FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Radiomics: Tumour Genotype Prediction

Catarina Manuela Fonseca Dias

Mestrado Integrado em Bioengenharia

Supervisor: Prof. Hélder Filipe Pinto de Oliveira

Co-Supervisor: Prof. António Manuel Trigueiros da Silva Cunha

July 16, 2019

© Catarina Dias, 2019

Radiomics: Tumour Genotype Prediction

Catarina Manuela Fonseca Dias

Mestrado Integrado em Bioengenharia

July 16, 2019

Resumo

Avanços em genómica levaram ao reconhecimento de que os tumores são constituídos por diferentes subclones de células malignas que influenciam a sua progressão. Portanto, o tratamento deve ser moldado com base na heterogeneidade tumoral individual. A heterogeneidade genómica espacial e temporal dos tumores tem vindo a ser uma barreira na área de medicina de precisão, uma vez que a metodologia-padrão para a análise genómica é a biópsia, que para além de invasiva, não permite capturar o tumor na sua totalidade num único exame.

A imagiologia médica, que já é uma fonte comum de informação na prática clínica, é uma potencial alternativa à biópsia. A imagem contém um grande número de características que, embora não sejam visíveis a olho nu, podem ser importantes para a caracterização do tumor. Assim, o recente campo da radiometria abre novas oportunidades para a análise genómica de um tumor, ao extrair centenas de características quantitativas de imagens médicas que, de maneira não invasiva, fornecem uma visualização completa do estado de um tumor a um nível macroscópico. Tirando partido de técnicas de *machine learning*, o conjunto de caraterísticas extraído pode ser usado para prever fatores genómicos importantes para o planeamento do tratamento oncológico.

Este estudo teve como objetivo investigar em que medida características extraídas de imagens médicas se relacionam com fatores genómicos úteis para a caracterização do tumor, como o estado de mutação do EGFR e do KRAS e o subtipo histológico de Cancro do Pulmão de Não Pequenas Células. Usando o conjunto de caraterísticas extraído, modelos preditivos foram desenvolvidos com recurso a técnicas de *machine learning*. Os modelos foram criados utilizando um conjunto de características semânticas, avaliadas visualmente por radiologistas, características radiómicas, diretamente extraídas da imagem, e características clínicas.

O desempenho dos modelos demonstrou que o estado de mutação do EGFR (AUC=0.79) e KRAS (AUC=0.63) pode ser diferenciado através de imagens médicas usando características radiómicas. No entanto, quando o conjunto de características semânticas foi utilizado, o estado de mutação do KRAS não se manifestou na mesma extensão (AUC=0.56) que o estado de mutação do EGFR (AUC=0.96). O excelente resultado da previsão do estado de mutação do EGFR usando características semânticas gerou a ambição de desenvolver um modelo capaz de prever a presença da característica semântica mais importante - enfisema - usando características radiómicas. O modelo alcançou uma AUC promissora de 0.64. Além disso, foi alcançado um modelo capaz de prever com confiança (AUC=0.70) o subtipo histológico de Cancro do Pulmão de Não Pequenas Células.

Abstract

Advances in genomics have driven to the recognition that tumours are populated by different minor subclones of malignant cells that control the way the tumour progresses. Therefore, the treatment must be shaped based on the individual tumour heterogeneity. The spatial and temporal genomic heterogeneity of tumours has been a hurdle in the field of precision medicine, since the standard methodology for genomic analysis is the biopsy, that besides being an invasive technique it does not capture the full spatial state of a tumour in a single exam.

Medical imaging, which is already a common source of information in clinical practice, is a potential alternative to the biopsy. It contains a large number of features that, although not visible to the naked eye, may be valuable for tumour characterisation. Thus, the recent field of radiomics allows new opportunities for the genomic analysis of a tumour, by extracting hundreds of quantitative features from medical images which, in a non-invasive way, provide a full state visualisation of a tumour at a macroscopic level. Taking advantage of machine learning techniques, the extracted set of features can be used to predict valuable genomic factors for treatment planning.

This study aimed to investigate in which extent features extracted from medical images are related to helpful genotype factors for tumour characterisation, namely EGFR and KRAS mutation status and the Non-Small Cell Lung Cancer histological subtype. Using the extracted set of features, predictive models were developed taking advantage of machine learning techniques. The models were created using a set of semantic features, visually assessed by radiologists, radiomic features, directly extracted from the image, and clinical features.

The performance of the models demonstrated that EGFR (AUC=0.79) and KRAS (AUC=0.63) mutation status can be differentiated through medical images using radiomic features. However, when the set of semantic features was used, the KRAS mutation status did not manifest in the same extent (AUC=0.56) as the EGFR mutation status (AUC=0.96). The outstanding result of EGFR mutation status prediction using semantic features raised the ambition of developing a model which predicts the presence of the most important semantic feature - emphysema - using radiomic features. The model achieved a promising AUC of 0.64. Furthermore, a reliable predictive model (AUC=0.70) was achieved regarding the classification of the Non-Small Cell Lung Cancer histological subtype.

Acknowledgements

The accomplishment of this dissertation would not have been possible without the support of many people. However, there are some people I want to thank in particular.

Foremost, I would like to express my gratitude to my advisor Professor Hélder P. Oliveira for his guidance and permanent availability. He consistently gave me the freedom to take my own decisions, but steered me in the right direction whenever he thought I needed it.

I would also like to thank Professor António Cunha for his permanent eager to help and constant enthusiasm, which motivated me to keep improving my work. I am also grateful for his wise words and ability to think positively even in the most stressful situations.

I extend my appreciation to INESC TEC for all the work conditions and technical support. A special thank to Gil Pinheiro for his assistance and availability to provide me valuable feedback whenever I needed.

Last but not least, I would like to thank my family for always giving me continuous encouragement and support throughout my studies. I would not be where I am without them.

Catarina Dias
Porto, June 2019

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-030263.

Dedicated to my family.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Objectives . . . 2
- 1.3 Contributions . . . 2
- 1.4 Document Structure . . . 3

2 Problem Contextualisation: Lung Cancer **5**
- 2.1 Types of Lung Cancer . . . 5
- 2.2 Tumour Heterogeneity and Clonal Evolution . . . 6
- 2.3 Precision Medicine . . . 7
- 2.4 Summary . . . 8

3 Radiomics **9**
- 3.1 The Concept . . . 9
- 3.2 Radiomic Features . . . 14
  - 3.2.1 Histogram Features (First-order statistics) . . . 14
  - 3.2.2 Morphological Features (Shape-based features) . . . 15
  - 3.2.3 Texture-based Features (Higher-order statistics) . . . 15
- 3.3 Summary . . . 16

4 Literature Review **17**
- 4.1 Genes Mutation Status Prediction . . . 17
  - 4.1.1 EGFR . . . 20
  - 4.1.2 KRAS . . . 27
- 4.2 Histological Subtype Prediciton . . . 28
- 4.3 Clinical Outcomes Prediction . . . 31
- 4.4 Public Datasets . . . 34
- 4.5 Discussion . . . 36
- 4.6 Summary . . . 37

5 Data Preparation **39**
- 5.1 Data Description . . . 39
  - 5.1.1 NSCLC- Radiogenomics Dataset . . . 40
  - 5.1.2 NSCLC- Radiomics Dataset . . . 41
- 5.2 Conversion from RTSTRUCT Contours to 3D Binary Mask . . . 41
- 5.3 CT Images Pre-Processing . . . 43
  - 5.3.1 Hounsfield Units . . . 44
  - 5.3.2 Resampling . . . 44

