação) e 183 pacientes (137 para treino, 46 para validação), verificou-se que os modelos treinados com dados equilibrados, seja por subamostragem da classe maioritária ou por geração de instâncias sintéticas para a classe minoritária através da técnica de SMOTE, obtiveram uma proeza superior aos modelos treinados com os dados originais desequilibrados. Verificou-se ainda que as variáveis radiómicas extraídas do VOI (volume de interesse) da glândula inteira produziram modelos extremamente mais confiáveis do que as variáveis radiómicas extraídas dos VOIs das lesões. Sugerindo que as áreas em volta das lesões tumorais oferecem informações relevantes sobre a classificação de Gleason que é atribuída a essa lesão. Além de sugerir que o trabalho monótono de segmentação das lesões realizado pelo radiologista pode não ser necessário ou mesmo prejudicar a assinatura radiómica.

Selecionaram-se 26 dos 288 modelos para validação interna e semi-externa. A primeira realizou-se através de *cross-validation* e a segunda realizou-se através de uma análise da volatilidade das métricas. Aqui foi possível avaliar quais os modelos que estavam mais *overfitted*, tendo-se observado que os modelos treinados com dados gerados pela técnica de SMOTE estavam significativamente mais *overfitted* do que os modelos treinados com dados resultantes da subamostragem da classe maioritária. Do mesmo modo, concluiu-se ainda que os modelos treinados com variáveis radiómicas extraídas dos VOIs das lesões estavam significativamente mais *overfitted* do que os modelos treinados com variáveis radiómicas extraídas do VOI da glândula inteira.

Estes resultados sugerem que áreas polarizantes do ramo de inteligência artificial na saúde como a realização de segmentação das lesões tumorais pode não ser necessário ou mesmo prejudicial para o modelo, bem como gerar variáveis radiómicas pouco reproduzíveis devido à variabilidade de segmentação entre radiologistas diferentes (aspeto também avaliado nesta dissertação).

Contents

Abbreviations

Here, we present the abbreviations used in this dissertation:

- AFS (or AS) - Anterior fibromuscular stroma
- TZ - Tranzitional zone
- PZ - Peripheral zone
- CZ - Central zone
- ED - Ejaculatory ducts
- U - Urethra
- PCa - Prostate cancer
- DRE - Digital rectal examination
- PSA - Prostate specific antigen
- BPH - Benign prostatic hyoertrophy
- TRUS - Trans-rectal ultrasound
- NPV - Negative predictive value
- PPV - Positive predictive value
- csPCa - clinically significant prostate cancer
- ciPCa - clinically insignificant prostate cancer
- mpMRI - multiparametric magnetic ressonance imaging
- T1W - T1-weighted imaging
- T2W - T2-weighted imaging

- DWI - Diffusion weighted imaging
- MRSI - magnetic ressonance spectroscopy
- DCE-MRI - dynamic contrast enhanced MRI
- ADC - Apparent diffusion coefficient
- PI-RADS - Prostate imaging reporting and data system
- CAD - computer aided diagnosis
- GS - Gleason score
- AI - Artificial intelegence
- ML - Machine learning
- VOI - Volume of Interest
- RFE - Recursive feature elimination
- mRMR - minimum redundancy maximum relevance
- ICC - Intraclass correlation coefficient
- LOF - Local outlier factor
- SVM - Support vector machine
- PCA - Principal component analysis
- NB - Naive Bayes
- LR - Logistic Regression
- LR-EN - Logistic Regression with Elastic Net Regularization
- DT - Adaboosted Decision Tree
- RF - Random Forest
- XGB - Exteme Gradient Boost
- TP - True positives
- TN - True negatives
- FP - False positives
- FN - False negatives

- ROC - Receiver Operating Characteristic curve
- AUC - Area under the ROC curve
- TPR - True positive rate
- FPR - False positive rate
- PRC - Precision recall curve
- AUPRC - Area under the precision recall curve
- CV - Cross-validation performance
- TS - Test set performance
- G - Model trained with the Gland dataset
- L - Model trained with the Lesion dataset
- Lp - Model trained with the Lesion Features with Anatomical Zone dataset
- D - Model trained with downsampled data
- S - Model trained with SMOTE data
- nS - Model trained with data that was not sampled
- FWHM - Full width at half maximum

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents the background, motivation, objectives and contributions of this dissertation, as well as the overall document structure.

1.1 Background

1.1.1 Prostate Anatomy

The prostate is a gland of the male reproductive system. It is situated between the bladder and the penis, just in front of the rectum. The main purpose of the prostate is to secrete fluid with proteolytic enzymes into the semen, which will nourish and protect sperm.

The gland is commonly divided into three main glandular zones: central zone (CZ), peripheral zone (PZ) and transitional zone (TZ); and one stromal zone: anterior fibromuscular stroma [28]. See Figure 1.1. The peripheral zone constitutes over 70% of the gland volume and it is known that approximately 70% of prostate tumours originate from here. From the transitional zone arise approximately 25% of prostate tumours. The central zone constitutes 25% of the gland volume and it is known that 8% of prostate tumours originate from here [29].

1.1.2 Prostate Cancer Diagnosis

Prostate cancer in its early stages does not cause any specific symptoms. A suspicion of PCa can arise from: an abnormality on digital rectal examination, DRE [5; 7; 18], or an elevated level of prostate-specific antigen (PSA) in the serum [8; 7].

PSA is an androgen-regulated glycoprotein serine protease that is encoded in the KLK3 gene. Its purpose is to cleave semenogelin, aiding the liquification of the ejaculate [34] and it is produced almost exclusively by the prostate. However, elevated PSA blood levels are not specific to PCa, making an appearance in conditions like BPH (benign prostatic hypertrophy or enlargement of the prostate) and prostatitis (inflammation of the prostate) [19]. Additionally, there is no definite threshold value for PSA

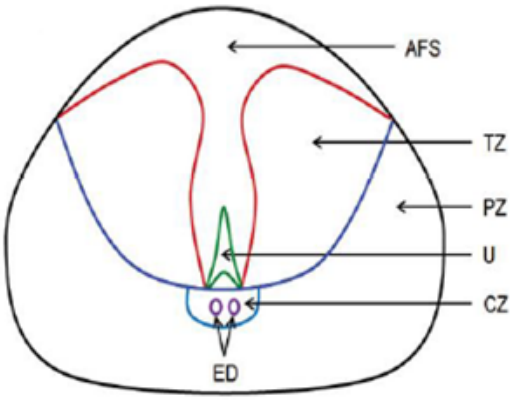

Figure 1.1: Representation of the zonal anatomy of the prostate, as described by [28]. Adapted from [25]. AFS – anterior fibromuscular stroma; TZ – transitional zone; PZ – peripheral zone; U – urethra; CZ – central zone; ED – ejaculatory ducts.

above which a man is guaranteed to have PCa or below which we can safely assume he doesn't [20; 37]. Similarly, an abnormality detected during DRE might be due to BPH or prostatitis in addition to lumps and nodules of PCa [31]. As mentioned before, while approximately 70% of PCas originate from the PZ, the large percentage that is left, does not, and, so, will not be palpable through DRE, due to its anatomical location. Nevertheless, it has been shown that the sensitivity, specificity, and positive predictive value for the detection of PCa by means of PSA is 72.1%, 93.2% and 25.1%, respectively; and by DRE is 53.2%, 83.6% and 17.8%, respectively [30].

The most widely used technique to confirm a suspicion of PCa is biopsy of the prostate gland guided by trans-rectal ultrasound (TRUS). In spite of this, TRUS presents several shortcomings. For a lesion to be detected by TRUS it needs to be hypo or hyperechoic. Although a large majority of PCas are ill-defined hypoechoic lesions [26], a study detected that close to 30% of patients had isoechoic lesions, decreasing TRUS's negative predictive value (NPV) [11]. Further shortcomings include the low specificity and positive predictive value (PPV) of TRUS [27]. In short, there is an elevated risk that a tumour is either missed or that the most aggressive part of the tumour is not targeted, leading to an over-diagnosis of clinically insignificant PCa (ciPCa) or under-diagnosis of clinically significant PCa (csPCa). This could lead to a necessity for repeated biopsies, with the risks that accompany it, an increased number of biopsy cores, an incorrect Gleason score or staging [3].

The current diagnostic approach, comprising PSA levels, DRE and TRUS guided biopsy, lacks both in sensitivity and specificity in PCa detection, in addition to offering limited information regarding aggressiveness and / or stage of the cancer [3].

Magnetic resonance imaging, on the other hand, allows for clear visualization of the zonal anatomy of the prostate, when compared to TRUS [15] and is, therefore, a promising tool for identification, characterization and staging of PCa.

1.1.3 Multiparametric MRI (mpMRI)

Multiparametric MRI (mpMRI) is a combination of functional and anatomical imaging methods: T1-weighted imaging (T1W), T2-weighted imaging (T2W), diffusion weighted imaging (DWI), MR spectroscopy (MRSI) and dynamic contrast enhanced MRI (DCE-MRI) [15]. mpMRI is able to provide morphologic and metabolic data as well as characterize tissue vascularity, showing promise in the detection of PCa.

It has been shown that for biopsy-naive patients, the sensitivity and specificity of mpMRI in detecting csPCa is approximately 85% and 72%, respectively [17; 22]. Additionally, mpMRI followed by targeted biopsy performed better in the detection of csPCa than TRUS-guided biopsy [9; 17].

For patients with a previous negative TRUS-guided biopsy and persistent elevated risk of PCa (elevated PSA and/or abnormal DRE), mpMRI followed by targeted biopsy identified more csPCa than repeated TRUS-guided biopsy. In this context, mpMRI demonstrated an overall sensitivity and specificity in detecting csPCa ranging from 68 to 100% and 41 to 91%, respectively [1; 17; 21; 33]. In addition, the high resolution obtained with mpMRI allows for a less invasive biopsy procedure, since fewer cores are obtained per patient than in repeated TRUS-guided biopsy [17].

1.1.3.1 T1-weighted imaging (T1W)

T1W imaging does not allow for accurate differentiation of zonal anatomy, showing a uniform signal within the prostate. This type of imaging technique is useful for depicting the outline of the prostate gland and for identification of haemorrhage, seen as hyperintense regions. Haemorrhage can appear to be PCa in T2W imaging since both cancerous lesions and haemorrhage appear as hypointense regions in T2W. Thus, a hypointense region in T2W paired with no hyperintense region in T1W can be used for PCa detection [3]. See Figure 1.2.

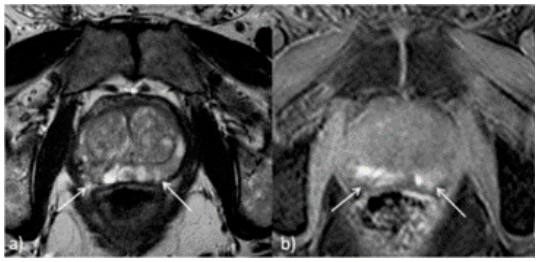

Figure 1.2: a) hypointense areas in T2W imaging. b) hyperintense areas in a fat-saturated T1W image. Extracted from [3]. The T1W image discards the suspicious hypointense areas in the T2W image as haemorrhages and not PCa.

1.1.3.2 T2-weighted imaging (T2W)

T2W imaging is considered to have high spatial resolution, allowing to clearly distinguish between anatomic zones (Figure 1.2a). On T2W imaging of a normal prostate, the peripheral zone (PZ) appears with high signal intensity because of the high content of water in the glandular tissue, whereas the transitional and central zones have often a lower signal intensity, while still being distinguishable from each other [3].

PCa in the PZ appears as a hypointense region in an otherwise hyperintense PZ. PCa in the TZ is not as distinguishable due to the overall lower signal intensity of the healthy TZ, as well as the possible presence of BPH nodules that might mimic PCa or be mixed with the cancerous tissue [3].

1.1.3.3 Diffusion-weighted imaging (DWI)

Diffusion-weighted imaging (DWI) assesses the diffusion of water molecules in the tissue. It is made sensitive to molecule diffusion by using a pair of opposite gradients and measuring the loss of signal. The first gradient introduces a phase-shift in spins and the second gradient, after a time interval Δ, re-phases the spins. If the molecules have not moved during that time interval, the re-phasing will be exact and there will be no loss of signal. However, if diffusion occurred, then the re-phasing will not be exact and there will be a loss of signal. The greater the amount of displacement of water molecules, the greater the signal loss will be. Thus, regions with restricted diffusion will appear bright on a DWI image. [25]

In a normal prostate, especially in the PZ, water molecules move relatively freely, without restriction. PCa contains more tightly packed cells causing restricted diffusion, which is represented in the DWI image by an area of high signal intensity.

The degree of sensitivity to diffusion depends on one parameter, the b-value. The higher the b-value, the greater will be the sensitivity to diffusion. As we can see in Figure 1.3, a low b-value allows to distinguish blood vessels, where diffusion is extremely elevated, but does not differentiate between normal cells and tightly packed cancerous cells. If we choose a higher b-value, the differentiation between blood vessels and healthy cells will not be as clear, however, we can more easily distinguish the tumour. A higher b-value will also diminish the T2 shine-through effect.

Therefore, if a lesion is found on a T2W image, haemorrhage can be discarded by looking at the corresponding T1W image, and a high signal intensity region is found on the corresponding high b-value DWI image, then there is a high probability of PCa.

1.1.3.4 Apparent Diffusion Coefficient (ADC)

The apparent diffusion coefficient (ADC) can be calculated for each voxel, given that DWI images have been taken for at least two b-values (b-value = 0 and high b-value).

The monoexponential model is commonly used in the literature, and it states that if, for each voxel, we plot the signal intensity on DWI on a logarithmic y-axis against the b-value on a linear x-axis, then

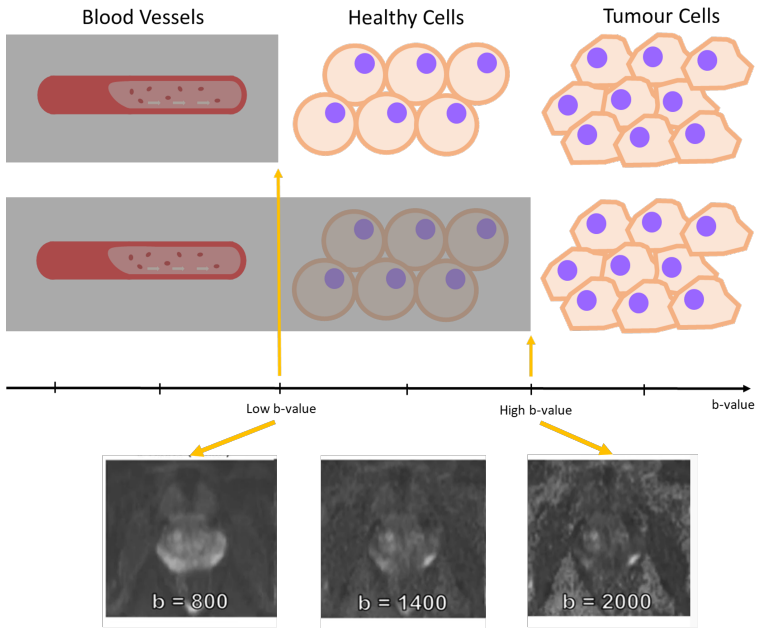

Figure 1.3: Representation of the degree of sensitivity to diffusion regulated by the b-value parameter. The shaded sections represent the dark regions in the DWI image.

the ADC for that voxel corresponds to the slope. The monoexponential model can be described as

$$S(b) = S0 \times e^{-b \times ADCm} \tag{1.1}$$

Where S is the signal intensity of the DWI image at a particular b-value, b is the b-value, S0 is the signal intensity at b-value = 0 s/mm2 and ADCm is the apparent diffusion coefficient of the monoexponential model.

A low ADC, or low slope, corresponds to a slow loss of signal and will be plotted dark on an ADC map. While normal cells' signal intensities on DWI decrease relatively rapidly as b-value increases, appearing bright on the ADC map, PCa's signal intensity should decrease fairly slower, resulting in a lower slope and, thus, lower ADC, appearing hypointense in the ADC map.

While the monoexponential model is commonly used in the literature, it describes the diffusion of pure water without any barriers, which is not accurate for complex biological tissues with cell membranes that create compartments and barriers to diffusion. The kurtosis model, a non-Gaussian model, addresses this issue and has been shown to have higher information content, higher fitting quality, similar repeatability and similar robustness to noise when compared to the monoexponential model [39]. The kurtosis model can be described as:

$$S(b) = S0 \times e^{-b \times ADCk + \frac{1}{6} \times b^2 \times ADCk^2 \times K} \tag{1.2}$$

Where S(b) is the signal intensity of the DWI image at a particular b-value, b is the b-value, S0 is the signal intensity at b-value = 0 s/mm2, ADCk is the diffusion coefficient of the kurtosis model and K is the kurtosis. As before, PCa appears hypointense in the ADC map.

1.1.3.5 Dynamic Contrast Enhanced MRI (DCE-MRI)

DCE-MRI consists of a series of T1W images taken before, during and after the intravenous injection of a contrast agent, commonly gadolinium. DCE-MRI evaluates the differences in the velocities and intensities of contrast agent uptake and washout by the tissue, allowing it to assess the status of tumour angiogenesis, the process of formation of new blood vessels [36].

For each voxel, a signal-vs-time curve is registered, which can be used to calculate parameters such as initial slope, time-to-peak, maximum signal enhancement, washout slope and area under the curve after a specified time. Pharmacokinetic properties can also be estimated. These include Ktrans (transfer constant), Ve (extravascular extracellular volume) and Kep (rate constant) [25].