5.3.3 Normalisation . . . 45
5.4 Extraction of Radiomic Features . . . 45
5.5 Summary . . . 47

6 Mutation Status Prediction **49**
6.1 Semantic Features . . . 49
6.1.1 Dataset . . . 50
6.1.2 Synthetic Minority Over-sampling Technique- Nominal and Continuous . . . 52
6.1.3 Random Forest . . . 53
6.1.4 Multi-layer Perceptron . . . 53
6.1.5 Results . . . 54
6.1.6 Discussion . . . 58
6.2 Radiomic Features . . . 58
6.2.1 Dataset . . . 58
6.2.2 Data Pre-Processing . . . 59
6.2.3 Feature Selection . . . 59
6.2.4 Classification . . . 61
6.2.5 Results . . . 61
6.2.6 Discussion . . . 68
6.3 Predict Emphysema Presence Using Radiomic Features . . . 69
6.3.1 Dataset and Methodology . . . 69
6.3.2 Results . . . 70
6.3.3 Discussion . . . 72
6.4 Summary . . . 72

7 Tumour Histological Subtype Prediction **75**
7.1 Dataset . . . 75
7.2 Methodology . . . 76
7.3 Results . . . 77
7.4 Discussion . . . 80
7.5 Summary . . . 81

8 Conclusions **83**

A List of Extracted Features **85**

B Correlation matrices **89**
B.1 **KRAS Mutation Status Prediction** . . . 89
B.2 **EGFR Mutation Status Prediction** . . . 90
B.3 **Emphysema Presence Prediction** . . . 90
B.4 **NSCLC Histological Subtype Prediction** . . . 91

C Set of Features Used for Emphysema Prediction **93**

D Set of Features Used for NSCLC Histological Subtype Prediction **95**

References **97**

List of Figures

2.1 Scheme of different types and subtypes of lung cancer. 5
2.2 Evolutionary trajectory of a tumour through phylogenetic tree analysis. 7

3.1 Workflow of radiomics. 10
3.2 Example of CT slice and 3D CT scan of the chest. 10
3.3 Tumour segmentation in CT image using a semi-automated algorithm. 11
3.4 Confusion matrix for a binary classification task. 13
3.5 3D representations of a round tumour and a sharp tumour. 15
3.6 Texture differences between tumours obtained by CT imaging. 16

4.1 ROC curves obtained for the prediction of EGFR mutation status, using clinical features alone, radiomic features alone and clinical and radiomic features combined. 20
4.2 ROC curve of the models obtained using the GLSZM radiomic feature size zone non-uniformity normalised and the one obtained using the combination of radiomic and clinical features to predict EGFR mutation status. 21
4.3 Architecture of the deep learning model. 23
4.4 ROC curves of the deep learning model, radiomics model, semantic model and clinical model in the training and validation cohorts. 23
4.5 Architecture of the CNN-based model. 24
4.6 Scheme of the study's framework. 25
4.7 ROC curves of the developed models. 26
4.8 Decision Tree obtained for the prediction of EGFR mutation status using five semantic features. 27
4.9 Decision Tree obtained for the prediction of KRAS mutation status. 28
4.10 Cluster consensus map. 30
4.11 Forest plot of radiomic features, basic metrics and clinical variables, regarding distant metastasis prediction. 34

5.1 Representations of the tumour segmentation mask and corresponding CT image. . 40
5.2 Overlap between a contour instance and the respective CT slice. 42
5.3 Representation which illustrates the meaning of origin and spacing. 43
5.4 Histogram of voxel HU values and respective CT slice. 44
5.5 Schematic of the 3D wavelet transform applied to each CT image. 46
5.6 Overview of the process of feature extraction via *Pyradiomics*. 47

6.1 Five-fold cross-validation grid-search for hyper-parameter tuning. 50
6.2 Example of how to one-hot encode a categorical feature. 52
6.3 Generation of new samples from the minority class using SMOTE-NC. 52
6.4 Random Forest classifier. 53

6.5 Multi-layer perceptron network with two hidden layers. 54
6.6 ROC curve of the RF model and MLP model using clinical and imaging features. 56
6.7 Confusion matrix of the RF and MLP models using clinical and imaging features. 57
6.8 Features importance scores. 57
6.9 Cumulative importance versus the number of features. 62
6.10 Features whose score is different from zero and respective score regarding KRAS mutation status prediction. 62
6.11 ROC curve and confusion matrix obtained by the model with the highest performance for predicting KRAS mutation status. 64
6.12 Cumulative importance versus the number of features regarding EGFR mutation status prediction. 65
6.13 Features whose score is different from zero and respective score regarding EGFR mutation status prediction. 65
6.14 Scatter plot of the two selected features for EGFR mutation status prediction before and after applying SMOTE. 66
6.15 ROC curve and confusion matrix obtained by the model with the highest performance for predicting EGFR mutation status. 67
6.16 SVM decision region boundary between EGFR mutant and wildtype. 68
6.17 Cumulative importance versus the number of features regarding emphysema presence prediction. 70
6.18 Top 10 most important features and respective importance scores regarding emphysema presence prediction. 71
6.19 ROC curve and confusion matrix obtained by the model created to predict emphysema presence. 72

7.1 Cumulative importance versus the number of features regarding histological subtype prediction. 77
7.2 Top 10 most important features and respective importance scores regarding histological subtype prediction. 78
7.3 ROC curves of the RF models created to predict NSCLC histological subtype. . . 79
7.4 Confusion matrix of the RF models created to predict NSCLC histological subtype. 80

B.1 Correlation matrices obtained before and after the removal of highly correlated features regarding KRAS mutation status prediction. 89
B.2 Correlation matrices obtained before and after the removal of highly correlated features regarding EGFR mutation status prediction. 90
B.3 Correlation matrices obtained before and after the removal of highly correlated features regarding emphysema presence prediction. 90
B.4 Correlation matrices obtained before and after the removal of highly correlated features regarding histological subtype prediction. 91

List of Tables

4.1 Overview of published studies regarding genes mutation status prediction. 19
4.2 Overview of the literature regarding histological subtype prediction. 29
4.3 Overview of published studies regarding clinical outcomes prediction. 32
4.4 Comparison of public available datasets for radiomics trials. 35
4.5 Main collected data types and the corresponding number of subjects available in the NSCLC- Radiogenomics dataset. 35
4.6 Main collected data types and the corresponding number of subjects available in the NSCLC-Radiomics-Genomics dataset. 36
4.7 Main collected data types and the corresponding number of subjects available in the NSCLC- Radiomics dataset. 36

5.1 Number of patients available to accomplish the proposed tasks. 48

6.1 Distribution of the semantic data used for KRAS and EGFR mutation status prediction experiments. 51
6.2 Set of hyper-parameter values used in the grid search for RF. 54
6.3 Set of hyper-parameter values used in the grid search for MLP. 55
6.4 Hyper-parameters of the RF model which achieved the highest F-measure in the 5-fold cross-validation. 55
6.5 Hyper-parameters of the MLP model which achieved the highest F-measure in the 5-fold cross-validation. 55
6.6 Results obtained in the different experiments regarding EGFR mutation status prediction. 55
6.7 Clinical data distribution for both KRAS and EGFR-related experiments using radiomic experiments. 59
6.8 Set of hyper-parameter values used in the grid search for GBM. 61
6.9 Set of hyper-parameter values used in the grid-search for LR, K-NN, SVM and RF. 63
6.10 Set of hyper-parameter values of the models which achieved the highest F-measure in the grid-search, regarding KRAS mutation status prediction. 63
6.11 Results obtained on the test set by the different models regarding KRAS mutation status prediction. 63
6.12 Set of hyper-parameter values of the models which achieved the highest F-measure in the grid-search, regarding EGFR mutation status prediction. 66
6.13 Results obtained on the test set by the different models regarding EGFR mutation status prediction. 67
6.14 Clinical data distribution for emphysema prediction experiments. 70
6.15 Set of hyper-parameter values used in the grid search for RF, regarding emphysema presence prediction. 71