The development of PCa includes the stimulation of angiogenesis and an increase in vascular permeability, resulting in a signal-vs-time curve with a high and early contrast enhancement peak followed by a rapid washout [3] and higher Ktrans, Ve, Kep when compared to healthy tissue [25].

1.1.3.6 PI-RADS score

One of the biggest challenges in the clinical use of mpMRI is that its interpretation is dependent on the radiologist's subjective opinion and, thus, is inevitably affected by a high rate of inter-reader variability in interpretation and lack of reliability. In order to reduce these effects, a standardized reporting system was developed, the Prostate Imaging Reporting and Data System (PI-RADS). The PI-RADS applies a set of rigid criteria to assign to each MRI sequence a specific score of suspicion out of a five-point suspicion scale (PI- RADS =1, very low suspicion; PI-RADS = 5, very high suspicion), with the final total score being dependent on the number of sequences used. [36]

Despite the improvements after the introduction of PI-RADS, there is still room for improvement in mpMRI reporting. Hence, efforts have been made to implement computer-aided diagnosis (CAD), with the aim to bypass interobserver variability.

1.1.4 Prostate Cancer Aggressiveness

The most widely used measure for PCa aggressiveness is the Gleason Score (GS) [16]. This grading system is assigned to a lesion after biopsy. The larger the GS the more likely it is that the cancer will grow and spread quickly. It ranges from 1 to 5, 1 meaning that the biopsy exposed near healthy tissue, and 5 that the biopsy revealed abnormal tissue (Figure 1.4).

Usually, two grades are given per patient. The primary grade represents the GS of the largest area of the tumour and the secondary grade describes the GS of the next largest area. The sum of the two

scores is taken to be the final GS. A recent modification to this system groups the GSs into five different categories [12]: group 1, $GS = 6$; group 2, $GS = 3 + 4 = 7$; group 3, $GS = 4 + 3 = 7$; group 4, $GS = 8$; group 5, $GS = 9$. A lesion is considered clinically significant for PCa when its GS is higher or equal to 7.

Figure 1.4: Prostate cancer histologic patterns for the grading system. Adapted from [16]

1.1.5 Radiomics

Radiomics is the analysis of medical images through the extraction of quantitative features. The hypothesis behind radiomics is that tissue characteristics might be reflected in the image and, thus, can be quantified by the extracted features. These are extremely valuable for their objectivity and reproducibility.

Radiomic features are of high importance since they are often used to train machine learning models, which can then be used to predict, for instance, the diagnosis, best treatment option or even survival of the patient.

1.1.6 Supervised Machine Learning

Machine learning is a branch of AI where algorithms use statistics to find patterns in data and latter apply said patterns to make predictions about new instances.

Machine learning algorithms can be divided into two main groups: supervised and unsupervised. Supervised machine learning takes labelled training data, or input-output pairs, and attempts to create a function that maps each input (or a vector of predictor variables) to an output (or target variable). On the

other hand, in unsupervised machine learning, the training data is not labelled, so the algorithm just looks for whatever patterns it can find and sorts the training samples into groups accordingly.

Variables can be classified as quantitative or qualitative (or categorical). Quantitative variables are continuous numeric values (for example: the weight of a cookie jar or the size of a house), while qualitative variables are discrete categories (for example: the colour of a cookie jar or the neighbourhood a house belongs to).

In this work, we will address a supervised binary classification machine learning problem, where the input is a vector of radiomic features and the output is the clinical significance of the tumour, described as True for clinically significant PCa or False for clinically insignificant PCa.

1.2 Motivation

Prostate cancer is one of the most prevalent cancers in Portugal, being among the top 4 malignant neoplasm causes of death in 2018, with a crude mortality rate of 38.23 deaths per 100 000 males. [DGS]

Prostate cancer diagnosis and classification is not ideal, relying on unspecific measures such as PSA levels and DRE, followed by biopsy, where an aggressiveness level is attributed in the form of Gleason score.

It has been shown that the Gleason score attributed to a lesion after biopsy is likely to change after radical prostatectomy [13], which confirms the shortcomings of TRUS-guided biopsy mentioned above. Therefore, a reliable, and preferably non-invasive, method for classification of PCa is in urgent demand. With this goal in mind, efforts have been made in the past to use CAD coupled with radiomics and machine learning to predict GS from clinical images, showing promising results.

In fact, texture features have shown potential as biomarkers for PCa aggressiveness [43]. Additionally, previous studies have reported a strong negative correlation between the GS and the ADC values calculated in the tumour region. Furthermore, an even stronger correlation has been found between the GS and the ADC ratio, or normalized ADC, which corresponds to the ADC value calculated for the tumour region divided by the ADC value calculated for the benign region [4; 42]. It is hypothesised that the ADC ratio shows a stronger correlation, because it levels out some of the individual variability, taking into account not only the tumour ADC but also the individual's prostate specific signal characteristics. In addition, the ADC ratio proves to be a more robust feature than the absolute ADC, when comparing different b-values [38].

1.3 Objectives

The hypothesis of this dissertation is that tumour characteristics that are the cause or consequence of Gleason score are reflected in the radiomic features extracted from the MRI image and, thus, can be used to build a classifier model capable of assessing this parameter. Hence, the main goals of this work are to:

1. Extract radiomic features from a set of prostate MRI sequences taking into account the respective segmentation mask.
2. Evaluate the stability of radiomic features with regards to segmentation margins.
3. Build supervised machine learning models that take as input stable radiomic features and predict disease aggressiveness in the form of Gleason score.
4. Validate the machine learning models constructed internally (by means of cross-validation and hold-out test set performances) and semi-externally (by means of a metric volatility analysis).

1.4 Document Structure

Additionally to the present introductory chapter, this document is structured in four chapters as follows:

- **Chapter 2** (Dataset Construction) describes the feature extraction process and subsequent construction of the datasets utilized in this dissertation.
- **Chapter 3** (Feature Reduction) describes the feature reduction steps taken, namely, stability to segmentation, near-zero variance, correlation and feature selection through RFE, mRMR, Boruta and Lasso.
- **Chapter 4** (Classifier Development) presents the work undertaken in the development of 288 classifiers, corresponding to different combinations of pipeline aspects, namely, type of input data (i.e. lesion features vs gland features), sampling strategy, feature selection method and machine learning algorithm.
- **Chapter 5** (Classifier Post-Development Analysis) presents the validation of the highest performing pipelines found in the previous chapter, by means of a metric volatility analysis.
- **Chapter 6** (Conclusion) discusses the main conclusions of this work, as well as some limitations and future work.

1.5 Methodology

The work of this dissertation was performed with both python and the software RapidMiner Studio (version 9.9; https://rapidminer.com/):

- The feature extraction and dataset engineering done in **Chapter 2** (Dataset Construction) was performed in python with the packages pyRadiomics[40] and scikit-learn (version 0.23.2; https://scikit-learn.org/).

- In **Chapter 3** (Feature Reduction) the stability to segmentation analysis was performed in Python, the near-zero variance analysis in R with the caret package (version 6.0-86; https://topepo.github.io/caret/) and the correlation and feature selection steps were performed with the software RapidMiner Studio (version 9.9; https://rapidminer.com/).
- The work of **Chapter 4** (Classifier Development) was performed with the software RapidMiner Studio (version 9.9; https://rapidminer.com/).
- In **Chapter 5** (Classifier Post-Development Analysis) retrieval of performances in the metric volatility analysis was done with the software RapidMiner Studio (version 9.9; https://rapidminer.com/), however the full statistical analysis that followed was performed in Python.

1.6 Contributions

The main contributions of this work are the following:

- Construction of a Rapidminer Studio extension with an operator capable of calculating the Fβ-score performance.
- Construction of a Rapidminer Studio operator capable of performing Boruta feature selection.
- Overview of which pipeline aspects might be more suited in this particular context.
- Further proof of the value of radiomic features extracted from prostate MRI in the prediction of prostate cancer aggressiveness in the form of Gleason score.
- Value of radiomic features extracted from the whole gland VOI over the ones extracted from the lesion VOI.

Chapter 2

Dataset Construction

This chapter describes the feature extraction process and subsequent construction of the datasets utilized in this dissertation.

2.1 Background

2.1.1 Types of Radiomic Features

As briefly described in the previous chapter, radiomics is the analysis of medical images by means of an advanced mathematical analysis that results in the extraction of a large number of quantitative features.

These quantitative features are hypothesised to be able to reflect information about disease-specific processes that are imperceptible to the human eye [41]. Through mathematical quantification of the spacial distribution of signal intensities and pixel interrelationships [41], radiomics can evaluate different dimensions of an image. These dimensions are reflected in the different perspectives provided by the various types of radiomic features, to name a few:[40]

- First-order or histogram based features describe the statistical distribution of voxel intensities within the segmented region. Some first-order features include: mean, median, maximum, minimum, variance, skewness, kurtosis, several percentiles, etc.
- Shape features describe the size and shape of the segmented region. Some shape features include: voxel volume, surface area, sphericity, flatness, maximum diameter, etc.
- Texture features describe interrelationships between pixels. The calculation of texture features begins by the construction of a matrix from which the features are later calculated. This matrix is built according to the type of texture features one wishes to calculate:
  - Grey Level Co-ocurrence Matrix describes the second-order joint probability function of the segmented region.

- Grey Level Size Zone Matrix quantifies grey level zones in the segmented region. Where a grey level zone corresponds to the number of connected voxels that share the same grey level intensity.
- Gray Level Run Length Matrix quantifies grey level runs in the segmented region. Where a grey level run corresponds to the length in number of consecutive voxels that share the same grey level intensity.
- Neighbouring Grey Tone Difference Matrix quantifies the difference between a grey value and the average grey value of its neighbourhood.
- Grey Level Dependence Matrix quantifies grey level dependencies in the segmented region. Where a grey level dependency corresponds to the number of connected voxels within a certain distance that are dependent on the centre voxel, or, in other words, that have a grey level close enough to the centre voxel.

2.2 Methods

2.2.1 Data Gathering

Our dataset consisted of T2W, DW and ADC data from the SPIE-AAPM-NCI PROSTATEx challenge (the data can be downloaded from https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656). The following description of the dataset was provided by the Challenge's organizers: "This collection is a retrospective set of prostate MR studies. All studies included T2-weighted (T2W), proton density-weighted (PD-W), dynamic contrast enhanced (DCE), and diffusion-weighted (DW) imaging. The images were acquired on two different types of Siemens 3T MR scanners, the MAGNETOM Trio, and Skyra. T2-weighted images were acquired using a turbo spin echo sequence and had a resolution of around 0.5 mm in plane and a slice thickness of 3.6 mm. The DWI series were acquired with a single-shot echo planar imaging sequence with a resolution of 2-mm in-plane and 3.6-mm slice thickness and with diffusion-encoding gradients in three directions. Three b-values were acquired (50, 400, and 800), and subsequently, the apparent diffusion coefficient (ADC) map was calculated by the scanner software. All images were acquired without an endorectal coil."

The dataset consisted of 281 lesions from 183 patients. The approximate location of the centroid of each lesion was provided in DICOM coordinates. Cancer was considered significant when the biopsy Gleason score was 7 or higher. The lesions were labelled with "TRUE" and "FALSE" for presence of clinically significant cancer, with a distribution of 67 True lesions and 214 False lesions. The lesions were labelled as belonging to peripheral zone (PZ), transitional zone (TZ), anterior stroma (AS) and seminal vesicles (SV). The distribution of lesions according to anatomic zone and clinical significance is described in Table 2.1.

As the number of lesions is higher than the number of patients, some patients had more than one cancerous lesion. Figure 2.1 shows the distribution of patients according to their number of lesions.

	True	False	Total
PZ	31	128	159
TZ	9	62	71
AS	27	23	50
SV	0	1	1
Total	67	214	281

Table 2.1: Distribution of lesions according to anatomical area and presence of clinically significant cancer. PZ – peripheral zone; TZ – transition zone; AS – anterior stroma; SV – seminal vesicles.

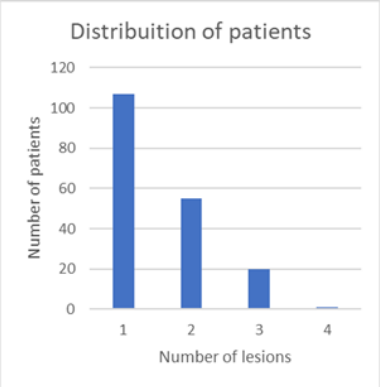

Figure 2.1: Distribution of patients according to their number of lesions.

2.2.2 Feature Extraction

As mentioned in the previous chapter, MRI interpretation is burdened by its subjectivity. Being a human dependent task, segmentation of tumorous lesions suffers from the same problem. In an attempt to overcome this, manual segmentations of the whole prostate gland and of each lesion were performed independently by two radiologists on T2W and DW maps separately. For each sample, one radiologist's volume of interest (VOI) was randomly chosen to be included in the final dataset.

Radiomic features were extracted using the package Pyradiomics (version 3.0) [40] in Python (v. 3.7.9; https://www.python.org/). 14 shape features, 18 first-order features and 75 texture features were extracted from the VOI of three MRI modalities, T2W, DWI and ADC, resulting in a total of 321 features extracted. In the feature extraction of the ADC map, the mask drawn on the DWI was used. The mathematical expressions and semantic meanings of the features extracted can be found at https://pyradiomics.readthedocs.io/en/lates
.

2.2.3 Datasets Construction

The features extracted from a lesion mask VOI constituted the Lesion Dataset. The features extracted from a whole gland mask VOI constituted the Gland Dataset. A Gland was considered to have clinically significant PCa if at least one of its lesions is clinically significant.

From the previous datasets, two additional datasets were constructed:

- Lesion Features with Anatomical Zone dataset – A dataset composed of lesion features plus features describing the anatomical location of the lesion. The possible values for anatomical location were peripheral zone (PZ), transitional zone (TZ), anterior stroma (AS) and seminal vesicles (SV). This categorical variable was encoded with the oneHotEncoder() function of the Python scikit-learn package (version 0.23.2; https://scikit-learn.org/).
- Single-Lesion Whole Gland Features dataset – A truncated dataset composed of patients from the Gland dataset that had one only lesion.

2.2.4 Train/Test Split

The train/test split was performed with the train_test_split() function of the Python scikit-learn package (version 0.23.2; https://scikit-learn.org/). The hold out test sets consisted of 25% randomly selected samples from the original datasets and the split was stratified so that both train and test sets have the same proportion of True labels.

2.3 Results

2.3.1 Datasets Description

The Lesion Dataset is composed of 321 features and 281 lesions, out of which, 67 lesions have a Gleason Score of 7 or higher and are considered clinically significant (True label) and 214 lesions have a Gleason Score lower than 7 and are considered clinically insignificant (False label).

The Gland Dataset is composed of 321 features and 183 patients. A gland was considered to have clinically significant cancer if at least one of its lesions was clinically significant. This resulted in 63 patients being considered as having clinically significant cancer (True label) and 120 patients being considered as having clinically insignificant cancer (False label).

The Lesion Features and Anatomical Zone Dataset is composed of 325 features and 281 lesions, out of which, 67 lesions have a Gleason Score of 7 or higher and are considered clinically significant (True label) and 214 lesions have a Gleason Score lower than 7 and are considered clinically insignificant (False label).

The Single-Lesion Whole Gland Features Dataset is composed of 321 features and 107 patients, out of which, 33 patients have a Gleason Score of 7 or higher and are considered clinically significant (True

	Train	Test
True	48	15
False	89	31
Total	137	46

Table 2.2: Size and label distribution of the train and test Gland Datasets.

	Train	Test
True	51	16
False	159	55
Total	210	71

Table 2.3: Size and label distribution of the train and test Lesion and Lesion with Anatomical Zone Datasets.

label) and 74 patients have a Gleason Score lower than 7 and are considered clinically insignificant (False label).

2.3.2 Train/Test Split

The sizes and label distribution of the train and test sets for the Gland dataset is described in Table 2.2. For both the lesion dataset and the lesion features with anatomical zone dataset refer to table 2.3 for the size and label distribution of the train and test sets, since they are identical. The single-lesion whole gland features dataset was not split into train and test set, due to its already reduced number of samples.

Chapter 3

Feature Reduction

Feature reduction is the process used to select the subset of features that will be used to train the predictive model. Relatively insignificant features may contribute little to the model or even add noise and decrease performance. The several phases of feature reduction done in this work are described in this chapter. See Figure 3.1

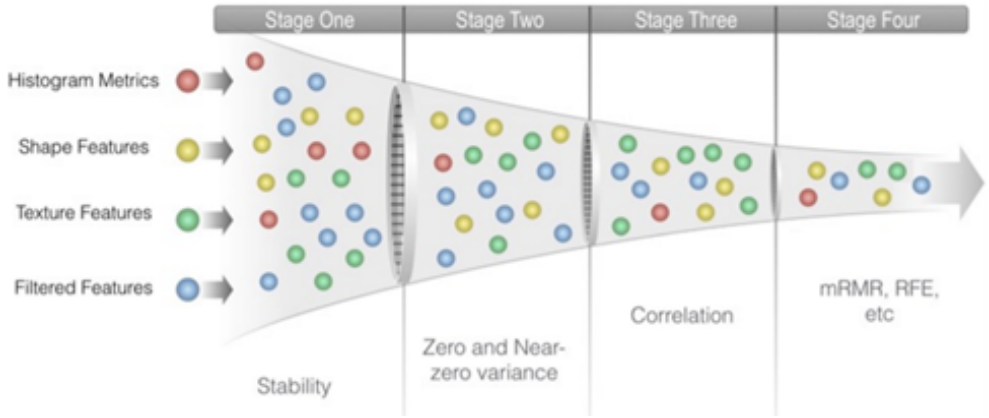

Figure 3.1: Phases of feature reduction performed in this work. Image extracted from [32]

3.1 Background

3.1.1 The Curse of Dimensionality

The curse of dimensionality occurs when a dataset has a lot more features, or predictor variables, than instances, or observations. Two popular aspects that explain why this is problem in the AI world are data sparsity and distance concentration.