6.16 Results obtained in the test set regarding emphysema presence prediction. 71

7.1 Statistics of the dimensions of the tumours in the two used datasets and in the total cohort, in millimetres. 76
7.2 Set of hyper-parameter values used in the grid search, regarding histological subtype prediction. 78
7.3 Hyper-parameters of the models which achieved the highest F-measure in the 5-fold cross-validation, regarding histological subtype prediction. 78
7.4 Results obtained by the two models regarding NSCLC histological subtype prediction. 79

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ALK	Anaplastic Lymphoma Kinase
AUC	Area Under The Curve
CNN	Convolutional Neural Network
CT	Computed Tomography
DFS	Disease-Free Survival
DICOM	Digital Imaging and Communications in Medicine
EGFR	Epidermal Growth Factor Receptor
FN	False Negative
FP	False Positive
GBM	Gradient Boosting Machine
GLCM	Grey Level Co-occurrence Matrix
GLDM	Grey Level Dependence Matrix
GLRLM	Grey Level Run Length Matrix
GLSZM	Grey Level Size Zone Matrix
HU	Hounsfield Unit
K-NN	K-Nearest Neighbour
KRAS	Kristen Rat Sarcoma Viral Oncogene Homolog
LoG	Laplacian of Gaussian
LR	Logistic Regression
MLP	Multi-layer Perceptron
NB	Naive Bayes
NSCLC	Non-Small Cell Lung Cancer
OS	Overall Survival
RF	Random Forest
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics
ROI	Region Of Interest
RTSTRUCT	Radiotherapy Structure Set
SVM	Support Vector Machines
TN	True Negative
TP	True Positive

Chapter 1

Introduction

Lung cancer is the most common cause of cancer death in the world, responsible for nearly 1.6 million deaths annually. It comprises 13% of the total new cancer cases, with a higher incidence in males [1]. According to Travis [2], around 85% of cases in men can be associated with tobacco smoking. In some Western countries, where the smoking habit became successful about 100 years ago, tobacco control programmes have managed to significantly decrease mortality. Sadly, the habit has expanded to many industrialised countries, especially in Asia and in Europe, and there is a worrying trend of rising smoking prevalence in young women. Despite smoking behaviour, other factors strongly contribute to lung cancer such as air pollution and exposure to carcinogens like asbestos, arsenic, radon, and polycyclic aromatic hydrocarbons [3].

Once diagnosed, lung cancer is often in an advanced stage, with 15% or less chance of a 5-year survival [4]. At that stage, tumours are already composed by multiple clonal subpopulations of cancer cells and, consequently, the treatment must be shaped based on the individual tumour heterogeneity. The traditional method of characterising the tumour is by extracting tumour tissue in a biopsy, which is then characterised using genomic-based approaches. In spite of being a successful approach in clinical oncology, repeated biopsies tend to increase medical complications. Further, this type of prognostic demands several biopsies, since the results can vary depending on the part of the tumour that is analysed [5].

1.1 Motivation

Since the biopsy is an invasive technique that does not capture the full state of a tumour, there is the urge to find an alternative way to shape the treatment accordingly to the tumour's heterogeneity.

Medical imaging modalities like Computed Tomography (CT), which are already used in clinical practice for diagnosis, stage assessment and response to treatment, are promising alternatives to invasive biopsy. Not only are medical images accessible at a low price, but they are also non-invasive. Medical images contain a large number of features, which may be valuable for tumour characterisation; however, most of them are not visible to the naked eye and end up being discarded [5]. Radiomics overcomes this situation by extracting a large number of quantitative

features from the medical images which are subsequently used for decision making. Therefore, radiomics is a promising approach to create clinical decision tools for improving personalisation of gene-based lung cancer therapy. By having a reliable predictive model, its results can be used as a second opinion or, even in some cases, to avoid a biopsy [6].

1.2 Objectives

This study aims to provide further advances in the radiomics field by improving the tumour characterisation and building predictive models that support the decision of the physician regarding the choice of treatment.

The goal is to find features extracted from medical images that are related to helpful genotype factors for tumour characterisation, namely Epidermal Growth Factor Receptor (EGFR) mutation status, Kristen Rat Sarcoma Viral Oncogene Homolog (KRAS) mutation status and Non-Small Cell Lung Cancer (NSCLC) histological subtype. In addition, using the relevant set of features, this study proposes to develop predictive models taking advantage of the learning abilities of machine learning techniques.

1.3 Contributions

This dissertation presents the following contributions:

- An experimental methodology to build predictive models useful for the radiomics field, which uses exclusively open-source tools;
- Predictive model of EGFR mutation status, using semantic and clinical features;
- Predictive model of EGFR mutation status using radiomic features;
- Predictive model of KRAS mutation status using radiomic features and clinical features;
- Predictive model of emphysema presence using radiomic and clinical features and clinical features;
- Predictive of the NSCLC histological subtype (Adenocarcinoma or Squamous Cell Carcinoma) using radiomic and clinical features;

Publication

This Dissertation involved the following publication during its period:

Dias, Catarina; Pinheiro, Gil; Cunha, António; P. Oliveira, Hélder. Radiogenomics: Lung Cancer-Related Genes Mutation Status Prediction. In Proceedings of the 9th Iberian Conference of Pattern Recognition and Image Analysis, IbPRIA 2019, Madrid, Spain, July 2019.

1.4 Document Structure

This document is divided into 8 chapters. In each chapter, there is an introductory note with the purpose of the chapter, specifying its content.

The present Chapter introduces the importance of treatment personalisation in lung cancer therapy and the motivation behind radiomic techniques. Moreover, the study's objectives and expected contributions are also presented. Chapter 2 briefly describes lung cancer, focusing on the subclonal heterogeneity and how this condition is currently handled. Chapter 3 introduces Radiomics and describes its standard workflow and the classes of features that are commonly extracted. In Chapter 4, it is covered the relevant radiomic studies that have been done so far. The review presented in Chapter 4 is mainly supported by recent papers to obtain a current view of the Radiomics research. Chapter 5 describes the methodology followed regarding data preparation before developing the predictive models. The methodology and respective results regarding genes mutation status prediction and histological subtype prediction are described separately in Chapters 6 and 7, respectively. At last, Chapter 8 provides a conclusion of the present work and the achieved results.

Chapter 2

Problem Contextualisation: Lung Cancer

In the present Chapter, a brief description of the different types of lung cancer is given in Section 2.1, following the explanation made in [7]. A description of tumour heterogeneity and its evolution over time is explained in Section 2.2, whereas Section 2.3 describes the potential role of precision oncology, the requirements to address tumour heterogeneity and possible solutions.

2.1 Types of Lung Cancer

Lung cancer can be divided into three main types: NSCLC, Small Cell Lung Cancer, and Lung Carcinoid tumour. It is essential to define the lung cancer type since it is related to the treatment options and prognosis. In Figure 2.1 is presented a scheme of the different lung tumour types and subtypes.

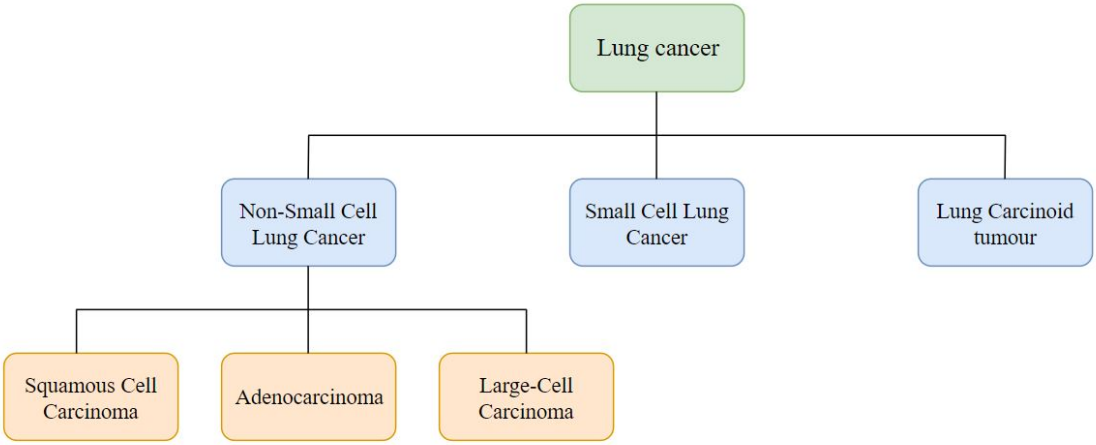


Figure 2.1: Scheme of different types and subtypes of lung cancer.

Of all lung cancers, 85% are NSCLC, with Adenocarcinoma type having the highest incidence. Small Cell Lung Cancer, also known as oat cells, is the lung cancer type that tends to spread more quickly, representing between 10% and 15% of all lung cancers. Holding less than 5% of the total number of lung cancer cases is the Lung Carcinoid tumour, which is known by a slow growth and a low tendency of spreading [7].

Since NSCLC is the most common type of lung cancer, a more detailed characterisation of its subtypes is presented below. Adenocarcinoma, which represents around 40% of lung cancers, normally starts in mucus-secreting cells, and it is found in the outer parts of the lung. Differently, from other types of cancer, Adenocarcinoma is more prone to occur in younger people and has a greater incidence in women than in men. The second most common type of NSCLC is Squamous Cell Carcinoma, representing around 25% of lung cancers. It tends to start in squamous cells, which are flat cells present inside of the airways in the lungs, and are found in the central part of the lungs, close to the main airway (bronchus). Large-Cell Carcinoma, which represents 10% to 15% of lung cancers, can develop in any part of the lung. It tends to grow and spread quickly, which supports the difficulty of treatment. A subtype of Large-Cell Carcinoma, known as Large-Cell Neuroendocrine Carcinoma, is fast-growing cancer that has similarities with Small Cell Lung Cancer [7].

2.2 Tumour Heterogeneity and Clonal Evolution

Lung cancer is a genetic disease that boosts due to the accumulation of multiple genetic mutations and epigenetic changes, which result in a cell proliferation out of control. As more abnormal cells develop, they can form into a tumour and spread through different body areas [8]. A significant genetic heterogeneity is present in tumours which accommodate various cell degrees of proliferation, differentiation, vascularity and inflammation, and/or invasiveness [9].

In Figure 2.2 is represented the evolutionary trajectory of tumour subclones, and their distribution (represented by different colours) within the tumour mass. Clonal mutations (truncal mutations on the evolutionary tree), represented by the blue cells in the tumour and blue trunk in the phylogenetic tree are present in all tumour regions and are the most recent common ancestor. Represented by red and yellow branches are subclonal mutations, which start to develop later in the tumour evolution course and exist only in a subset of regions or cells within the tumour. Hence, the likelihood of extracting the different subclonal mutations in a single biopsy is low.

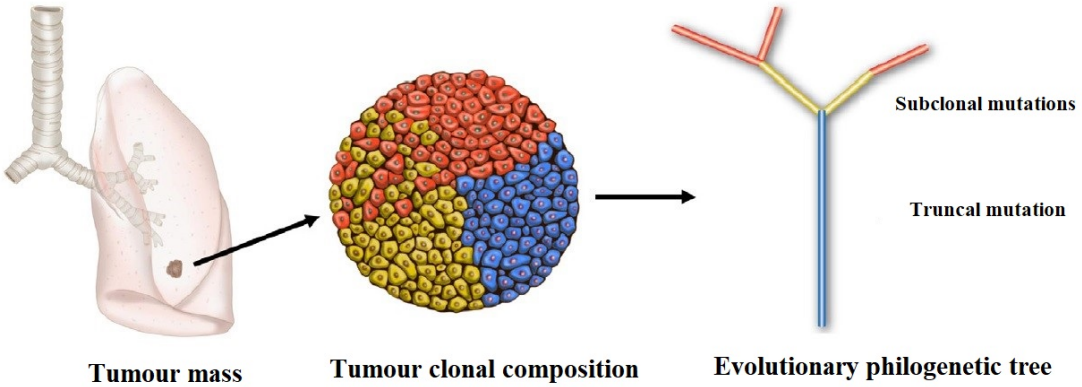


Figure 2.2: Evolutionary trajectory of a tumour through phylogenetic tree analysis. Adapted from Lee et al. [8]

The idea of genomically diverse minor subclones of malignant cells can explain phenomena like partial tumour responses, the seeding of metastatic cells or the development of drug-resistant malignant cells. Indeed, previous works have demonstrated how the presence of minor subclones can influence tumour progression [10–12]. This raises the relevance of extracting valuable genomic information about the entire tumour in order to clearly understand each patient's genomic stratification and individual situation.

2.3 Precision Medicine

Precision medicine tailors practices and/or therapies to individual patients by taking into account the individual variability of genes. Most of the research in this field is centred on molecular characterisation of tumours using genomic-based approaches, which demand tissue extraction by tumour biopsies [8, 13]. However, taking multiple tumour biopsies to properly determine the tumour clonal composition is not the most desirable solution due to its invasiveness. Moreover, taking into account the spatial intra-tumour heterogeneity and the tumour subclones stratification over time as a consequence of cancer treatment therapies, the potential sampling bias linked to a single tumour biopsy would limit oncology capacity to identify the best treatment option [8].

Medical image analysis can help to solve some clinical challenges related to biopsy, by providing tools for measuring functional, structural and physiological characteristics of a tumour, which are relevant for tumour diagnosis, clinical staging and treatment planning. Since image analysis can be subjective, precision medicine requires robust and standardised image features which can be used to predict molecular characteristics, prognosis and treatment response. These standardised features can be obtained by semantic annotations acquired from human observers or by a radiomic approach, which extracts the features directly from the image through a software [14]. The latter path will be deeply explained in Chapter 3.

2.4 Summary

The NSCLC, especially its subtype Adenocarcinoma, is the most incident type of lung cancer. It is normally found in the outer parts of the lung, and contrarily to other types of cancer, it is more prone to occur in younger people and has a greater incidence in women than in men. Generally, in a tumour, there is a heterogeneity of cells in different degrees of proliferation and/or differentiation, which form minor subclones of malignant cells. The presence of these minor subclones can influence tumour progression and, consequently, there is the need to search for the most beneficial treatment. Precision medicine is the field that aims to customise the treatment to individual patients, mostly centred on molecular characterisation of tumours using genomic-based approaches. Although tissue extraction from biopsies is the most regularly used approach today, medical imaging also has the potential to meet the needs of precision medicine, in a non-invasive way.