As the number of dimensions, or predictor variables, increases, the number of possible combinations that can be found in the data will also increase, but in a geometric way. This means that the higher

the dimension of our dataset, the more observations we will need to gather so as to cover all possible combinations of features. When the training samples available do not capture all possible combinations we have a data sparsity problem. This will lead to the overfitting condition, since the model will not accurately predict the target of feature combinations that it has not come into contact with in the training data.

The distance concentration problem refers to the fact that the distances between observations converge to the same value as dimensionality increases [2]. Since observations appear equidistant, no meaningful relations can be extracted from the data.

To overcome the issues associated with high dimensional data, feature reduction techniques are used. Some of these will be described in the following sections.

3.1.2 Feature Reduction Phases

Lesion or Gland segmentation, like any other human dependent activity, is subject to human error and high inter-reader variability. Hence, features that are highly dependent on segmentation margins, will not be stable predictors, since they easily change depending on the radiologist that performed the segmentation. The first step in the feature reduction performed in this work was to find and remove these unstable features from the dataset.

Similarly, features with zero or near-zero variance across the dataset offer slight information regarding label distinction and, so, should be found and excluded from the data. This was performed as a second step of feature reduction.

Outliers are data points that differ significantly from the remaining observations in the dataset. The presence of outliers in the data can badly affect the mean and standard deviation of features and lead to the development of less precise models. Therefore, they should be identified and excluded. Although this is not a feature reduction step, it is described in this section since the presence of outliers can affect the feature correlation analysis and, so, should be done prior to it.

Two features are correlated when one can be used to predict the other with high accuracy. The presence of correlated features in the dataset can mask useful interactions between features and lead to the development of unstable models, in addition to heightening the curse of dimensionality. The removal of correlated features was the third step of feature reduction.

Feature selection algorithms are classified into three different categories: wrapper methods, filter methods and embedded methods. Wrapper methods are feature selection algorithms that compute different subsets of features until they find the optimal set. This optimal subset is determined through a feature weighing algorithm that is “wrapped” within the main algorithm. Some wrapper-type feature selection algorithms include forward selection (starts with zero features and successively adds features with the greatest improvement to the model), backward elimination (starts with all features and successively removes the least useful features) and stepwise selection (hybrid approach that starts with zero features and successively adds relevant features or removes previously relevant features that are no longer useful).

Filter methods select a subset of features by ranking them according to some useful descriptive measure. Filter-type feature selection algorithms include Spearman's correlation coefficient and ANOVA. Embedded methods are feature selection algorithms that are an integrant part of the machine learning algorithm. These include LASSO and Ridge regressions, as well as Decision Trees.

3.2 Methods

The several phases of feature reduction done in this work were applied only on the train sets and will be described in the following sections.

3.2.1 Feature Stability to Segmentation

Features extracted from the VOIs created by both radiologists were compared with Intraclass correlation coefficient (ICC). The ICC used was a two-way, single rater, absolute agreement ICC model (ICC - 2,1) [23]. Features with ICC 95% confidence interval lower limit over 0.8 were considered to be robust to segmentation and were kept for further analysis.

The assessment of feature stability to segmentation was performed in Python, outside of cross-validation.

3.2.2 Zero and Near-zero Variance Features

Zero and near-zero variance analysis was performed outside of cross validation with the nearZeroVar() function of the R caret package (version 6.0-86; https://topepo.github.io/caret/). This function makes use of the frequency of the most prevalent value over the second most frequent value (which would be near one for well-behaved predictors and very large for highly-unbalanced data) and the percentage of unique values, so as not to exclude predictors that, in spite of having low granularity, are evenly distributed [24].

3.2.3 Outlier Detection and Removal

In order to identify outliers, the local outlier factor (LOF) was used. This algorithm calculates the density of a given subject. Where the density is given by the distance of that subject to its k nearest neighbours. The further away the neighbours are, the smaller the density will be and there will be a higher probability that this subject is an outlier. Since scale affects the distance function, the data was normalized before applying the LOF algorithm.

Samples with LOF over 2 were removed from the original not normalized dataset. Outlier detection and removal was performed inside cross validation with the software RapidMiner Studio (version 9.9; https://rapidminer.com/).

3.2.4 Feature Correlation Analysis

The feature correlation analysis was performed inside cross validation on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Remove Correlated Attributes". This operator uses the Pearson correlation coefficient to compute the correlation between each pair of features. If a pair of features is found to have a correlation higher than the threshold, one of the features is randomly eliminated. The correlation threshold was a hyperparameter optimized during model training.

3.2.5 Feature Selection

In this work, in order to find the optimal feature set (fourth step of feature reduction), four feature selection algorithms were applied separately, and their performance compared. These algorithms were recursive feature elimination (RFE), Boruta algorithm, minimum redundancy maximum relevance algorithm (mRMR) and LASSO regularization.

3.2.5.1 Recursive Feature Elimination

Recursive feature elimination (RFE) is a wrapper-type feature selection algorithm, more specifically, it is a form of backward selection. The weighing method that is "wrapped" within RFE can be chosen according to each situation. In this work, three feature weighing methods were combined with RFE and their performance was evaluated. These feature weighing methods were Support Vector Machine (SVM), Tree importance and Principal Component Analysis (PCA).

The Support Vector Machine's weights are given by the coefficients of the hyperplane calculated. Here we used a SVM with a linear kernel, where the C parameter was a hyperparameter optimized during model training. The C parameter regulates how much misclassification the hyperplane should allow and, consequently, moves along the bias-variance curve. This analysis was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Weight by SVM".

Tree Importance was extracted from the criterion information gain ratio of a Random Forest. This analysis was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Weight by Tree Importance". As described by the operator creators: "each node of each tree is visited and the benefit created by the respective split is retrieved. This benefit is summed per attribute, that had been used for the split. The mean benefit over all trees is used as importance".

Principal Component Analysis weights are given by the coefficients of the first principal component. This analysis was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Weight by PCA".

The three feature weighing methods wrapped in RFE were evaluated on the Lesion and Gland Datasets in a cross validation setting as illustrated in Figure 3.2. Six machine learning algorithms were chosen for this analysis: Naïve Bayes, Logistic Regression, Logistic Regression with Elastic Net regularization, Adaboosted Decision Tree, Random Forest and Extreme Gradient Boost. These will be further described

in the next chapter.

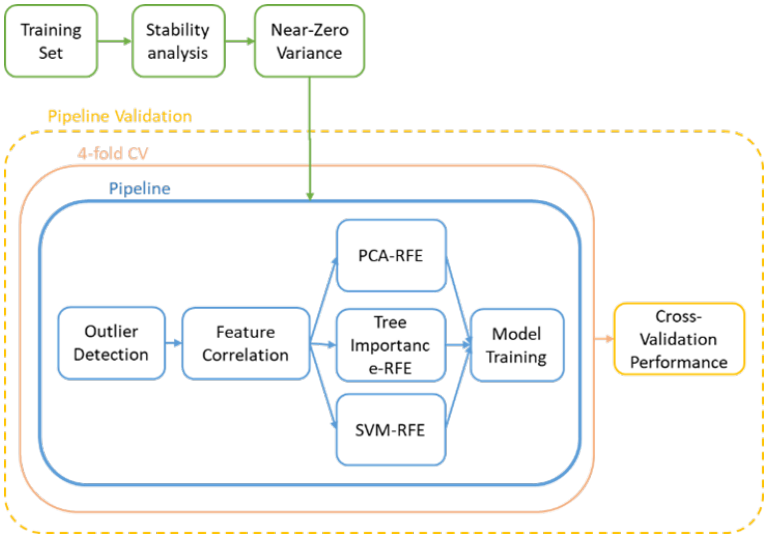

Figure 3.2: Process through which PCA, Tree Importance and SVM were evaluated as weighing methods wrapped in RFE.

The number of features selected by the algorithm was a hyperparameter optimized during model training.

As will be described in the next chapter, Kappa is a powerful metric in imbalanced data settings as we have here. Thus, this metric was chosen to evaluate the performance of the different pipelines.

3.2.5.2 Boruta

In short, the Boruta algorithm selects features that are better predictors than a randomized shuffled version of themselves. Initially, a "shadow" dataset is constructed by randomly shuffling each feature. This shadow dataset is then added to the original dataset. Next, a random forest model is fitted on the new dataset and the importance of each feature is retrieved. Finally, the importance of each original feature is compared to the highest feature importance recorded among the shadow features. If a feature has higher importance then the best shadow feature, then it is selected.

Boruta feature selection was not previously available in RapidMiner Studio, so an operator capable of performing Boruta feature selection was created using the "Python Transformer" operator. This takes a python script where Boruta feature selection was performed with the python package BorutaPy (version 0.3; https://github.com/scikit-learn-contrib/boruta_py).

	T2W	DWI	ADC	Total
Shape	6.49%	3.25%	3.25%	12.99%
First-order	5.19%	3.25%	5.84%	14.29%
Texture	18.18%	38.96%	15.58%	72.73%
Total	29.87%	45.45%	24.68%	100%

Table 3.1: Distribution of the unstable Lesion features across MRI modalities and feature types.

3.2.5.3 Minimum Redundancy Maximum Relevance

Minimum redundancy maximum relevance (mRMR) is a wrapper-type feature selection algorithm, more specifically, it is a form of forward selection. The weighing method that is "wrapped" within mRMR selects features that are the most relevant to the prediction of the target variable and are the least redundant with respect to the features that have been selected in previous iterations.

mRMR feature selection was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Select by MRMR / CFS" of the extension "Feature-Selection-Extension". The number of features selected by the algorithm was a hyperparameter optimized during model training.

3.2.5.4 LASSO Regularization

Regularization is a technique that reduces overfitting by making the model less sensitive to the training data or, in other words, by introducing a small amount of bias so, in return, we get a significant drop in variance. Lasso regularization (L1) reduces the coefficients of each feature in the linear equation so as to reduce the impact a change in that feature could have in the final prediction. The advantage of Lasso is that it can reduce these coefficients all the way to zero, excluding useless features from the equation.

LASSO feature selection was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Logistic Regression". The parameter "use regularization" was selected and alpha was set to 0, indicating Lasso regularization (L1). The feature weights were retrieved and used to select features from the dataset.

3.3 Results

3.3.1 Feature Stability to Segmentation

In the Lesion Dataset, 154 features were found to be unstable, out of the total 321 features. The distribution, in terms of percentage, of these 154 features across MRI modality and feature type is described in Table 3.1.

The feature groups that were found to be most unstable to segmentation were texture features extracted from DWI images (38.96% of unstable features were texture features extracted from DWI). The feature

	T2W	DWI	ADC	Total
Shape	1.56%	0%	0%	1.56%
First-order	0%	3.13%	10.94%	14.06%
Texture	1.56%	10.94%	71.88%	84.38%
Total	3.13%	14.06%	82.81%	100%

Table 3.2: Distribution of the unstable Gland features across MRI modalities and feature types.

type that seemed to be the least robust to segmentation was texture, with 72.73% of unstable features being texture features. The features extracted from DWI images showed a lower stability than the remaining MRI modalities (45.45% of unstable features came from DWI).

Additionally, 23 features were found to be unstable across all three MRI modalities. Of these 23, 17 were texture features, 2 were first order features and 4 were shape features.

In the Gland Dataset, 64 features were found to be unstable, out of the total 321 features. The distribution, in terms of percentage, of these 64 features across MRI modality and feature type is described in Table 3.2.

The feature groups that were found to be most unstable to segmentation were first-order and texture features extracted from ADC maps (10.94% and 71.88% of the unstable features, respectively) and texture features extracted from DWI images (10.94% of the unstable features). Among the feature types, texture features seem to be the most unstable to segmentation (84.38% of the unstable features). Regarding MRI modalities, the features extracted from ADC maps showed a lower stability (82.81% of the unstable features) than the remaining modalities.

On both datasets the texture features seem to be the least stable to segmentation.

3.3.2 Zero and Near-zero Variance Features

In the Lesion Dataset, out of the total 169 stable features, 2 features were found to have near-zero variance: DWI_original_glszm_GrayLevelNonUniformity and ADC_original_glszm_GrayLe-velNonUniformity. While, in the Gland Dataset, no features were found to have near-zero variance.

3.3.3 Feature Selection - Recursive Feature Elimination

The cross-validation performance in terms of Kappa is described in Figure 3.3 with the respective mean values and standard deviation presented in Table 6. The average Kappa score across the six machine learning algorithms for each feature selection method were very close to each other with the highest average performance belonging to SVM-RFE.

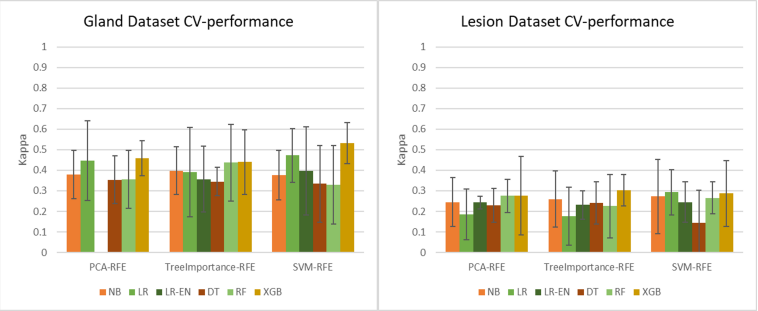

Figure 3.3: Cross-validation performance results clustered by feature selection method. The graph on the left describes Kappa performance on models trained with the Gland Dataset. The graph on the right describes Kappa performance on models trained with the Lesion Dataset.

	Gland Dataset		Lesion Dataset	
	mean	std	mean	std
PCA-RFE	0.3988	0.1306	0.2423	0.1047
Tree Importance - RFE	0.3945	0.1513	0.2393	0.1137
SVM-RFE	0.4065	0.1578	0.251	0.131

Table 3.3: Mean values and standard deviation of the Kappa cross-validation performance of the models described in Figure 3.3

3.4 Discussion

The whole-gland features seem to be considerably more robust to segmentation than lesion features (approximately 50% of lesion features were found to be unstable, compared to approximately 20% of gland features being unstable). This is expected since it is much more challenging for a radiologist to determine lesion borders when compared to determining whole gland borders. Hence, there is a lot more inter-reader variability in lesion segmentation and, consequently, a higher number of unstable features.

Regarding the choice of weighing method wrapped within RFE, a slightly higher performance was observed in the pipelines that performed SVM-RFE. This result coupled with its wide use in the literature confirmed the decision to select SVM-RFE for the remaining analysis.

Chapter 4

Classifier Development and Performance Evaluation

This chapter presents the work undertaken in the development of 288 classifiers, corresponding to different combinations of pipeline aspects, namely, type of input data (i.e. lesion features vs gland features), sampling strategy, feature selection method and machine learning algorithm.

4.1 Background

4.1.1 Bias/Variance Trade off

The bias can be defined as the difference between a model's prediction for a certain instance and its ground truth. A model with high bias makes assumptions about the data, in order to make the target function easier to learn. This can lead to underfitting, since the model is unable to capture the underlying pattern of the data. Some examples of high-bias machine learning algorithms are linear and logistic regressions.

The variance describes how much the target function changes when different training data is used. If the data comes from the same distribution, then the algorithm should have low variance. A model with high variance will be overfitted, since it captures noise along with the underlying pattern of the data. An example of a high-variance machine learning algorithm is the Decision Tree.

Achieving a low-bias and low-variance classifier should ensure that our machine learning model is successful at making predictions for new instances. However, there is no escaping that a decrease in bias will lead to an increase in variance and vice-versa. This trade-off in model complexity is the bias/variance trade off.

4.1.2 Machine Learning Algorithms

In this work we attempt to solve a supervised classification problem. The ”no free lunch” theorem states that there is no ”best” learning strategy [44]. With that in mind, some machine learning algorithms were chosen so as to cover a wide range of machine learning algorithm types:

- Linear classifiers – Logistic Regression with or without regularization of type Elastic Net. As described in the previous chapter, regularization is a technique that reduces overfitting by making the model less sensitive to the training data or, in other words, by introducing a small amount of bias so, in return, we get a significant drop in variance. Ridge regularization (L2) reduces the coefficients of each feature in the linear equation so as to reduce the impact a change in that feature could have in the final prediction. Lasso regularization (L1) is similar to Ridge however, while Ridge regularization can only reduce the coefficients asymptotically close to zero, Lasso can reduce them all the way to zero, excluding useless features from the equation. Elastic Net regularization combines lasso (L1) and ridge (L2) regularizations in a way that allows us to control the weight of each type of regularization.

- Bayesian classifiers – Gaussian Naïve Bayes classifier. The Bayes classifier calculates the most probable classification for a new instance. It is considered optimal since, theoretically, no other algorithm working on the same data can outperform it on average. Hence, its misclassification error is considered the minimal possible error that can be achieved. The Naïve Bayes algorithm is a simplification of the Bayes optimal classifier, where features are considered to be conditionally independent from each other.