Chapter 3

Radiomics

Medical imaging is a commonly used and easily obtainable source of information in clinical oncology. Unlike biopsies, it is non-invasive and contains information with regard to the entire tumour [5]. Tumour response is currently assessed in clinical practice using RECIST (Response Evaluation Criteria In Solid Tumours) [15], which classifies changes in tumour diameter on CT images before and after therapy. However, the response to therapy can be slow [16], which decreases the efficiency of this procedure by giving a small margin for clinicians to shape the treatment. Therefore, the development of automatic and reproducible methodologies to obtain more information from medical images is a necessity.

Radiomics addresses this problem by extracting valuable quantitative features from medical images for the customisation of cancer care, which adds a new and hopeful dimension to precision medicine. Furthermore, it can also capture the tumour heterogeneity, which is considered an essential factor in tumour characterisation [6, 10–12].

The present Chapter is dedicated to describing this emerging and promising field. Section 3.1 mainly focus on the concept and the steps that comprise the workflow of radiomics, and afterwards, in Section 3.2 the most important types of radiomic features are described.

3.1 The Concept

Radiomics, first introduced by Lambin et al. [6] in 2012, hypothesises that medical images provide essential information regarding tumour physiology, which can lead to the improvement of cancer diagnosis and prognosis. Therefore, high-throughput data is extracted, and quantitative features are computed from medical images [6]. The aim of radiomics is to use predictive models, that can include clinical and molecular data, to help provide valuable diagnostic and prognostic information.

The workflow of radiomics (see Figure 3.1) can be divided into five steps: (I) image acquisition, (II) segmentation of the region of interest (ROI), (III) feature extraction and selection, (IV) model development and (V) model evaluation. The workflow is outlined as follows:

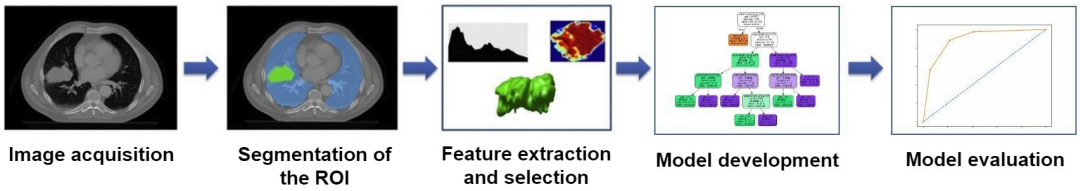


Figure 3.1: Workflow of radiomics. On the acquired medical images, a segmentation of the ROI is performed, followed by the extraction of features from the segmented region (e.g. features based on tumour intensity, shape, and texture). Later, these features are analysed in terms of their ability to predict valuable information for treatment planning. Adapted from Lambin et al. [6].

I. **Image acquisition:** Different imaging modalities suchlike CT, Positron Emission Tomography-Computed Tomography and Magnetic Resonance Imaging are managed in radiomics [17]. This work will focus on radiomics applied to CT images.

   Conceived in 1972 by Hounsfield, CT is the imaging modality of choice for analysing the thorax, since no other radiology method can achieve such anatomic detail. In fact, the only other imaging modality method that allows greater anatomic detail than CT is tissue examination, which is far more invasive than CT [18]. CT scanning is an imaging procedure in which a narrow beam of X-rays revolves around the patient, generating signals that are processed by the CT scanner's computer to produce cross-sectional images (slices) of the body (see the left image in Figure 3.2). Once a number of successive slices is obtained, they are digitally aggregated together to build a three-dimensional (3D) image of the patient (see the right image in Figure 3.2), which enables a clearer identification of basic structures such as tumours or abnormalities [19].

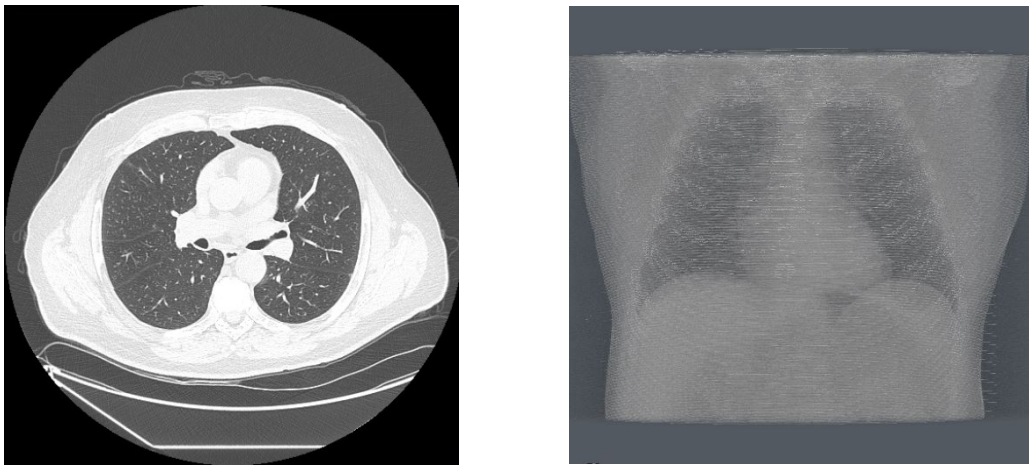


Figure 3.2: Example of CT slice (left) and 3D CT scan (right) of the chest, obtained with Aliza Medical Imaging DICOM Viewer [20]. Image from the NSCLC Radiogenomics dataset available at The Cancer Imaging Archive database [21, 22].

II. **Segmentation of the ROI:** The ROI segmentation is a central process within radiomics since the extracted features depend on the segmented volumes. The tumour's ROI is not necessarily restricted to the tumour's borders, which may, in some cases, be ambiguous. It can also include some area around the tumour that contains valuable information for tumour characterisation. Tumour segmentation can be accomplished with the resort to manual or automated methods, i.e. through manual delineation performed by an experienced radiologist or using an automated segmentation algorithm. As an alternative, a semi-automatic approach can also be used, encompassing not only automatic segmentation but also a posterior manual correction, if required [5]. In Figure 3.3 is presented a CT image where the tumour was segmented using a semi-automatic method and a representation of the segmentation in three dimensions.