- Tree-based classifiers – Adaboosted decision tree, random forest and extreme gradient boost. Tree-based models make use of ‘if-then’ rules to make predictions, for instance if weight is higher than 120 Kg, then patient is obese. Decision trees are the base of all tree-based models and are built in the following manner: first, the features on which to split the data are selected in order to maximize information gain; the data is split multiple times until, finally, a decision is made on when to stop splitting the tree. A very large tree will likely be overfitted, in the sense that it is very specific to the dataset that it was trained on and doesn’t generalize well for new data. This can be avoided by pruning the tree – a technique where the lower sections of the tree are removed. In this work, this was done by setting a maximum tree depth. Adaboost is an ensemble method that builds multiple trees where each tree is focused on correcting the error of the previous tree. Random forest models also build multiple trees however, each tree is trained on a sampled dataset and each node is only allowed to split from a subset of the total feature set. This ensures variety and reduces overfitting. The Gradient boost algorithm follows the same concept as adaboost but utilizes gradient descent for optimization.

		Actual	
		True	**False**
Predicted	**True**	TP	FP or Type I error
	False	FN or Type II error	TN

Table 4.1: Confusion matrix

4.1.3 Performance Metrics

When a model attempts to predict the clinical significance of a given patient's lesion or gland, one of four outcomes occurs:

- The model predicts TRUE when the label is in fact TRUE, these correspond to the true positive results (TP);
- The model predicts TRUE when the label is actually FALSE, these correspond to the false positive results (FP);
- The model predicts FALSE when the label is in fact FALSE, these correspond to the true negative results (TN);
- The model predicts FALSE when the label is actually TRUE, these correspond to the false negative results (FN).

With this information, a confusion matrix can be built (Figure 4.1). Most metrics used to evaluate a model's performance are calculated from the confusion matrix.

A commonly used metric is classifier accuracy. This is given by:

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

Even though, accuracy is wildly used as a metric of model performance, it is not appropriate in imbalanced training problems, as is common in the clinical setting. Here, we often have a minority class, which represents the harshest situation for the patient, and which we wish to accurately predict. A model that predicts the majority class for all samples in the validation or test set will have a relatively high accuracy, corresponding to the percentage of samples belonging to the majority class. Since accuracy is not able to distinguish between the correctly classified examples in the different classes, it might lead to an overestimation of model performance.[6; 14]

In the same lines, the commonly used metric AUC, or area under the receiver operating characteristic curve (ROC), is also not appropriate for imbalanced data. The ROC curve plots TPR versus FPR,

$$True\ Positive\ Rate\ /\ Recall\ /\ Sensitivity = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

with the desired plot having high TPR and low FPR. In an imbalanced setting, the FPR is pulled down due to a large number of true negatives (majority class). Hence, the AUC-ROC may overestimate performance. [35]

On the other hand, precision recall curves (PRC) have been shown to be more informative than ROC when dealing with imbalanced data, since precision is influenced by both classes (TP and FP) [35].

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

Another important aspect of the clinical setting is the cost of misclassifications. Classifying a patient as positive for clinically significant cancer when it is not, will only lead to further examination. However, classifying a patient as negative when in fact they have clinically significant cancer might prevent the patient from getting the necessary treatment. Hence, a FN result, or type II error, has much more drastic consequences than a FP result, or type I error. So, it is important for our classifier to focus on minimizing the type II error, instead of treating both errors with equal importance. The performance metric that takes FN (type II error) into account is the recall and the one that accounts for FP (type I error) is precision. So, our classifier should prioritize a higher recall rather than a higher precision, in order to minimize type II error.

The performance metric that takes into account both precision and recall is the Fβ-measure.

$$F_\beta = \left(1 + \beta^2\right) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

For a $\beta = 1$, the same weight is put on precision and recall. For a β between 0 and 1, more weight is given to precision. For a β higher than 1, more weight is given to recall.

Cohen's Kappa is also a metric that can handle imbalanced data problems. It ranges from 0 to 1 and tells us how much better our classifier is at predicting the class label, when compared to a classifier that makes a random prediction according to the frequency of each class.

4.1.4 Hyperparameter Optimization through Nested Cross Validation

Nested cross-validation is a hyperparameter optimization algorithm, that attempts to reduce overfitting by finding the optimal hyperparameters for multiple subsections of the data space and later making a decision regarding the final set of best hyperparameters.

Nested cross-validation begins by splitting the data into k different folds. One of these folds is held out of the training process, while the remaining k-1 folds are used for hyperparameter tuning. These k-1 folds are further divided into j different folds. One of these is again held out, while the remaining data is trained on every possible hyperparameter combination. Each trained classifier is validated by quantifying its performance on the held-out fold. The process is repeated so that each of the j folds is

used exactly once for validation. For each hyperparameter combination, a mean performance is obtained. The hyperparameter combination with the highest mean performance is applied to train the full k-1 folds and this classifier is evaluated by the held-out fold. Again, this process is repeated until each of the k folds has been used once for validation. The hyperparameter combination that performed highest on the outer fold is chosen as optimal. The nested cross-validation algorithm described above is shown in Figure 4.1.

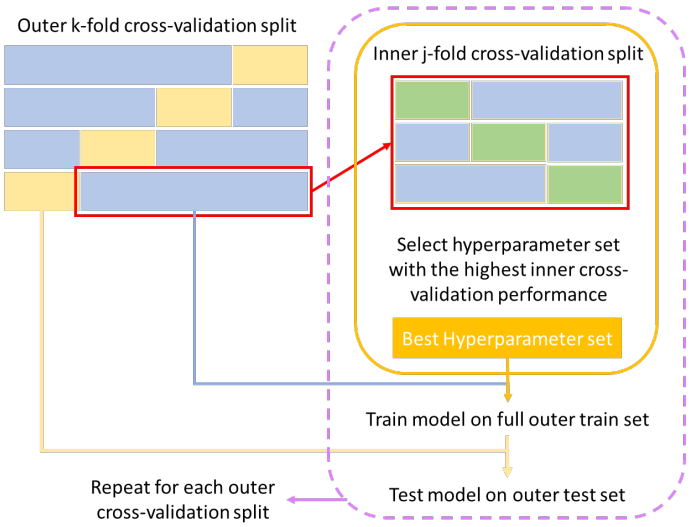

Figure 4.1: Nested cross-validation algorithm.

4.2 Methods

In this work, different aspects of model development were assessed and compared. The different combinations are described in Figure 4.2.

In total, 288 pipelines were produced. Each was trained and validated according to the diagram in Figure 4.3.

4.2.1 Sampling Strategies

The undersampling of the majority class was done in a random fashion outside of cross validation. All minority class samples were kept, and samples from the majority class randomly chosen so as to match the number of samples in the minority class. This was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator “Sample”.

In both lesion trainsets, the minority class constituted of 51 lesions. Therefore, 51 lesions were randomly selected from the majority class pool, making the final sampled dataset 102 lesions long. While

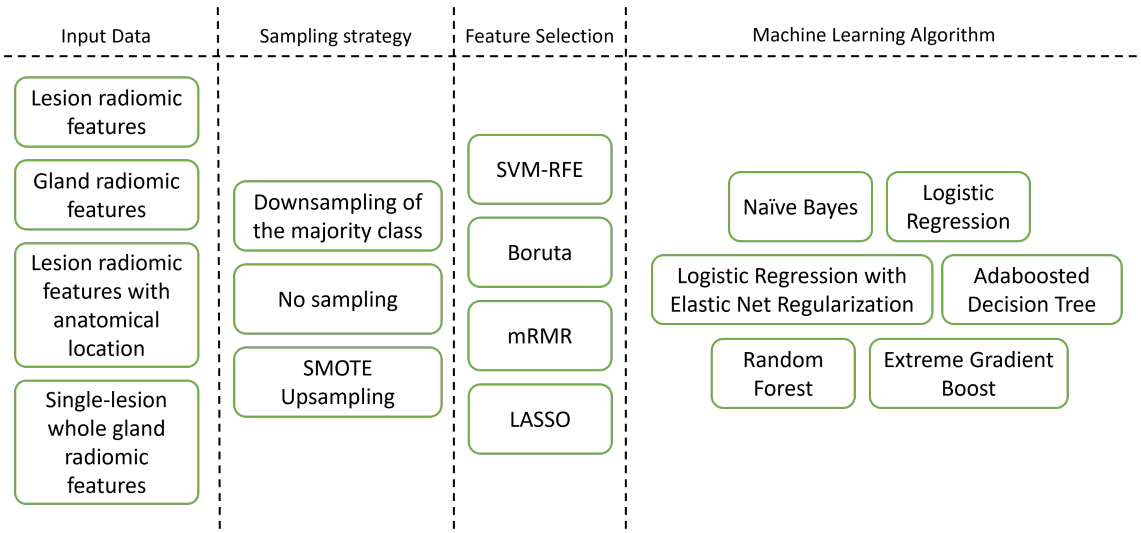

Figure 4.2: Different model dimensions explored in this study.

in the gland trainset, the minority class constituted of 48 patients. Therefore, 48 patients were randomly selected from the majority class pool, making the final sampled dataset 96 patients long. Finally, in the single-lesion whole gland dataset, the minority class constituted 33 patients. Therefore, 33 patients were randomly selected from the majority class pool, making the final sampled dataset 66 patients long.

The SMOTE algorithm generates synthetic samples for the minority class. It works by choosing a minority class sample at random, finding its k nearest neighbours, randomly choosing one of those neighbours and, finally, generating a synthetic sample somewhere in the high-dimensional ”line” that connects those two samples. In this work, the number of nearest neighbours considered was 5. SMOTE upsampling was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator “SMOTE Upsampling”.

In both lesion trainsets, the majority class constituted of 159 lesions and the minority class of 51 lesions. Therefore, 108 lesions were generated with SMOTE, making the final sampled dataset 318 lesions long. While in the gland trainset, the majority class constituted of 89 patients and the minority class of 48 patients. Therefore, 41 patients were generated with SMOTE, making the final sampled dataset 178 patients long. Finally, in the single-lesion whole gland dataset, the minority class constituted 33 patients. Therefore, 41 patients were randomly selected from the majority class pool, making the final sampled dataset 148 patients long.

4.2.2 Machine Learning Algorithms

The RapidMiner Studio (version 9.9; https://rapidminer.com/) implementation of the chosen machine learning algorithms was utilized.

A Naïve Bayes classifier (NB) with laplace correction was trained with the operator “Naive Bayes”.

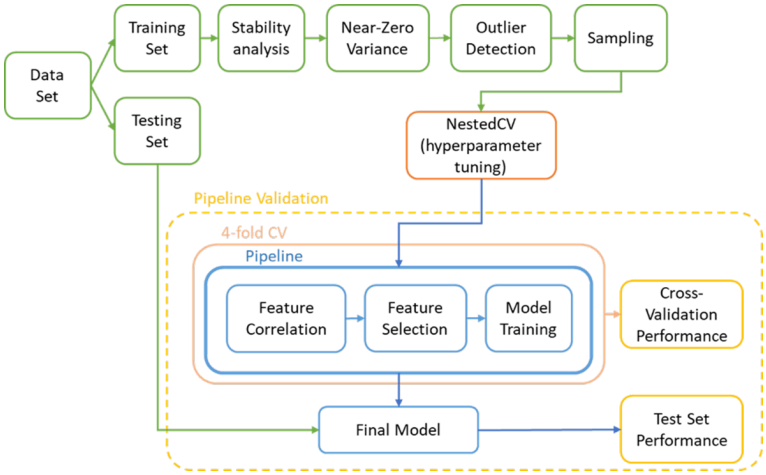

Figure 4.3: Overall pipeline followed in this study to train and validate models.

A logistic regression classifier (LR) was trained with the operator "Logistic Regression". The parameters "standardize", "add intercept" and "remove collinear columns" were selected and the "solver" parameter was set to "AUTO". In the case of logistic regression with elastic net regularization (LR-EN), the parameter "use regularization" was selected and alpha was a hyperparameter optimized during model training. Alpha ranges from 0 to 1, 0 corresponding to Lasso regularization (L1) and 1 to Ridge regularization (L2).

An Adaboosted Decision Tree classifier (DT) was trained with the operators "AdaBoost" and "Decision Tree". The number of iterations in the AdaBoost operator was set to 10, the criterion according to which features are selected in the Decision Tree was set to "gain_ratio", corresponding to information gain ratio, a criterion related to the entropy of a feature. The "maximal depth" parameter was a hyperparameter optimized during model training.

A Random Forest classifier (RF) was trained with the operator "Random Forest". The criterion according to which features are selected was again set to "gain_ratio". The "maximal depth" and "number of trees" parameters were hyperparameters optimized during model training. The voting strategy by which the forest makes a decision was set to "confidence vote".

An extreme gradient boost classifier was trained with the operator "Gradient Boosted Trees". The "maximal depth" and "number of trees" parameters were hyperparameters optimized during model training. The remaining parameters were left with the default values set by RapidMiner.

4.2.3 Performance Metrics

In this work, we have chosen to optimize the F2-score and we report Cohen's Kappa and the area under the precision recall curve (AUPRC) as measures of model performance. Additionally, standard ROC-AUC was calculated for literature comparison purposes.

Hyperparameter		Possible values
Correlation threshold		[0.8, 0.9, 1.0]
SVM-RFE	C	[0.01, 0.1, 1, 10, 100]
SVM-RFE/mRMR	number of features	[10, 12, 14, 16, 18, 20, 22, 24]
LASSO	lambda	[0.2, 0.4, 0.6, 0.8, 1.0]
LR-EN	alpha	[0, 0.2, 0.4, 0.6, 0.8, 1.0]
DT-AdB	Tree depth	[2, 3, 4, 5, 6, 7, 8, 9, 10]
RF / XGB	Tree depth	[9, 11, 12, 14]
	Maximum number of trees	[80, 90,100]

Table 4.2: List of hyperparameters explored in this study.

ROC-AUC and Cohen's Kappa calculation was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/ with the operators "Performance Binomial Classification". AUPRC was calculated on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Performance (AUPRC)" from the extension "Operator Toolbox".

Fβ-score performance was not previously available in RapidMiner Studio, so an operator capable of calculating the metric was built in Java (version 8.0.2810.9), and the extension was installed in RapidMiner.

4.2.4 Hyperparameter Optimization and Classifier Validation

Hyperparameter tuning was done in a nested cross-validation fashion with an exhaustive grid search. This was performed on RapidMiner Studio (version 9.9; https://rapidminer.com/) with the operator "Optimize Parameters (Grid)". The list of hyperparameters can be found on Table 4.2.

4.2.5 Best Classifier Selection

The best classifiers were selected according to their cross-validation F2 and Kappa performance, following the rule:

$$CV_{F2} > 0.8 \cap CV_{Kappa} > 0.5$$

These were applied to the hold-out test set for validation.

The purpose of the single-lesion whole gland dataset was to more accurately compare the performance of models trained on lesion data with the ones trained on gland data. In addition, this dataset was not sufficiently large to divide it by creating a hold out test set. Thus, these models were not considered for the best classifiers.

4.3 Results

4.3.1 Feature Selection Methods

In Figure 4.4, we can see the cross-validation F2-score and Cohen's Kappa performance results grouped by feature selection method for the pipelines trained on the gland (G), lesion (L) and lesion with anatomical zone (Lp) datasets.

Overall, the Boruta algorithm did not perform as well as expected. Despite having a high cross-validation F2, most kappa values were extremely low, especially for pipelines trained on whole gland features. Pipelines trained with data that underwent SVM-RFE achieved an average cross-validation F2 of 0.7226 and Kappa of 0.3781. While the feature sets that underwent mRMR achieved average performances of 0.7071 on F2 and 0.4095 on Kappa. Overall, at this stage, SVM-RFE and mRMR pipelines show a similar average performance. Pipelines trained with data that underwent Lasso feature selection achieved an average cross-validation F2 of 0.643 and Kappa of 0.347, not performing, on average, as high as SVM-RFE and mRMR.

4.3.2 Sampling

In Figure 4.5, we can see the cross-validation F2-score and Cohen's Kappa performance results grouped by sampling method for the pipelines trained on the gland (G), lesion (L) and lesion with anatomical zone (Lp) datasets.

We can see that the average cross-validation performance results were higher on the models trained with sampled data on both F2 and Kappa, with average F2 of 0.7541 and Kappa of 0.3659 on the models trained with downsampled data and F2 of 0.8094 and Kappa of 0.3666 on the models trained with SMOTE data. As expected, the pipelines trained with the original imbalanced dataset performed lower with F2 of 0.4779 and Kappa of 0.2626.

4.3.3 Machine Learning Algorithms

In Figure 4.6 we can see the cross-validation F2-score and Cohen's Kappa performance results grouped by machine learning algorithm for the pipelines trained on the gland (G), lesion (L) and lesion with anatomical zone (Lp) datasets. On average the Naïve Bayes classifier achieved an F2 of 0.6573 and a Kappa of 0.3016, the Logistic regression classifier achieved an F2 of 0.6569 and a Kappa of 0.3058, the Logistic regression classifier with Elastic Net regularization achieved an F2 of 0.6984 and a Kappa of 0.3002, the Adaboosted Decision Tree classifier achieved an F2 of 0.6784 and a Kappa of 0.2931, the Random Forest classifier achieved an F2 of 0.6725 and a Kappa of 0.3914 and, finally, the Extreme Gradient Boost classifier achieved an F2 of 0.7226 and a Kappa of 0.3885. Overall, the Random Forest and Extreme Gradient Boost classifiers performed, on average, significantly higher in terms of Kappa than the remaining machine learning algorithms. In terms of F2, the average results were similar across

Figure 4.4: Cross-validation F2 and Kappa performance results grouped by feature selection method.