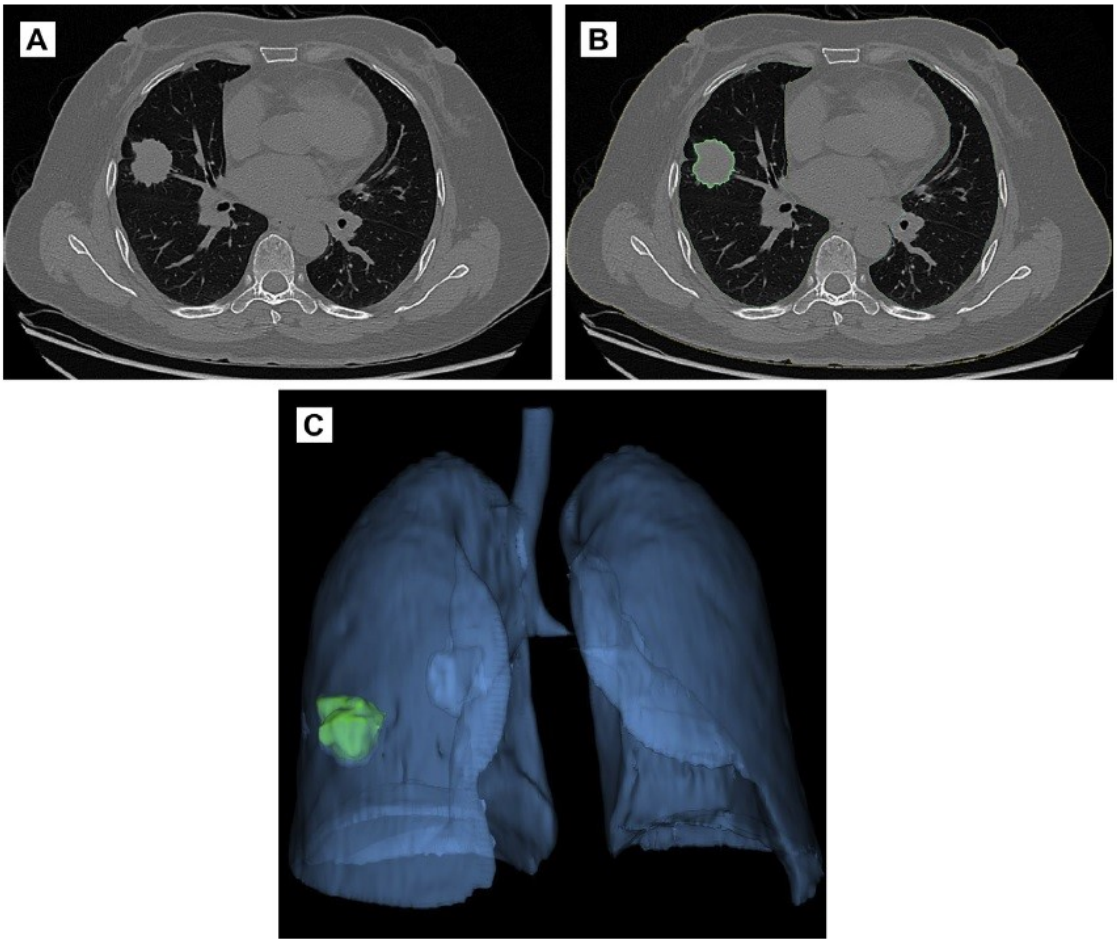


Figure 3.3: Tumour segmentation in CT image using a semi-automated algorithm. (A) CT slice of chest with a tumour in the right middle lobe. (B) Tumour outlined in green by a radiologist in a CT slice. (C) 3D view of the lung and the segmented tumour. Adapted from Liu et al. [23].

III. **Feature extraction and selection:** Quantitative features are extracted from the tumour region defined in the previous step. The different types of radiomic features will be further described in Section 3.2. Even though a vast number of features may be extracted, not all

of these features are necessarily useful for a specific task. In fact, the usage of all the features can lead to the model's power reduction and overfitting. Overfitting occurs whenever a model is extensively optimised for the training data set and performs poorly on new data. Therefore, a feature selection approach shall be applied in order to decrease the features space dimensionality and consequently enhance the model's generalisation ability and speed up the learning process. The feature selection step returns the most informative features on the basis of independence with other features and importance on the data [6]. The feature selection techniques can be roughly divided into two groups [24]:

- Filters: search for relations between the feature and the target attribute and ranks the features according to their relation;
- Wrappers: make an exhaustive search to look for the set of features which provides the highest predictive performance using a certain classifier.

IV. **Model development:** The selected radiomic features are used to create models able to predict valuable information for treatment planning, prognosis or diagnosis. The extracted features are used to build a model that can later be used to predict the output on new data. A machine learning model can either follow a supervised or unsupervised learning approach. In supervised learning, the training set includes a set of features paired with the corresponding label, and the job of the model is to learn the patterns between features and labels, in order to attribute the right label given a new set of features. Differently, the unsupervised learning approach comprises a set of unlabeled data and a learning algorithm, which learns the structure in the data [25].

V. **Model evaluation:** The model needs to be evaluated in terms of its generalisation ability, as far as possible, in an independent dataset. In case an independent dataset is not available, the available data must be split into a training set to train the model and a test set to evaluate its performance [5].

   In some cases, when it is useful to follow the model's performance during the training phase, for hyper-parameters tuning and overfitting control, some samples from the training set are held in a different set, called the validation set. These validation samples are not used to train the model, but for evaluating the model's performance in order to make adjustments in the training process (e.g. hyper-parameters tuning, early stopping). It is vital that the test set remains untouched until the final model is created, allowing its generalisation ability to be fairly evaluated.

   In classification tasks, where are included a great number of the radiomic studies, the model performance is evaluated with regard to the correspondence between the observed and predicted labels and its power to discriminate between classes. The success classification rate obtained by a model must be superior to the one obtained by classifying all objects in the class with the largest number of objects, also known as the majority class.

Diverse measures can be used to evaluate the predictive performance of a classification model. Even though most of these measures were developed for binary classification tasks, they can easily be adjusted for multiclass tasks. The main classification metrics are derived from a matrix, which shows for a set of objects how many were correctly classified and misclassified [24]. In Figure 3.4 is presented a confusion matrix followed by a representation which illustrates the meaning of each cell of the confusion matrix. The two columns represent the correct classes, whereas the two rows represent the predicted classes. The main diagonal holds the correct predictions, i.e. True Positives (TP) and True Negatives (TN), while the incorrect predictions are included in the secondary diagonal, i.e. False Positives (FP) and False Negatives (FN).

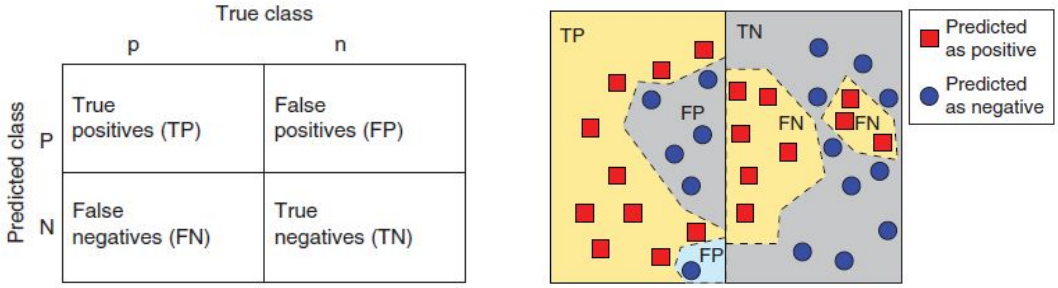


Figure 3.4: Confusion matrix for a binary classification task (left) and representation of TP, TN, FP and FN in a binary classification problem (right). Adapted from Moreira et al.[24].

The metrics that are normally used to evaluate classification models are a combination of these four simple measures - TP, TN, FP, FN. Commonly used metrics which derive from the confusion matrix are presented in Equations (3.1) to (3.5).

$$Recall = \frac{TP}{TP+FN} \tag{3.1}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3.2}$$

$$FalsePositiveRate = \frac{FP}{FP+TN} \tag{3.3}$$

$$Precision = \frac{TP}{TP+FP} \tag{3.4}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.5}$$

The recall, also known as True Positive Rate or sensitivity, returns the percentage of objects that are correctly classified as positive by the model, among all that were classified as positive. Specificity or True Negative Rate is the equivalent to recall, but regarding the negative class. Plus, the complement of specificity is the False Positive Rate. Differently, precision returns the percentage of objects classified as positive among all the objects that are truly positive. The balance between precision and recall can be achieved using the F-measure (Equation 3.6), which is a valuable metric especially in problems with class imbalance [24].