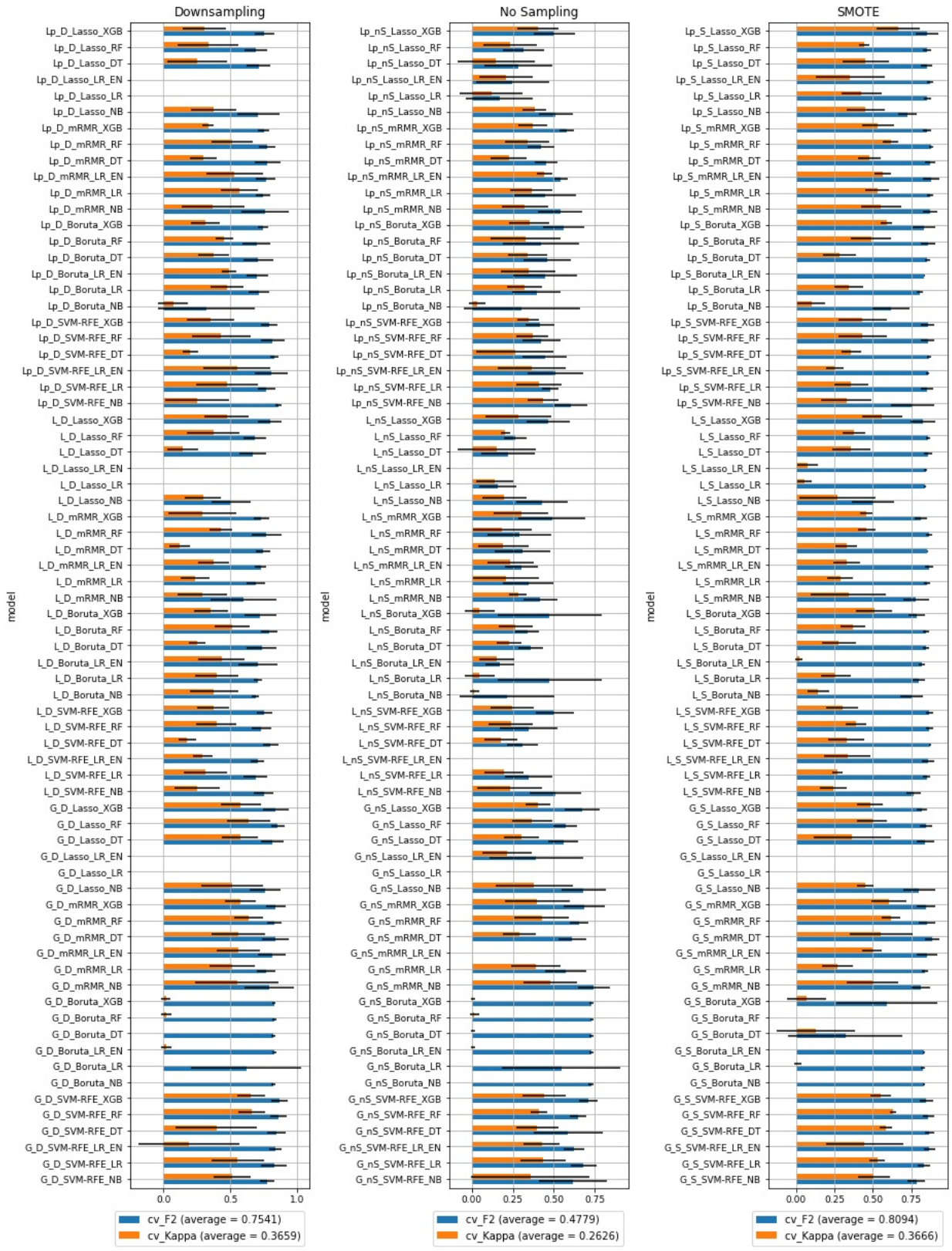

Figure 4.5: Cross-validation F2 and Kappa performance results grouped by sampling method.

machine learning algorithms with the exception of the Extreme Gradient Boost classifier, that performed slightly higher.

Figure 4.6: Cross-validation F2 and Kappa performance results grouped by machine learning algorithm.

4.3.4 Type of Input Data

In Figure 4.7, we can see the cross-validation F2-score and Cohen's Kappa performance results grouped by type of input data. On average, classifiers trained with whole Gland radiomic features achieved a cross-validation performance of 0.7426 on F2 and of 0.351 on Kappa. While classifiers trained with the Lesion Dataset achieved an average cross-validation F2 of 0.6344 and a Kappa of 0.2749. The classifiers trained with the Lesion features with anatomical zone dataset achieved an average cross-validation F2

of 0.6682 and a Kappa of 0.3687. And, finally, the classifiers trained with the single-lesion whole gland features dataset achieved an average cross-validation F2 of 0.7508 and a Kappa of 0.3806. Overall, the pipelines trained with whole gland features performed, on average, higher than the ones trained on lesion features, both in terms of Kappa and of F2.

Figure 4.7: Cross-validation F2 and Kappa performance results grouped by type of input data.

4.3.5 Best Classifiers Selection and Validation

Figure 4.8 shows the 26 models that satisfied the condition: F2>0.8 AND Kappa>0.5. 65% of these are models trained on whole gland features. All of the best models were trained on data that underwent some kind of sampling: 42% downsampled data and 58% SMOTE data. Regarding feature selection, 31% of the pipelines included SVM-RFE, 50% included mRMR, 15% included Lasso and 4% included Boruta. As for the machine learning algorithm, the large majority of best models are tree-based algorithms (73%) and the remaining models are logistic regressions with or without elastic net regularization and one Naïve Bayes pipeline.

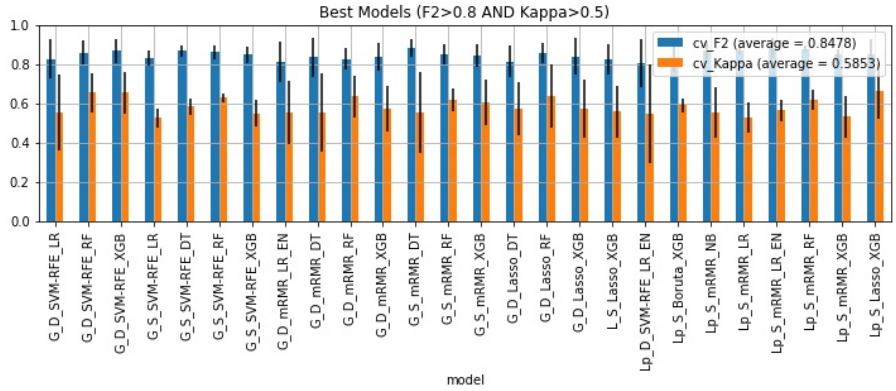

Figure 4.8: Classifiers that performed highest in terms of Kappa and F2.

Figure 4.9 shows the performance of these 26 models on the hold out test set.

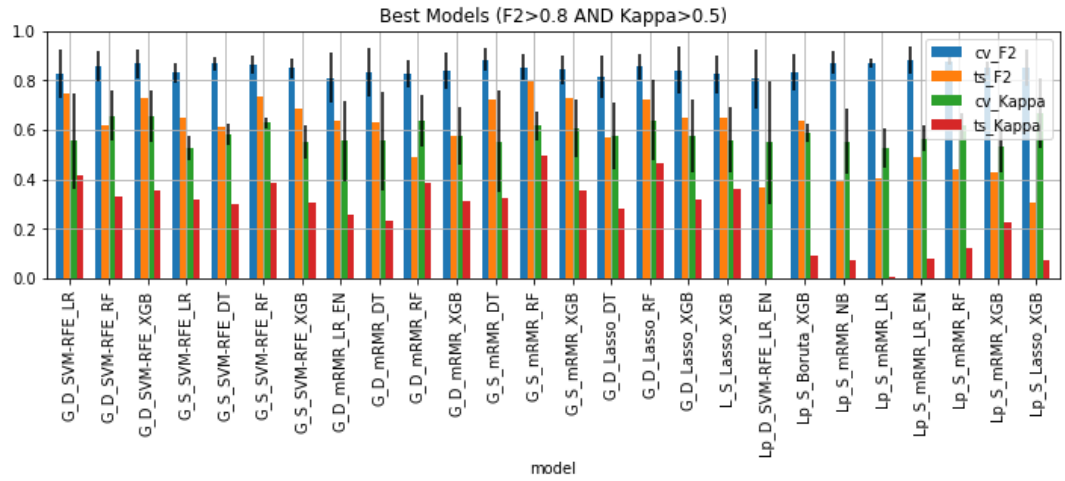

Figure 4.9: Performance of the best models on the hold out test set in terms of F2 and Kappa.

Table 4.3 shows the performance of the best models on the cross-validation setting and on the hold

out test set in terms of F2, Kappa, ROC-AUC and AUPRC. In addition, it shows the difference between cross-validation and test set performance. The models where this difference is closest to zero are the least overfitted models.

Model	cv_F2	ts_F2	cv_Kappa	ts_Kappa	cv_AUC	ts_AUC	cv_AUPRC	ts_AUPRC	Δ F2	Δ Kappa	Δ AUC	Δ AUPRC
G_D_SVM-RFE_LR	0.826	0.745	0.555	0.416	0.765	0.772	0.68	0.518	0.081	0.139	-0.007	0.162
G_D_SVM-RFE_RF	0.859	0.618	0.657	0.333	0.857	0.843	0.787	0.734	0.241	0.324	0.014	0.053
G_D_SVM-RFE_XGB	0.868	0.729	0.655	0.354	0.859	0.753	0.792	0.545	0.139	0.301	0.106	0.247
G_S_SVM-RFE_LR	0.831	0.652	0.528	0.32	0.798	0.766	0.742	0.655	0.179	0.208	0.032	0.087
G_S_SVM-RFE_DT	0.868	0.611	0.584	0.301	0.806	0.746	0.545	0.449	0.257	0.283	0.06	0.096
G_S_SVM-RFE_RF	0.862	0.737	0.629	0.385	0.873	0.788	0.841	0.576	0.125	0.244	0.085	0.265
G_S_SVM-RFE_XGB	0.849	0.684	0.551	0.308	0.847	0.728	0.805	0.504	0.165	0.243	0.119	0.301
G_D_mRMR_LR_EN	0.812	0.638	0.557	0.26	0.789	0.755	0.724	0.53	0.174	0.297	0.034	0.194
G_D_mRMR_DT	0.836	0.632	0.556	0.231	0.767	0.634	0.636	0.404	0.204	0.325	0.133	0.232
G_D_mRMR_RF	0.827	0.488	0.636	0.385	0.789	0.757	0.683	0.737	0.339	0.251	0.032	-0.054
G_D_mRMR_XGB	0.84	0.575	0.576	0.314	0.808	0.719	0.718	0.485	0.265	0.262	0.089	0.233
G_S_mRMR_DT	0.884	0.722	0.554	0.325	0.778	0.691	0.405	0.271	0.162	0.229	0.087	0.134
G_S_mRMR_RF	0.853	0.798	0.618	0.494	0.841	0.847	0.8	0.642	0.055	0.124	-0.006	0.158
G_S_mRMR_XGB	0.844	0.729	0.607	0.354	0.814	0.783	0.766	0.576	0.115	0.253	0.031	0.19
G_D_Lasso_DT	0.815	0.568	0.574	0.282	0.808	0.71	0.696	0.346	0.247	0.292	0.098	0.35
G_D_Lasso_RF	0.855	0.722	0.638	0.466	0.826	0.824	0.754	0.659	0.133	0.172	0.002	0.095
G_D_Lasso_XGB	0.84	0.652	0.576	0.32	0.856	0.7	0.798	0.447	0.188	0.256	0.156	0.351
L_S_Lasso_XGB	0.826	0.652	0.56	0.363	0.855	0.755	0.844	0.54	0.174	0.197	0.1	0.304
Lp_D_SVM-RFE_LR_EN	0.806	0.368	0.55	0.001	0.786	0.581	0.706	0.812	0.438	0.549	0.205	-0.106
Lp_S_Boruta_XGB	0.833	0.64	0.591	0.091	0.874	0.646	0.861	0.874	0.193	0.5	0.228	-0.013
Lp_S_mRMR_NB	0.873	0.389	0.554	0.075	0.836	0.55	0.793	0.713	0.484	0.479	0.286	0.08
Lp_S_mRMR_LR	0.872	0.404	0.528	0.006	0.853	0.53	0.804	0.783	0.468	0.522	0.323	0.021
Lp_S_mRMR_LR_EN	0.882	0.49	0.566	0.078	0.849	0.667	0.783	0.862	0.392	0.488	0.182	-0.079
Lp_S_mRMR_RF	0.879	0.44	0.617	0.124	0.881	0.58	0.868	0.805	0.439	0.493	0.301	0.063
Lp_S_mRMR_XGB	0.85	0.427	0.534	0.227	0.864	0.697	0.845	0.871	0.423	0.307	0.167	-0.026
Lp_S_Lasso_XGB	0.852	0.305	0.667	0.073	0.904	0.634	0.907	0.846	0.547	0.594	0.27	0.061

Table 4.3: Best classifiers' cross-validation and test set performances, as well as the difference between cross-validation and test set performance, Δ. The performance columns are color coded from highest value in green, to lowest value in white. The Δ columns are color coded from lowest value in green to highest value in red.

4.4 Discussion

Regarding feature selection, a low performance was unexpectedly observed from the pipelines that applied Boruta feature selection. These showed a high F2, because the model would classify the large majority of samples as the minority class, leading to a high recall. However, the low Kappa score makes it clear that these were not useful models. It was observed that the Boruta algorithm found very few features that were better predictors than the random versions of themselves. Hence, it is hypothesised that the number of features selected by the Boruta algorithm (around 3 features) was not enough to build a meaningful radiomics signature, which led to the poor results.

The pipelines where sampling was applied performed higher than the pipelines where no sampling was done, whether it was downsampling of the majority class or upsampling of the minority class with SMOTE. This was expected since training a model with balanced data gives it equal opportunities to

learn from both classes.

In terms of input data, it was observed that the performance results obtained with the Gland Dataset were higher than the ones obtained with the lesion Datasets. This might suggest that the areas surrounding the tumour lesions offer relevant information regarding the Gleason Score that is ultimately attributed to that lesion. In addition to suggesting that the monotonous lesion segmentation work performed by radiologists may not be necessary or even be harming to the radiomics signature. However, it is of note that a few patients had more than one lesion. If these multiple lesions have the same clinical significance (same target label), then it seems reasonable that the model performs higher with gland features since it has more information pointing to the correct label. In order to make a fair comparison between the performance of both types of input data, the single-lesion whole gland dataset was created, including only patients with a single lesion. The performance results obtained with this smaller dataset confirm the suspicions above, that whole gland features produce more reliable machine learning models than lesion features.

As a final note, it is important to point out that given so many pipeline combinations we have to assume that it is possible to find one that performs well by chance. Statistically speaking, we could remedy this by doing something similar to a multiple comparisons p-value correction. However, at this point, we are not aware of such a correction for machine learning performance metrics.

Chapter 5

Classifier Post-Development Analysis

This chapter presents the validation of the highest performing pipelines found in the previous chapter, by means of a metric volatility analysis.

5.1 Background

Comparably to other technologies used in the medical field, the importance of clinical validation of machine learning models cannot be overstated. This can be assessed in terms of classifier performance, patient outcome, cost-benefit analysis, etc.

The reliability of a classifier's real-world clinical performance is often estimated during cross-validation, which calculates the test set performance by repeatedly holding out a subset of the training samples from the fitting process and then applying the classifier to those held out observations. Another way of estimating this real-world performance is by applying our trained classifier to a hold-out test set, a random subsample of the original dataset.

The issue with both of these approaches is selection bias, which is the idea that we may get an extremely high or low test set performance due to chance or that our collection of samples is not representative of the real-world distribution and, consequently, leads to erroneous performance results that are not reflective of the classifier's performance in the "wild". This is especially concerning when doing a retrospective study, due to the data drift phenomenon.

To assess this concern, and in the absence of an external validation dataset, a volatility analysis was performed on the highest-ranking classifiers found in the previous chapter. This analysis will be described in the following sections.

5.2 Methods

5.2.1 Volatility Analysis

The Gland, Lesion and Lesion with anatomical location Datasets were each randomly split in training and testing sets in 50 different ways, according to 50 different random seeds. Each of the highest-ranking classifiers was then trained on each of the 50 training sets and validated through both cross-validation and each of the 50 hold-out testing sets. The distribution of cross-validation and test set performance results was recorded for further analysis.

Mean and standard deviation values were calculated for each performance metric and each classifier. The difference between cross-validation and test set performance of each random split was calculated and is presented as Δ. This value represents how overfitted the model is.

The collection of performance results was performed in RapidMiner Studio and the statistical metrics were calculated in Python. This analysis was based on the metric volatility analysis performed by the Probatus package (https://ing-bank.github.io/probatus/).

Lines were fit to the plotted histograms of cross-validation and test set performance distributions, respectively, and, for each, the full width at half maximum metric was calculated. The former was performed with the Seaborn package (version 0.11.1; https://seaborn.pydata.org) and the latter was calculated as bellow,

$$FWHM = 2\sqrt{2\ln 2}\sigma \tag{5.1}$$

Where σ is the standard deviation of the distribution of performances.

5.2.2 Normality Tests

All performance distributions were tested for normality using the Shapiro-Wilk test and the D'Agostino K^2 test. The Shapiro-Wilk test evaluates the likelihood that a sample was drawn from a Gaussian distribution and was performed with the shapiro() function of the SciPy package (version 1.5.2; https://docs.scipy.org). The D'Agostino K^2 test calculates the kurtosis (how much of the distribution belongs to the tails) and skewness (a measure of distribution asymmetry) of the data, in order to determine if it differs significantly from the normal distribution. D'Agostino K^2 test was performed with the normaltest() function of the SciPy package (version 1.5.2; https://docs.scipy.org).

Both tests behave like common hypothesis tests in the sense that there is a null and alternative hypothesis and as a result we get a test statistic and a p-value that will tell us if we have significant statistical evidence to reject the null hypothesis. In both tests, the hypotheses were as follow:

H_0: *the distribution of performances is Gaussian*
H_1: *the distribution of performances is not Gaussian*

The significance level, α, was chosen to be 0.05. Therefore, a p-value lower than 0.05 will lead to a decision to reject the null hypothesis since there is sufficient statistical evidence that the sample does not

belong to a Gaussian distribution. On the other hand, a p-value higher than 0.05, will lead to a decision to fail to reject the null hypothesis, since there is not sufficient statistical evidence that the sample does not belong to a Gaussian distribution.

5.2.3 Distribution Comparison Tests

For each classifier, the distribution of cross-validation performances was compared to the distribution of test set performances, to assess whether they belonged to the same distribution. Two statistical tests were used: the paired t-test and the Kolmogorov-Smirnov test.

The paired t-test compares the mean and standard deviation of two paired groups to determine whether there is a significant difference between the two. In our specific situation, such a statistical test is appropriate due to the paired nature of our samples, since from each train test split resulted one cross-validation performance and one test set performance. The paired t-test was performed with the ttest_rel() function of the SciPy package (version 1.5.2; https://docs.scipy.org).

The Kolmogorov-Smirnov test is a non-parametric test that evaluates the empirical cumulative distribution functions of each sample to measure whether they are similar enough to belong to the same distribution. The Kolmogorov-Smirnov test was performed with the kstest() function of the SciPy package (version 1.5.2; https://docs.scipy.org).

Both tests behave like common hypothesis tests in the sense that there is a null and alternative hypothesis and as a result we get a test statistic and a p-value that will tell us if we have significant statistical evidence to reject the null hypothesis. In both tests, the hypotheses were as follow:

H_0: *the distributions of cross-validation and test set performances are identical*
H_1: *the distributions of cross-validation and test set performances are different*

The significance level, α, was chosen to be 0.05. Therefore, a p-value lower than 0.05 will lead to a decision to reject the null hypothesis since there is sufficient statistical evidence that the samples do not belong to identical distributions. On the other hand, a p-value higher than 0.05, will lead to a decision to fail to reject the null hypothesis, since there is not sufficient statistical evidence that the samples do not belong to the same distribution. A decision to reject the null hypothesis will then lead to the conclusion that the model is overfitted.

5.2.4 Comparison with Dummy Classifier

The models where no significant difference was found between the cross-validation and test set performance distributions were compared with a dummy classifier. This was created with the DummyClassifier() function of the Python scikit-learn package (version 0.23.2; https://scikit-learn.org/) and the strategy used to generate predictions was set to ”stratified”, which means that the classifier will make predictions according to the train set’s label distribution.

5.3 Results

5.3.1 Volatility Analysis

The mean and standard deviation values calculated for each performance metric and each classifier are presented in Table 5.1, as well as the Δ values, which represent how overfitted the model is.

In Table 5.2, only the Δ values are shown. Each column is individually color-coded from lowest value, in green, to highest value, in red. As previously, there seems to be a cluster of overfitted models on the bottom of the table (in darker red). These correspond to the pipelines trained with Lesion data. Three clusters of lower Δ can be found in green, these correspond to the pipelines where downsampling of the majority class was performed.

In Table 5.3, only the mean values are presented for each performance metric and each column is individually color-coded from highest value, in green, to lowest value, in red. At first glance, we can see that a few of the highest cross-validation performances are in the bottom of the table, while the highest test set performances are higher in the table. This was expected since Table 5.2 showed that these models were the most overfitted. Additionally, from Table 5.3, two pipelines stand out as performing well across all performance metrics: G_S_SVM-RFE_LR and G_S_mRMR_RF.

In Figures 5.1 and 5.2, you can see the plotted distribution of F2 and Kappa performances respectively and the full width at half maximum value.

The last nine graphs show the volatility analysis of the models trained on lesion data. Here, we can clearly distinguish two different peaks, which confirms the previous results that these were the most overfitted models.

As expected, the test set performance distribution is overall shorter and wider than the cross-validation performance distribution, which is taller and thinner. This is clear by the difference in FWHM values.

5.3.2 Normality Tests

In Tables 5.4 and 5.5, the results of the F2 and Kappa performance distribution normality tests are displayed.

Out of 54 F2 distributions (26 test set plus 26 cross-validation performance distributions), 46 were considered, by both tests, not to be significantly different from the Guassian distribution. Out of the remaining 6 F2 distributions, 2 were found to be significantly different from Guassian on both tests, 2 were found to be significantly different from Gaussian only on the Shapiro-Wilk test, 1 was found to be significantly different from Gaussian only on the D'Agostino's K^2 test and 1 was inconclusive.

Out of 54 Kappa distributions, 52 were considered, by both tests, not to be significantly different from the Guassian distribution. One of the remaining distributions was found to be significantly different from Guassian on both tests and the other was found to be significantly different from Gaussian only on the D'Agostino's K^2 test, accompanied by a rather low p-value on the Shapiro-Wilk test.

Models	F2				Kappa				AUC				AUPRC				Δ (CV - TS)			
	CV		TS		CV		TS		CV		TS		CV		TS					
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	F2	Kappa	AUC	AUPRC
G_D_SVM-RFE_LR	0.6447	0.0582	0.6060	0.0910	0.3103	0.1095	0.2754	0.1177	0.7041	0.0607	0.7130	0.0646	0.6308	0.0465	0.6281	0.1745	0.0388	0.0349	-0.0089	0.0027
G_D_SVM-RFE_RF	0.6734	0.0651	0.6207	0.0957	0.3278	0.1174	0.2678	0.1199	0.7195	0.0782	0.7076	0.0688	0.6502	0.0584	0.6286	0.1638	0.0527	0.0601	0.0119	0.0215
G_D_SVM-RFE_XGB	0.6538	0.0832	0.6309	0.1085	0.3168	0.1286	0.3033	0.1080	0.7074	0.0722	0.7072	0.0640	0.6371	0.0565	0.6159	0.1838	0.0230	0.0135	0.0002	0.0212
G_S_SVM-RFE_LR	0.8011	0.0302	0.6944	0.0822	0.4376	0.0683	0.2762	0.1193	0.7721	0.0324	0.7102	0.0690	0.7156	0.0410	0.6245	0.1752	0.1067	0.1614	0.0619	0.0911
G_S_SVM-RFE_DT	0.7939	0.0312	0.6484	0.0990	0.3893	0.0827	0.1906	0.1280	0.7357	0.0487	0.6336	0.0824	0.5059	0.0912	0.4347	0.1954	0.1454	0.1987	0.1021	0.0712
G_S_SVM-RFE_RF	0.7967	0.0278	0.6318	0.0963	0.4422	0.0671	0.2311	0.1195	0.8122	0.0278	0.6838	0.0747	0.7662	0.0272	0.5959	0.1722	0.1649	0.2110	0.1284	0.1703
G_S_SVM-RFE_XGB	0.7509	0.0462	0.5627	0.1085	0.4599	0.0752	0.2364	0.1450	0.7986	0.0318	0.6718	0.0733	0.7469	0.0366	0.5786	0.1704	0.1881	0.2235	0.1268	0.1683
G_D_mRMR_LR_EN	0.6445	0.0546	0.6109	0.0731	0.3174	0.1084	0.2740	0.1015	0.7135	0.0699	0.7097	0.0624	0.6415	0.0585	0.6274	0.1777	0.0336	0.0434	0.0039	0.0141
G_D_mRMR_DT	0.6043	0.1085	0.6168	0.1652	0.2307	0.1250	0.2207	0.0960	0.6583	0.0834	0.6567	0.0606	0.5721	0.0812	0.5641	0.1742	-0.0125	0.0100	0.0016	0.0080
G_D_mRMR_RF	0.6985	0.0635	0.6594	0.0778	0.3715	0.1103	0.3279	0.1120	0.7330	0.0621	0.7360	0.0648	0.6578	0.0527	0.6570	0.1593	0.0390	0.0436	-0.0030	0.0007
G_D_mRMR_XGB	0.6381	0.0607	0.6188	0.1029	0.3091	0.1076	0.2907	0.1295	0.7058	0.0646	0.6976	0.0862	0.6343	0.0538	0.6175	0.1671	0.0193	0.0184	0.0082	0.0169
G_S_mRMR_DT	0.8257	0.0309	0.6744	0.0757	0.4070	0.0907	0.2260	0.0954	0.7191	0.0434	0.6411	0.0585	0.4480	0.0857	0.3913	0.1666	0.1513	0.1809	0.0780	0.0567
G_S_mRMR_RF	0.8204	0.0296	0.6669	0.0782	0.4850	0.0617	0.2757	0.1116	0.8318	0.0298	0.7283	0.0595	0.7810	0.0305	0.6645	0.1473	0.1535	0.2093	0.1035	0.1165
G_S_mRMR_XGB	0.7490	0.0487	0.5749	0.0967	0.4706	0.0825	0.2607	0.1193	0.8041	0.0357	0.6764	0.0699	0.7544	0.0377	0.5802	0.1835	0.1741	0.2099	0.1276	0.1741
G_D_Lasso_DT	0.6755	0.0738	0.6785	0.0972	0.2788	0.1207	0.2810	0.1136	0.6756	0.0688	0.6784	0.0638	0.5255	0.0748	0.4893	0.1560	-0.0030	-0.0021	-0.0028	0.0362
G_D_Lasso_RF	0.7027	0.0673	0.6779	0.0884	0.3570	0.1144	0.3266	0.1153	0.7213	0.0725	0.7376	0.0728	0.6489	0.0617	0.6530	0.1565	0.0249	0.0304	-0.0163	-0.0041
G_D_Lasso_XGB	0.6590	0.0681	0.6317	0.0985	0.3173	0.0988	0.2875	0.1177	0.7117	0.0703	0.7060	0.0712	0.6386	0.0598	0.6218	0.1838	0.0273	0.0298	0.0058	0.0168
L_S_Lasso_XGB	0.7987	0.0242	0.4141	0.0960	0.5295	0.0457	0.1490	0.0990	0.8500	0.0190	0.6176	0.0610	0.8208	0.0229	0.4277	0.1933	0.3846	0.3805	0.2324	0.3931
Lp_D_SVM-RFE_LR_EN	0.6065	0.0759	0.5396	0.1112	0.2910	0.0984	0.2417	0.1041	0.6913	0.0562	0.7028	0.0715	0.6280	0.0542	0.4944	0.1902	0.0668	0.0493	-0.0115	0.1336
Lp_S_Boruta_XGB	0.7907	0.0277	0.2506	0.2442	0.5480	0.0520	0.0046	0.1402	0.8617	0.0205	0.4881	0.1042	0.8365	0.0230	0.2717	0.1541	0.5401	0.5434	0.3736	0.5648
Lp_S_mRMR_NB	0.7704	0.0453	0.5169	0.1168	0.4336	0.0780	0.2827	0.1331	0.7850	0.0309	0.7045	0.0638	0.7399	0.0313	0.4364	0.2120	0.2534	0.1510	0.0805	0.3035
Lp_S_mRMR_LR	0.8427	0.0178	0.5514	0.1006	0.3930	0.0662	0.2398	0.1102	0.7892	0.0264	0.6803	0.0676	0.7446	0.0374	0.4711	0.1940	0.2913	0.1532	0.1089	0.2735
Lp_S_mRMR_LR_EN	0.8397	0.0194	0.5388	0.1064	0.3781	0.0794	0.2328	0.1240	0.7814	0.0300	0.6840	0.0708	0.7360	0.0373	0.4729	0.1942	0.3009	0.1453	0.0973	0.2631
Lp_S_mRMR_RF	0.8454	0.0203	0.5705	0.0807	0.4940	0.0711	0.2463	0.0945	0.8556	0.0216	0.6925	0.0618	0.8274	0.0228	0.4832	0.1935	0.2749	0.2477	0.1631	0.3441
Lp_S_mRMR_XGB	0.8000	0.0319	0.5152	0.1058	0.5530	0.0528	0.1861	0.1203	0.8560	0.0236	0.6608	0.0732	0.8260	0.0248	0.4621	0.1957	0.2848	0.3669	0.1952	0.3639
Lp_S_Lasso_XGB	0.8012	0.0234	0.5124	0.0990	0.5588	0.0465	0.1614	0.0955	0.8642	0.0211	0.6505	0.0594	0.8364	0.0228	0.4513	0.1910	0.2888	0.3974	0.2137	0.3851

Table 5.1: Mean and standard deviation values calculated for each performance metric and each classifier during the volatility analysis.

Models	Δ (CV - TS)			
	F2	Kappa	AUC	AUPRC
G_D_SVM-RFE_LR	0.0388	0.0349	-0.0089	0.0027
G_D_SVM-RFE_RF	0.0527	0.0601	0.0119	0.0215
G_D_SVM-RFE_XGB	0.0230	0.0135	0.0002	0.0212
G_S_SVM-RFE_LR	0.1067	0.1614	0.0619	0.0911
G_S_SVM-RFE_DT	0.1454	0.1987	0.1021	0.0712
G_S_SVM-RFE_RF	0.1649	0.2110	0.1284	0.1703
G_S_SVM-RFE_XGB	0.1881	0.2235	0.1268	0.1683
G_D_mRMR_LR_EN	0.0336	0.0434	0.0039	0.0141
G_D_mRMR_DT	-0.0125	0.0100	0.0016	0.0080
G_D_mRMR_RF	0.0390	0.0436	-0.0030	0.0007
G_D_mRMR_XGB	0.0193	0.0184	0.0082	0.0169
G_S_mRMR_DT	0.1513	0.1809	0.0780	0.0567
G_S_mRMR_RF	0.1535	0.2093	0.1035	0.1165
G_S_mRMR_XGB	0.1741	0.2099	0.1276	0.1741
G_D_Lasso_DT	-0.0030	-0.0021	-0.0028	0.0362
G_D_Lasso_RF	0.0249	0.0304	-0.0163	-0.0041
G_D_Lasso_XGB	0.0273	0.0298	0.0058	0.0168
L_S_Lasso_XGB	0.3846	0.3805	0.2324	0.3931
Lp_D_SVM-RFE_LR_EN	0.0668	0.0493	-0.0115	0.1336
Lp_S_Boruta_XGB	0.5401	0.5434	0.3736	0.5648
Lp_S_mRMR_NB	0.2534	0.1510	0.0805	0.3035
Lp_S_mRMR_LR	0.2913	0.1532	0.1089	0.2735
Lp_S_mRMR_LR_EN	0.3009	0.1453	0.0973	0.2631
Lp_S_mRMR_RF	0.2749	0.2477	0.1631	0.3441
Lp_S_mRMR_XGB	0.2848	0.3669	0.1952	0.3639
Lp_S_Lasso_XGB	0.2888	0.3974	0.2137	0.3851

Table 5.2: Delta values calculated for each performance metric and each classifier during the volatility analysis. Each column is individually color-coded from lowest value, in green, to highest value, in red.

5.3.3 Distribution Comparison Tests

In Table 5.6, we can see the results of the comparison between cross-validation F2 performance distribution and test set F2 performance distribution. Out of 26 classifiers, 19 classifiers displayed a significant difference between the test set performance distribution and the cross-validation performance distribution, 5 classifiers displayed no significant difference between the test set performance distribution and the cross-validation performance distribution, 1 classifier displayed a significant difference on the Kolmogorov-Smirnov test but no difference on the paired t-test and 1 classifier displayed a significant difference on the Kolmogorov-Smirnov test but inconclusive results on the paired t-test.

In Table 5.7, we can see the results of the comparison between cross-validation Kappa performance distribution and test set Kappa performance distribution. Out of 26 classifiers, 15 classifiers displayed a significant difference between the test set performance distribution and the cross-validation performance distribution, 8 classifiers displayed no significant difference between the test set performance distribution and the cross-validation performance distribution, 1 classifier displayed a significant difference on the Kolmogorov-Smirnov test but no difference on the paired t-test, 1 classifier displayed a significant difference on the paired t-test test but no difference on the Kolmogorov-Smirnov and 1 classifier displayed a significant difference on the Kolmogorov-Smirnov test but inconclusive results on the paired t-test.