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.6}$$

However, for most classifiers, different threshold values to decide the class of a new object, result in different performances. Therefore, the area under the curve (AUC) is also a great estimator of the classifier's predictive performance, since it takes into consideration the performance of the classifier at different threshold values, by computing the area under the receiver operating characteristics (ROC) curve, which is the recall versus the False Positive Rate.

3.2 Radiomic Features

The fundamental part of radiomics is the high-throughput extraction of quantitative features that define the ROI. For a better understanding of the variety of radiomic features, they are divided into three major classes:

- Histogram features (First-order statistics);
- Morphological features (Shape-based features);
- Texture-based features (Higher-order statistics).

Radiomic features can remarkably vary in terms of complexity. At the most basic level, radiomic features aspire to reproduce features traditionally assessed visually, such as nodule size and shape. On the other hand, histogram-related and texture features are more complex forms of radiomics [26]. In this Section, it will be presented the representative features of the three major classes listed above, following the description made in [8, 17, 27].

3.2.1 Histogram Features (First-order statistics)

A tumour histogram is a plot of CT attenuation coefficients on the x axis versus the frequency of each attenuation coefficient in the y axis. Therefore, it provides information about the range of attenuation coefficients and the frequency of each attenuation coefficient within the defined ROI [28]. Median, mean, standard deviation, skewness, kurtosis, entropy, energy, variance and

uniformity are some of the features that can be obtained from a histogram. Although the histogram analysis enables the extraction of hidden information at the voxel level, which facilitates the discovery and evaluation of subtle changes in lung tumours, it is not able to retain spatial information about each voxel.

3.2.2 Morphological Features (Shape-based features)

Morphological features define a tumour with regard to shape and volume. Surface area and surface-to-volume ratio are examples of commonly used morphological features. Surface area is obtained by a procedure name triangulation, which generates a mesh of triangles that fully surrounds the tumour surface. Regarding the surface-to-volume ratio, a lower ratio represents a rounder and smoother tumour, while a larger ratio represents a sharp and irregular tumour. Figure 3.5 illustrates a 3D representation of two tumours with rather different surface-to-volume ratios.

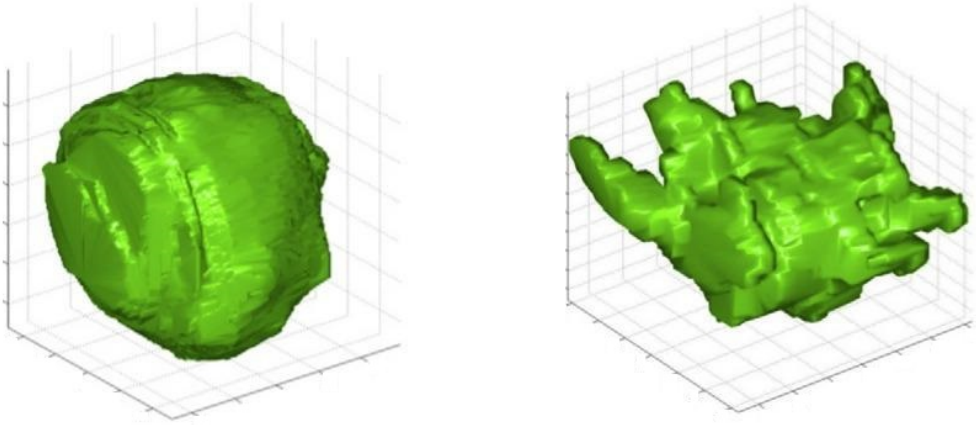


Figure 3.5: 3D representations of tumours with rather different surface-to-volume ratios. Left: round tumour with a low surface-to-volume ratio. Right: sharp tumour with a high surface-to-volume ratio. Adapted from Lambin et al. [6].

As for tumour volume, the number of voxels within the tumour's ROI is counted and the result is multiplied by the voxel volume.

3.2.3 Texture-based Features (Higher-order statistics)

Texture features, also known as higher-order statistics, take advantage of matrix features to classify the tumour in terms of texture. Four types of matrices are commonly used: grey level co-occurrence matrix (GLCM), grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM) and grey-level dependence matrix (GLDM).

Contrarily to histogram features, texture features hold spatial information between pixels, which enables to differentiate tumour subregions with different textural characteristics. Figure 3.6 presents two tumours with different textures.

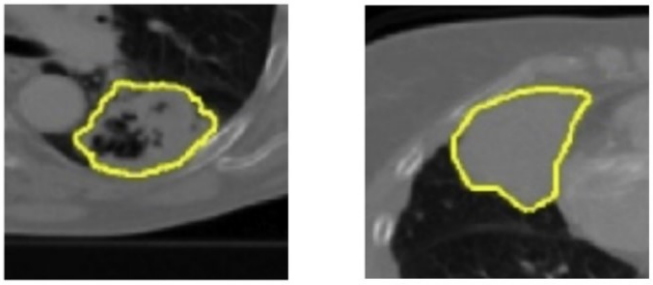


Figure 3.6: Texture differences between tumours obtained by CT imaging: more heterogeneous (left) and more homogeneous (right). Adapted from Lambin et al. [6].

3.3 Summary

The present chapter described how radiomics transforms radiographic medical images into a high dimensional feature space by quantifying tumour phenotypic characteristics. The workflow of radiomics comprises five steps: image acquisition, segmentation of the ROI, feature extraction and selection, model development and model evaluation.

Radiomics offers a vast group of quantitative features that can be extracted from the tumour's ROI. The available features vary in terms of complexity as well as in the type of tumour characteristics. The set of features can be divided into a number of families, including histogram-based, morphological and texture features. More complete information about each one of the features can be found in the Imaging Biomarker Standardization Initiative (IBSI) [27].

Chapter 4

Literature Review

Radiomics extracts a large number of medical imaging features that quantify the tumour's phenotypic characteristics. Multiple studies have shown the discriminating abilities of radiomic features for diverse clinical outcomes (e.g. overall survival (OS), disease-free survival (DFS) and the occurrence of metastasis) as well as to the tumour's histological subtype. Furthermore, some studies have analysed the association between radiomic features and the presence of mutations in specific genes. From hundreds of features extracted from the tumour's ROI in the CT image, a smaller number of features is selected according to their redundancy, relevancy, and stability, and afterwards, the group of features is evaluated for its ability to predict clinical factors.

In the type of studies mentioned above, machine learning is a valuable tool, which combines fundamental concepts of computer science with ideas from statistics, optimisation and probabilities. As defined by Mohri et al. [29], machine learning is a computational method that uses experience (data) to enhance the performance or to make accurate predictions. In a more practical sense, machine learning consists of designing efficient and reliable predictive algorithms, by learning from the data available. The success of the learning algorithm is dependent on the data used, which emphasises the importance of a robust feature engineering step.

A description that covers the most relevant radiomic studies that have been published so far is made in the present Chapter. For a clearer understanding, the Chapter is divided into four sections. Section 4.1 focus on studies with the aim of predicting genes mutation status, Section 4.2 presents the studies that have been published so far regarding the prediction of the tumour's histological subtype, and Section 4.3 briefly describes studies which aimed to predict prognostic factors. Finally, an analysis of the publicly available radiomics-related datasets is given in Section 4.4 followed by a discussion of the main conclusions of this Chapter in Section 4.5.