5 classifiers displayed no significant difference between the cross-validation performance and the test

Models	mean F2		mean Kappa		mean AUC		mean AUPRC	
	CV	TS	CV	TS	CV	TS	CV	TS
G_D_SVM-RFE_LR	0.64473	0.60595	0.31029	0.27535	0.70412	0.71304	0.63085	0.62812
G_D_SVM-RFE_RF	0.6734	0.6207	0.32782	0.26775	0.71951	0.70757	0.65015	0.62862
G_D_SVM-RFE_XGB	0.65382	0.63087	0.31676	0.30325	0.70742	0.70717	0.63706	0.6159
G_S_SVM-RFE_LR	0.80108	0.69442	0.4376	0.27623	0.77209	0.71023	0.71558	0.62452
G_S_SVM-RFE_DT	0.79387	0.64844	0.38932	0.1906	0.73567	0.6336	0.50591	0.43468
G_S_SVM-RFE_RF	0.79673	0.63181	0.44216	0.23111	0.8122	0.68378	0.76622	0.59593
G_S_SVM-RFE_XGB	0.75087	0.56273	0.45986	0.23636	0.79862	0.67183	0.74692	0.57859
G_D_mRMR_LR_EN	0.64451	0.61088	0.31738	0.27399	0.71352	0.70965	0.64155	0.62744
G_D_mRMR_DT	0.60428	0.61678	0.23072	0.22071	0.6583	0.65665	0.57206	0.56409
G_D_mRMR_RF	0.69847	0.65942	0.37147	0.32789	0.73303	0.73601	0.65777	0.65705
G_D_mRMR_XGB	0.63805	0.61879	0.30911	0.29069	0.70579	0.69763	0.63435	0.61748
G_S_mRMR_DT	0.82574	0.6744	0.40696	0.22604	0.71907	0.64106	0.448	0.39126
G_S_mRMR_RF	0.82041	0.66691	0.48496	0.27566	0.8318	0.72826	0.78096	0.66447
G_S_mRMR_XGB	0.74902	0.57491	0.47058	0.26069	0.80407	0.67643	0.75436	0.58023
G_D_Lasso_DT	0.67547	0.67849	0.27881	0.28095	0.67559	0.67843	0.52553	0.48929
G_D_Lasso_RF	0.70273	0.67787	0.357	0.32662	0.7213	0.73759	0.64891	0.65298
G_D_Lasso_XGB	0.65899	0.63168	0.31733	0.28748	0.71175	0.70599	0.63859	0.62182
L_S_Lasso_XGB	0.7987	0.41407	0.52949	0.14902	0.84997	0.61757	0.82082	0.42773
Lp_D_SVM-RFE_LR_EN	0.60645	0.53964	0.29102	0.24173	0.69131	0.70282	0.62803	0.49442
Lp_S_Boruta_XGB	0.79073	0.25063	0.54801	0.00459	0.86166	0.4881	0.8365	0.27165
Lp_S_mRMR_NB	0.77035	0.51693	0.43364	0.28268	0.78504	0.70452	0.73995	0.43642
Lp_S_mRMR_LR	0.84266	0.55141	0.39304	0.2398	0.7892	0.68034	0.74462	0.47112
Lp_S_mRMR_LR_EN	0.83969	0.53881	0.3781	0.2328	0.78136	0.68403	0.73596	0.47288
Lp_S_mRMR_RF	0.8454	0.57052	0.494	0.24625	0.85555	0.69247	0.82736	0.48323
Lp_S_mRMR_XGB	0.80004	0.51524	0.55298	0.18609	0.85604	0.66083	0.82599	0.46209
Lp_S_Lasso_XGB	0.80116	0.51236	0.55877	0.16138	0.86418	0.65045	0.83639	0.45127

Table 5.3: Mean values calculated for each performance metric and each classifier during the volatility analysis. Each column is individually color-coded from highest value, in green, to lowest value, in red.

set performance on both performance metrics, these were: G_D_SVM-RFE_XGB, G_D_mRMR_XGB, G_D_Lasso_DT, G_D_Lasso_RF and G_D_Lasso_XGB. These were also among the classifiers found to be least overfitted in the previous section, supporting those results.

5.3.4 Comparison with Dummy Classifier

For further validation of the results, the 4-fold cross-validation performance of the 5 classifiers found in the previous section was compared with the 4-fold cross-validation performance of a ”dummy” classifier. These results across all four performance metrics can be found in Figure 5.3.

We can confirm that our 5 classifiers perform higher than the dummy classifier across all four performance metrics.

5.4 Discussion

In this context, the Lesion-based models seem to be the most susceptible to selection bias, as they are the most overfitted. This result supports the findings of the previous chapter, in that the features extracted from the lesion VOI do not produce as reliable classifiers as the ones extracted from the whole gland VOI.

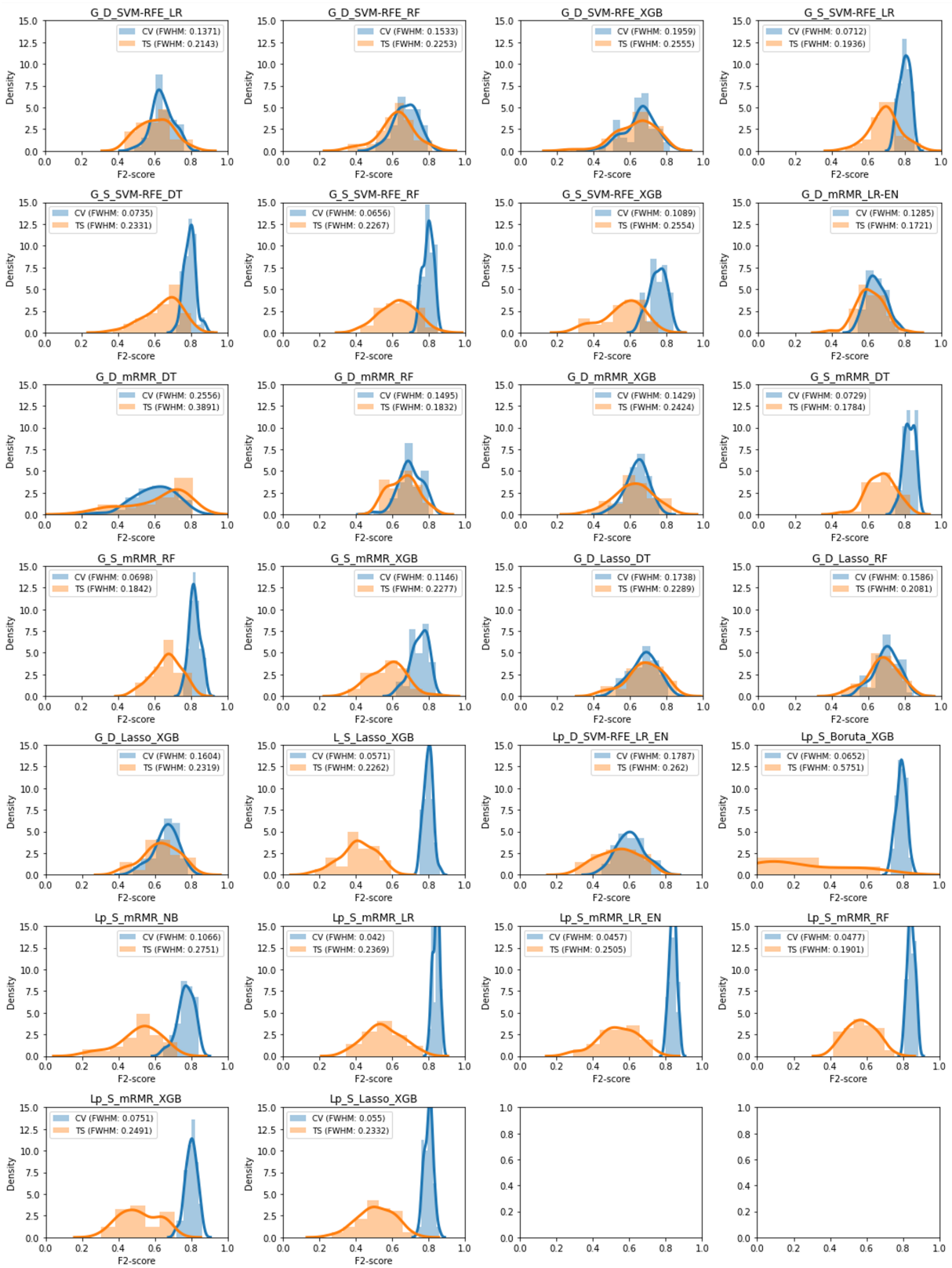

Figure 5.1: Distribution of F2 performances obtained during the volatility analysis for each classifier.

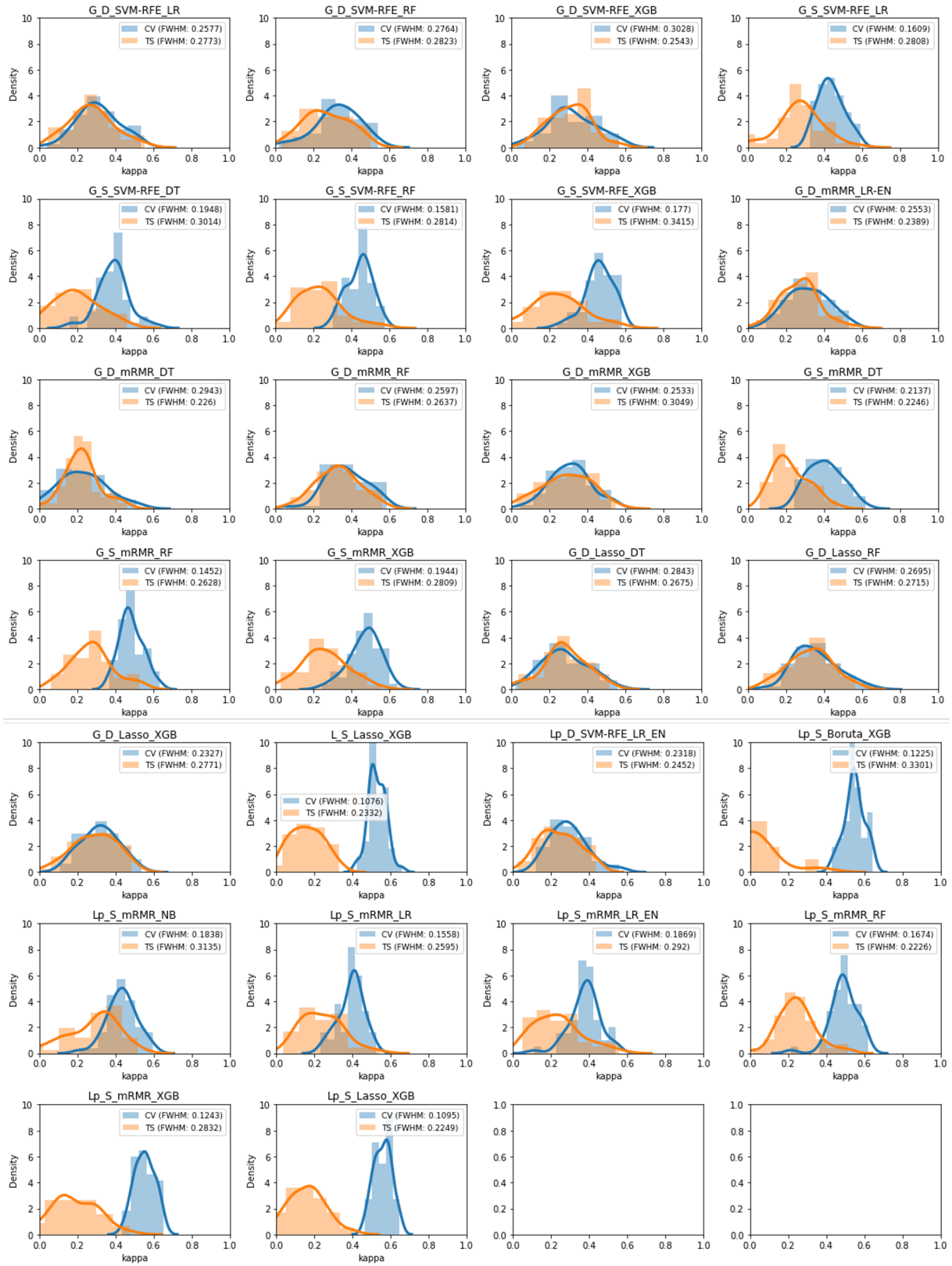

Figure 5.2: Distribution of Kappa performances obtained during the volatility analysis for each classifier.

Models		F2 Distribution Normality tests		Decision
		Shapiro-Wilk	D'Agostino's K^2	
G_D_SVM-RFE_LR	CV	0.39	0.909	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.542	0.503	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_SVM-RFE_RF	CV	0.702	0.545	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.231	0.145	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_SVM-RFE_XGB	CV	0.135	0.308	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.017	0.018	Reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_LR	CV	0.291	0.159	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.839	0.629	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_DT	CV	0.465	0.484	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.048	0.09	
G_S_SVM-RFE_RF	CV	0.135	0.191	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.894	0.862	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_XGB	CV	0.234	0.335	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.034	0.153	
G_D_mRMR_LR_EN	CV	0.441	0.569	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.799	0.457	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_DT	CV	0.593	0.464	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0	0.031	Reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_RF	CV	0.079	0.152	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.31	0.363	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_XGB	CV	0.725	0.372	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.795	0.767	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_mRMR_DT	CV	0.127	0.411	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.691	0.6	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_mRMR_RF	CV	0.742	0.892	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.348	0.486	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_mRMR_XGB	CV	0.054	0.132	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.749	0.944	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_Lasso_DT	CV	0.435	0.527	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.137	0.196	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_Lasso_RF	CV	0.547	0.694	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.427	0.376	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_Lasso_XGB	CV	0.167	0.171	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.392	0.527	Fail to reject the null hypothesis that the distribution is Gaussian
L_S_Lasso_XGB	CV	0.811	0.971	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.28	0.424	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_D_SVM-RFE_LR_EN	CV	0.706	0.806	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.218	0.156	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_Boruta_XGB	CV	0.529	0.842	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	1	nan	
Lp_S_mRMR_NB	CV	0.067	0.093	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.055	0.064	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_LR	CV	0.244	0.598	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.784	0.785	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_LR_EN	CV	0.688	0.857	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.322	0.501	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_RF	CV	0.729	0.518	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.792	0.533	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_XGB	CV	0.763	0.547	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.115	0.043	
Lp_S_Lasso_XGB	CV	0.695	0.937	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.91	0.743	Fail to reject the null hypothesis that the distribution is Gaussian

Table 5.4: Results of the F2 performance distribution normality tests for each classifier.

Models		Kappa Distribution Normality tests		Decision
		Shapiro-Wilk	D'Agostino's K^2	
G_D_SVM-RFE_LR	CV	0.861	0.965	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.736	0.701	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_SVM-RFE_RF	CV	0.259	0.167	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.498	0.284	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_SVM-RFE_XGB	CV	0.512	0.933	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.693	0.982	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_LR	CV	0.36	0.428	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.556	0.446	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_DT	CV	0.142	0.161	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.954	0.97	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_RF	CV	0.4	0.539	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.136	0.067	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_SVM-RFE_XGB	CV	0.097	0.101	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.846	0.885	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_LR_EN	CV	0.68	0.568	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.57	0.379	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_DT	CV	0.577	0.621	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.168	0.273	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_RF	CV	0.349	0.786	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.919	0.95	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_mRMR_XGB	CV	0.669	0.957	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.581	0.486	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_mRMR_DT	CV	0.62	0.368	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.114	0.234	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_mRMR_RF	CV	0.563	0.591	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.188	0.479	Fail to reject the null hypothesis that the distribution is Gaussian
G_S_mRMR_XGB	CV	0.43	0.241	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.71	0.77	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_Lasso_DT	CV	0.855	0.736	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.347	0.879	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_Lasso_RF	CV	0.22	0.321	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.842	0.924	Fail to reject the null hypothesis that the distribution is Gaussian
G_D_Lasso_XGB	CV	0.846	0.683	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.464	0.48	Fail to reject the null hypothesis that the distribution is Gaussian
L_S_Lasso_XGB	CV	0.419	0.612	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.317	0.362	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_D_SVM-RFE_LR_EN	CV	0.13	0.09	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.331	0.303	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_Boruta_XGB	CV	0.295	0.855	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	1	nan	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_NB	CV	0.93	0.498	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.051	0.172	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_LR	CV	0.245	0.372	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.304	0.317	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_LR_EN	CV	0.077	0.022	
	TS	0.051	0.168	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_RF	CV	0.004	0	Reject the null hypothesis that the distribution is Gaussian
	TS	0.289	0.1	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_mRMR_XGB	CV	0.6	0.4	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.675	0.799	Fail to reject the null hypothesis that the distribution is Gaussian
Lp_S_Lasso_XGB	CV	0.554	0.425	Fail to reject the null hypothesis that the distribution is Gaussian
	TS	0.891	0.688	Fail to reject the null hypothesis that the distribution is Gaussian

Table 5.5: Results of the Kappa performance distribution normality tests for each classifier.

Models	F2 Distribution Comparison tests		Decision
	paired t-test	Kolmogorov-Smirnov	
G_D_SVM-RFE_LR	0.021	0.006	Reject the null hypothesis that the distributions are identical
G_D_SVM-RFE_RF	0.003	0.022	Reject the null hypothesis that the distributions are identical
G_D_SVM-RFE_XGB	0.248	0.396	Fail to reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_LR	0	0	Reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_DT	0	0	Reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_RF	0	0	Reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_XGB	0	0	Reject the null hypothesis that the distributions are identical
G_D_mRMR_LR_EN	0.014	0.012	Reject the null hypothesis that the distributions are identical
G_D_mRMR_DT	0.633	0.022	
G_D_mRMR_RF	0.007	0.039	Reject the null hypothesis that the distributions are identical
G_D_mRMR_XGB	0.252	0.112	Fail to reject the null hypothesis that the distributions are identical
G_S_mRMR_DT	0	0	Reject the null hypothesis that the distributions are identical
G_S_mRMR_RF	0	0	Reject the null hypothesis that the distributions are identical
G_S_mRMR_XGB	0	0	Reject the null hypothesis that the distributions are identical
G_D_Lasso_DT	0.846	0.549	Fail to reject the null hypothesis that the distributions are identical
G_D_Lasso_RF	0.141	0.396	Fail to reject the null hypothesis that the distributions are identical
G_D_Lasso_XGB	0.085	0.112	Fail to reject the null hypothesis that the distributions are identical
L_S_Lasso_XGB	0	0	Reject the null hypothesis that the distributions are identical
Lp_D_SVM-RFE_LR_EN	0	0.006	Reject the null hypothesis that the distributions are identical
Lp_S_Boruta_XGB	nan	0	
Lp_S_mRMR_NB	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_LR	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_LR_EN	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_RF	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_XGB	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_Lasso_XGB	0	0	Reject the null hypothesis that the distributions are identical

Table 5.6: Results of statistical tests comparing the distributions of F2 performance between the cross-validation and test set setting.

With regards to sampling strategy, while the pipelines where SMOTE upsampling was performed seem to outperform downsampling of the majority class, the latter are consistently less overfitted and more reliable. Regarding feature selection, there don't seem to be significant differences in the metrics' volatility.

It is known that the difference between two means will follow a normal distribution if the samples are drawn from populations that also follow a normal distribution. However, the central limit theorem states that, even if the parent populations are not Gaussian, the differences will tend towards normality as sample size increases. Since we have a relatively high sample size of 50 and most of our problematic distributions were found to be significantly different from Gaussian on only one of the normality tests, we felt confident assuming normality in the remaining analysis.

As expected, the five models where no significant difference was found between the cross-validation and test set performance distributions (Tables 5.6 and 5.7) were also among the least overfitted models found in Table 5.2. These were all models trained with data that underwent downsampling of the majority class, in addition to all being tree-based machine learning algorithms. The validity of the 5 models with no significant overfitting was further confirmed with their comparison with a dummy classifier.

Models	Kappa Distribution Comparison tests		Decision
	paired t-test	Kolmogorov-Smirnov	
G_D_SVM-RFE_LR	0.163	0.272	Fail to reject the null hypothesis that the distributions are identical
G_D_SVM-RFE_RF	0.025	0.022	Reject the null hypothesis that the distributions are identical
G_D_SVM-RFE_XGB	0.574	0.717	Fail to reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_LR	0	0	Reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_DT	0	0	Reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_RF	0	0	Reject the null hypothesis that the distributions are identical
G_S_SVM-RFE_XGB	0	0	Reject the null hypothesis that the distributions are identical
G_D_mRMR_LR_EN	0.063	0.039	
G_D_mRMR_DT	0.641	0.396	Fail to reject the null hypothesis that the distributions are identical
G_D_mRMR_RF	0.077	0.179	Fail to reject the null hypothesis that the distributions are identical
G_D_mRMR_XGB	0.483	0.396	Fail to reject the null hypothesis that the distributions are identical
G_S_mRMR_DT	0	0	Reject the null hypothesis that the distributions are identical
G_S_mRMR_RF	0	0	Reject the null hypothesis that the distributions are identical
G_S_mRMR_XGB	0	0	Reject the null hypothesis that the distributions are identical
G_D_Lasso_DT	0.924	0.717	Fail to reject the null hypothesis that the distributions are identical
G_D_Lasso_RF	0.225	0.717	Fail to reject the null hypothesis that the distributions are identical
G_D_Lasso_XGB	0.164	0.396	Fail to reject the null hypothesis that the distributions are identical
L_S_Lasso_XGB	0	0	Reject the null hypothesis that the distributions are identical
Lp_D_SVM-RFE_LR_EN	0.042	0.068	
Lp_S_Boruta_XGB	nan	0	
Lp_S_mRMR_NB	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_LR	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_LR_EN	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_RF	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_mRMR_XGB	0	0	Reject the null hypothesis that the distributions are identical
Lp_S_Lasso_XGB	0	0	Reject the null hypothesis that the distributions are identical

Table 5.7: Results of statistical tests comparing the distributions of Kappa performance between the cross-validation and test set setting.

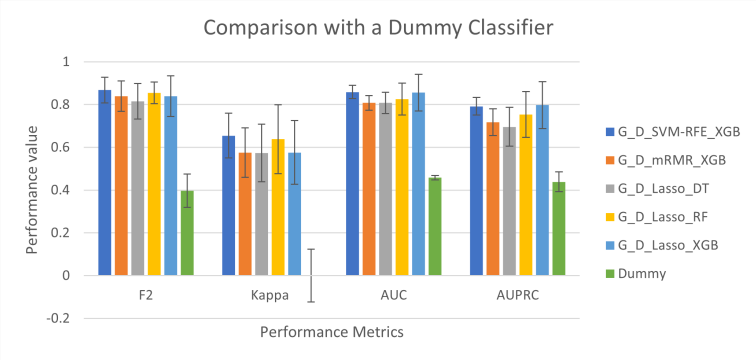

Figure 5.3: Comparison of 4-fold cross validation performance of 5 classifiers with no significant overfitting with the 4-fold cross validation performance of a dummy classifier.

Chapter 6

Conclusion

In this work, an extensive analysis of different dimensions of a machine learning pipeline were assessed and their performance compared. Since there is little consensus on what is the "right way" to perform AI in the context of medical imaging, it is interesting to test which aspects lead to a higher model performance and reliability, especially with such a widely used dataset.

Polarizing areas of AI in medical imaging such as whether or not to perform lesion segmentation or whether to sample the data in contrast to allowing the model to learn from the real label distribution were assessed in this study. And while we should proceed with caution when extrapolating to different settings, these results are still worth analysing.

Among the most interesting findings is the higher performance of models trained with radiomic features extracted from the whole gland VOI, as well as their higher reliability and lower overfitting. This suggests that the areas surrounding tumorous lesions might offer relevant information regarding their overall aggressiveness in the form of Gleason score. It is of note though that a much higher number of features was excluded from the Lesion Dataset during the stability to segmentation analysis than from the Gland Dataset. Despite being of low robustness to segmentation margins, these excluded features might have brought forth useful information and be partly at fault for the lower performance of the models trained with the Lesion Dataset.

The metric volatility analysis performed in this study is not commonly found in the literature. Despite this, we felt it added valuable insight into how the model would perform in the "wild", since multiple hold-out test sets were not available. An interesting result found here was that the widely used SMOTE technique results in models that are more overfitted than models trained with data that went through a simple downsampling of the majority class. This can be explained by the fact that SMOTE generates synthetic samples from the existing samples in the dataset. Thus, we are forcing the model to learn more from the same data, increasing the model's confidence in random variability, or noise, present in the data. Which results in the overfitted behaviour.

Despite these efforts, proper assessment of real-world clinical performance is only possible through external validation. An appropriately built external dataset is one that represents all relevant variations

of patient spectrum (for example: patient demographics, MRI scanner brand, patient age, disease aggressiveness, etc.). Hence the importance of validating with data from multiple external institutions. This important validation step will be addressed in future work.

This study has several limitations. First, this was a retrospective study and, so, a multicentre prospective analysis should be carried out to validate these results and investigate the impact these predictive models have on patient outcome. Second, only T2W, DWI and ADC sequences were used. Other sequences, such as MR spectroscopy and dynamic contrast enhanced MRI, could be worth exploring. Third, only one set of MRI sequences was evaluated per patient, so we were unable to evaluate the temporal stability of the radiomic features. Fourth, although the overall class imbalancement was addressed through downsampling of the majority class or SMOTE upsampling of the minority class, we did not address the imbalanced nature of the anatomical location of lesions, with the large majority of lesions belonging to the PZ. It would be interesting to investigate the model's performance on the different anatomical zones independently. Fifth, the use of a publicly available dataset increased transparency but limited our access to clinical data, such as PSA levels, patient age or PI-RADS score, which are a fundamental component of a clinician's assessment, but could not be included in our model. Finally, inherent to the Gleason system is the subjectivity of cancer grading, so we must keep in mind that the gold standard used in this study is subject to human error and inter or intra-observer variability.

In conclusion, our preliminary results further confirm the validity of MRI-based radiomic features in the identification of clinically significant prostate cancer lesions. The proposed noninvasive models, based on T2W, DWI and ADC maps, showed potencial for aiding clinical decision-makings for patients with a suspicion of prostate cancer.

References

[1] Abd-Alazeez, M., Ahmed, H. U., Arya, M., Charman, S. C., Anastasiadis, E., Freeman, A., Emberton, M., and Kirkham, A. (2014). The accuracy of multiparametric mri in men with negative biopsy and elevated psa level—can it rule out clinically significant prostate cancer? In *Urologic Oncology: Seminars and Original Investigations*, volume 32, pages 45–e17. Elsevier. 3

[2] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer. 18

[3] Boesen, L. (2017). Multiparametric mri in detection and staging of prostate cancer. *Danish medical journal*, 64(2). 2, 3, 4, 6

[4] Boesen, L., Chabanova, E., Løgager, V., Balslev, I., and Thomsen, H. S. (2015). Apparent diffusion coefficient ratio correlates significantly with prostate cancer gleason score at final pathology. *Journal of Magnetic Resonance Imaging*, 42(2):446–453. 8

[5] Borkenhagen, J. F., Eastwood, D., Kilari, D., See, W. A., Van Wickle, J. D., Lawton, C. A., and Hall, W. A. (2019). Digital rectal examination remains a key prognostic tool for prostate cancer: a national cancer database review. *Journal of the National Comprehensive Cancer Network*, 17(7):829–837. 1

[6] Branco, P., Torgo, L., and Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*. 27

[7] Catalona, W. J., Richie, J. P., Ahmann, F. R., Hudson, M. A., Scardino, P. T., Flanigan, R. C., Dekernion, J. B., Ratliff, T. L., Kavoussi, L. R., Dalkin, B. L., et al. (1994). Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. *The Journal of urology*, 151(5):1283–1290. 1

[8] Catalona, W. J., Smith, D. S., Ratliff, T. L., Dodds, K. M., Coplen, D. E., Yuan, J. J., Petros, J. A., and Andriole, G. L. (1991). Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *New England Journal of Medicine*, 324(17):1156–1161. 1

[9] Delongchamps, N. B., Peyromaure, M., Schull, A., Beuvon, F., Bouazza, N., Flam, T., Zerbib, M., Muradyan, N., Legman, P., and Cornud, F. (2013). Prebiopsy magnetic resonance imaging and prostate cancer detection: comparison of random and targeted biopsies. *The Journal of urology*, 189(2):493–499. 3

[DGS] DGS. Mortalidade em portugal. 8

[11] Egawa, S., Wheeler, T., Greene, D., and Scardino, P. (1992). Unusual hyperechoic appearance of prostate cancer on transrectal ultrasonography. *British journal of urology*, 69(2):169–174. 2

[12] Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., and Humphrey, P. A. (2016). The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252. 7

[13] Epstein, J. I., Feng, Z., Trock, B. J., and Pierorazio, P. M. (2012). Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades. *European urology*, 61(5):1019–1024. 8

[14] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484. 27

[15] George, A. K., Turkbey, B., Valayil, S. G., Muthigi, A., Mertan, F., Kongnyuy, M., and Pinto, P. A. (2016). A urologist's perspective on prostate cancer imaging: past, present, and future. *Abdominal Radiology*, 41(5):805–816. 2, 3

[16] Gleason, D. F. (1992). Histologic grading of prostate cancer: a perspective. *Human pathology*, 23(3):273–279. 6, 7

[17] Haider, M., Yao, X., Loblaw, A., and Finelli, A. (2016). Multiparametric magnetic resonance imaging in the diagnosis of prostate cancer: a systematic review. *Clinical oncology*, 28(9):550–567. 3

[18] Halpern, J. A., Oromendia, C., Shoag, J. E., Mittal, S., Cosiano, M. F., Ballman, K. V., Vickers, A. J., and Hu, J. C. (2018). Use of digital rectal examination as an adjunct to prostate specific antigen in the detection of clinically significant prostate cancer. *The Journal of urology*, 199(4):947–953. 1

[19] Haythorn, M. R. and Ablin, R. J. (2011). Prostate-specific antigen testing across the spectrum of prostate cancer. *Biomarkers in medicine*, 5(4):515–526. 1

[20] Hernández, J. and Thompson, I. M. (2004). Prostate-specific antigen: a review of the validation of the most commonly used cancer biomarker. *Cancer*, 101(5):894–904. 2

[21] Hoeks, C. M., Schouten, M. G., Bomers, J. G., Hoogendoorn, S. P., Hulsbergen-van de Kaa, C. A., Hambrock, T., Vergunst, H., Sedelaar, J. M., Fütterer, J. J., and Barentsz, J. O. (2012). Three-tesla magnetic resonance–guided prostate biopsy in men with increased prostate-specific antigen and repeated, negative, random, systematic, transrectal ultrasound biopsies: detection of clinically significant prostate cancers. *European urology*, 62(5):902–909. 3

[22] Komai, Y., Numao, N., Yoshida, S., Matsuoka, Y., Nakanishi, Y., Ishii, C., Koga, F., Saito, K., Masuda, H., Fujii, Y., et al. (2013). High diagnostic ability of multiparametric magnetic resonance imaging to detect anterior prostate cancer missed by transrectal 12-core biopsy. *The Journal of urology*, 190(3):867–873. 3

[23] Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163. 19

[24] Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26. 19

[25] Kumar, V., Bora, G. S., Kumar, R., and Jagannathan, N. R. (2018). Multiparametric (mp) mri of prostate cancer. *Progress in nuclear magnetic resonance spectroscopy*, 105:23–40. 2, 4, 6

[26] Lee, F., McLeary, R., Kumasaka, G., Borlaza, G., Straub, W., Gray, J., Meadows, T., Lee Jr, F., Solomon, M., McHugh, T., et al. (1985). Transrectal ultrasound in the diagnosis of prostate cancer: location, echogenicity, histopathology, and staging. *The Prostate*, 7(2):117–129. 2

[27] Lee, F., Torp-Pedersen, S., Littrup, P., McLeary, R., McHugh, T., Smid, A., Stella, P., and Borlaza, G. (1989). Hypoechoic lesions of the prostate: clinical relevance of tumor size, digital rectal examination, and prostate-specific antigen. *Radiology*, 170(1):29–32. 2

[28] McNeal, J. E. (1981). The zonal anatomy of the prostate. *The prostate*, 2(1):35–49. 1, 2

[29] McNeal, J. E., Redwine, E. A., Freiha, F. S., and Stamey, T. A. (1988). Zonal distribution of prostatic adenocarcinoma. correlation with histologic pattern and direction of spread. *The American journal of surgical pathology*, 12(12):897–906. 1

[30] Mistry, K. and Cable, G. (2003). Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *The Journal of the American Board of Family Practice*, 16(2):95–101. 2

[31] Pal, R. P., Maitra, N. U., Mellon, J. K., and Khan, M. A. (2013). Defining prostate cancer risk before prostate biopsy. In *Urologic Oncology: Seminars and Original Investigations*, volume 31, pages 1408–1418. Elsevier. 2

[32] Papanikolaou, N., Matos, C., and Koh, D. M. (2020). How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging*, 20(1):1–10. 17

[33] Pepe, P., Garufi, A., Priolo, G., Candiano, G., Pietropaolo, F., Pennisi, M., Fraggetta, F., and Aragona, F. (2013). Prostate cancer detection at repeat biopsy: can pelvic phased-array multiparametric mri replace saturation biopsy? *Anticancer research*, 33(3):1195–1199. 3

[34] Peter, A., Lilja, H., Lundwall, Å., and Malm, J. (1998). Semenogelin i and semenogelin ii, the major gel-forming proteins in human semen, are substrates for transglutaminase. *European journal of biochemistry*, 252(2):216–221. 1

[35] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432. 28

[36] Stabile, A., Giganti, F., Rosenkrantz, A. B., Taneja, S. S., Villeirs, G., Gill, I. S., Allen, C., Emberton, M., Moore, C. M., and Kasivisvanathan, V. (2020). Multiparametric mri for prostate cancer diagnosis: current status and future directions. *Nature Reviews Urology*, 17(1):41–61. 6

[37] Thompson, I. M., Pauler, D. K., Goodman, P. J., Tangen, C. M., Lucia, M. S., Parnes, H. L., Minasian, L. M., Ford, L. G., Lippman, S. M., Crawford, E. D., et al. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level≤ 4.0 ng per milliliter. *New England Journal of Medicine*, 350(22):2239–2246. 2

[38] Thörmer, G., Otto, J., Reiss-Zimmermann, M., Seiwerts, M., Moche, M., Garnov, N., Franz, T., Do, M., Stolzenburg, J.-U., Horn, L.-C., et al. (2012). Diagnostic value of adc in patients with prostate cancer: influence of the choice of b values. *European radiology*, 22(8):1820–1828. 8

[39] Toivonen, J., Merisaari, H., Pesola, M., Taimen, P., Boström, P. J., Pahikkala, T., Aronen, H. J., and Jambor, I. (2015). Mathematical models for diffusion-weighted imaging of prostate cancer using b values up to 2000 s/mm2: Correlation with gleason score and repeatability of region of interest analysis. *Magnetic resonance in medicine*, 74(4):1116–1124. 5

[40] Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107. 9, 11, 13

[41] van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., and Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*, 11(1):1–16. 11

[42] Vargas, H. A., Akin, O., Franiel, T., Mazaheri, Y., Zheng, J., Moskowitz, C., Udo, K., Eastham, J., and Hricak, H. (2011). Diffusion-weighted endorectal mr imaging at 3 t for prostate cancer: tumor detection and assessment of aggressiveness. *Radiology*, 259(3):775–784. 8

[43] Wibmer, A., Hricak, H., Gondo, T., Matsumoto, K., Veeraraghavan, H., Fehr, D., Zheng, J., Goldman, D., Moskowitz, C., Fine, S. W., et al. (2015). Haralick texture analysis of prostate mri: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. *European radiology*, 25(10):2840–2850. 8

[44] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82. 26