4.1 Genes Mutation Status Prediction

In the procedures of precision medicine, it is a standard practice to do mutational tests on selected genes to determine the response of patients to targeted therapy [30]. Most of the studies focus on the mutations of the EGFR, a cell receptor that activates growth and survival [31], and KRAS,

which activates the same pathway as EGFR when mutated [32]. The detection of Anaplastic Lymphoma Kinase (ALK) gene rearrangement is also important when patients have been diagnosed with NSCLC, since when rearranged ALK-positive tumours are extremely sensitive to therapy with ALK-targeted inhibitors, such as crizotinib [33].

Table 4.1 covers in a compiled mode a description, the main results and the most relevant features concerning studies that have been published regarding EGFR and KRAS mutation status prediction. Thereafter, each one of the studies presented in Table 4.1 is described in Section 4.1.1 regarding EGFR mutation status and in Section 4.1.2 respecting KRAS mutation status.

Table 4.1: Overview of published studies regarding genes mutation status prediction.

Publication	Description	Results	Relevant features
Liu et al. [23]	Evaluate the capability of radiomic and clinical features to predict EGFR mutation status in a cohort of patients with Adenocarcinoma.	Radiomic features combined with clinical features have a better performance predicting EGFR mutation status (AUC=0.709) compared to when used separately.	Radiomic features: main-direction, energy, wavelet features. Clinical features: gender, smoking status, histologic subtype, pathologic grade.
Mei et al. [34]	Evaluate the ability of radiomic and clinical features to predict EGFR mutation status in general and in exons 19 and 21, using a dataset with balanced classes.	Radiomic features combined with clinical features have a better performance predicting EGFR mutation status (AUC=0.664).	Radiomic features: GLSZM size zone non-uniformity normalised. Clinical features: gender, smoking status.
Ozkan et al. [35]	Investigate the correlation of 5 grey-level texture features with EGFR mutation status, in a small cohort of patients with Adenocarcinoma.	Contrast, correlation and difference momentum are correlated with EGFR mutation status.	Contrast, correlation and difference momentum.
Digumarthy et al. [36]	Prediction of EGFR mutation status using a set of clinical and radiomic features extracted from the CT slice in which the tumour has the maximal dimension.	Radiomic features combined with clinical features can help to properly distinguish from EGFR mutant and wildtype (AUC=0.863).	Radiomic feature: kurtosis. Clinical feature: smoking status.
Wang et al.[37]	Compare the capability of a Convolutional Neural Network to predict EGFR mutation status over models developed using semantic, clinical and radiomic features.	The Convolutional Neural Network was the model which achieved the best performance (Only Adenocarcinoma patients: AUC=0.81; other histological subtypes: AUC=0.77).	Not specified.
Li et al. [38]	Evaluate the capability of a radiomic approach using Random Forest and/or a Convolutional Neural Network approach to detect EGFR mutations.	The best performance was achieved with the model that combines clinical, radiomic and Convolutional Neural Network-based features (AUC=0.834).	Not specified
Gevaert et al. [39]	Prediction of EGFR mutation status using semantic features.	Decision Tree that predicts EGFR mutation status (AUC=0.89).	Emphysema, airway abnormality, percentage of ground glass component and tumour margin.
Gevaert et al. [39]	Prediction of KRAS mutation status using semantic features.	No relation between semantic features and KRAS mutation status (AUC=0.55).	No feature reached the significance threshold.
Weiss et al. [40]	Prediction of KRAS mutation status using texture features.	Decision Tree that differentiates KRAS mutation status (accuracy=89.6%).	Kurtosis, skewness, mean, standard deviation.

4.1.1 EGFR

In Adenocarcinomas, the most common type of NSCLC, EGFR mutations were discovered in 15-20% of the cases [41]. Patients with mutant EGFR have a higher response rate to tyrosine kinase inhibitors (TKIs), the first targeted drugs used for the treatment of NSCLC, than those with non-mutated EGFR [42]. Hence, patients with mutated EGFR lung cancer, who receive treatments with targeted TKIs, such as erlotinib, gefitinib, or afatinib are expected to have a longer progression-free survival in comparison to chemotherapy treatment [43]. However, if gefitinib is administered in cases with non-mutated EGFR, the patient will undergo a shorter progression-free survival [44]. On that basis, identifying patients with mutated EGFR is highly important in precision medicine.

Liu et al. [23] investigated the association between radiomic features extracted from CT images and EGFR mutation status in an Asian cohort of patients with Adenocarcinoma. From 219 features extracted, only 11 were found to have significant statistical differences between EGFR mutant and wildtype groups. These 11 features can be divided into 5 different feature classes: 2 features related to tumour location, 1 feature describing airspace, 2 histogram-based features, 2 co-occurrence and run-length features, and 3 texture-based features. The same analysis was performed concerning clinical features, which revealed that sex, smoking status, histologic subtype and pathological grade are related to EGFR mutation status. With multiple Logistic Regression (LR) analysis, radiomic features proved to increase the predictive performance of EGFR mutation status when combined with clinical features. In Figure 4.1 is represented the ROC curve for the three models analysed: (i) with clinical features alone; (ii) with radiomic features alone; (iii) both types of features combined. Clinical features had a better performance (AUC=0.667) than radiomic features (AUC=0.647) predicting EGFR mutation status, though, when these two groups of features are combined a significant improvement on the predictive power is achieved (AUC=0.709).

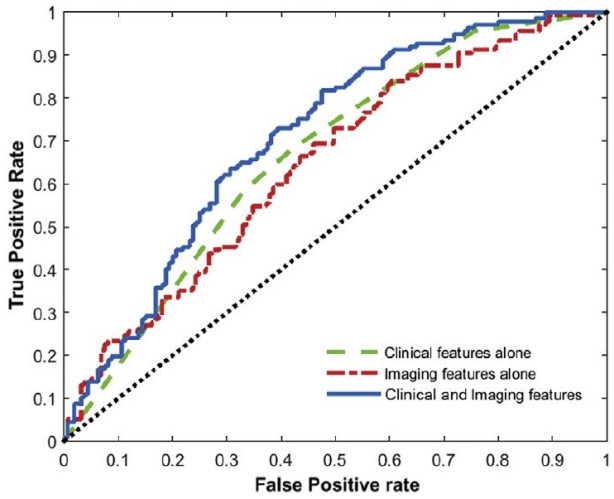


Figure 4.1: ROC curves obtained for the prediction of EGFR mutation status, using clinical features alone (green and dash line), radiomic features alone (red dash-dot line) and clinical and radiomic features combined (blue and continuous line). Adapted from Liu et al. [23].

Using a slightly larger sample size than Liu et al., Mei et al. [34] explored whether radiomic features extracted from pre-treatment CT images are biomarkers of EGFR mutation status in a cohort of 296 patients. Differently from the other studies, Mei et al. used a balanced dataset, which comprises 51% of EGFR mutated cases and 49% wildtype cases. Moreover, this study uses *Pyradiomics* [45], an open-source package, to extract radiomic features, rather than a software developed in-house. Hence, it simplifies the process of reproduction and comparison of results. Ninety-four radiomic features and three clinical features (age, gender and smoking status) were statistically analysed, and then all the features with statistical significance were used in an LR analysis. The performance was higher using the combination of radiomic and clinical features (AUC=0.664), rather than exclusively radiomic features (AUC=0.575). The logistic analysis identified that female sex, non-smoking status and a GLSZM radiomic feature named size-zone non-uniformity normalised are predictors of EGFR mutation status. Figure 4.2 presents the ROC curve obtained using only the GLSZM radiomic feature and the combination of radiomic and clinical features (i.e. gender and smoking status). A secondary mark in this study was to find radiomic and clinical predictors that are specifically correlated with mutations on exons 19 and 21 of EGFR.

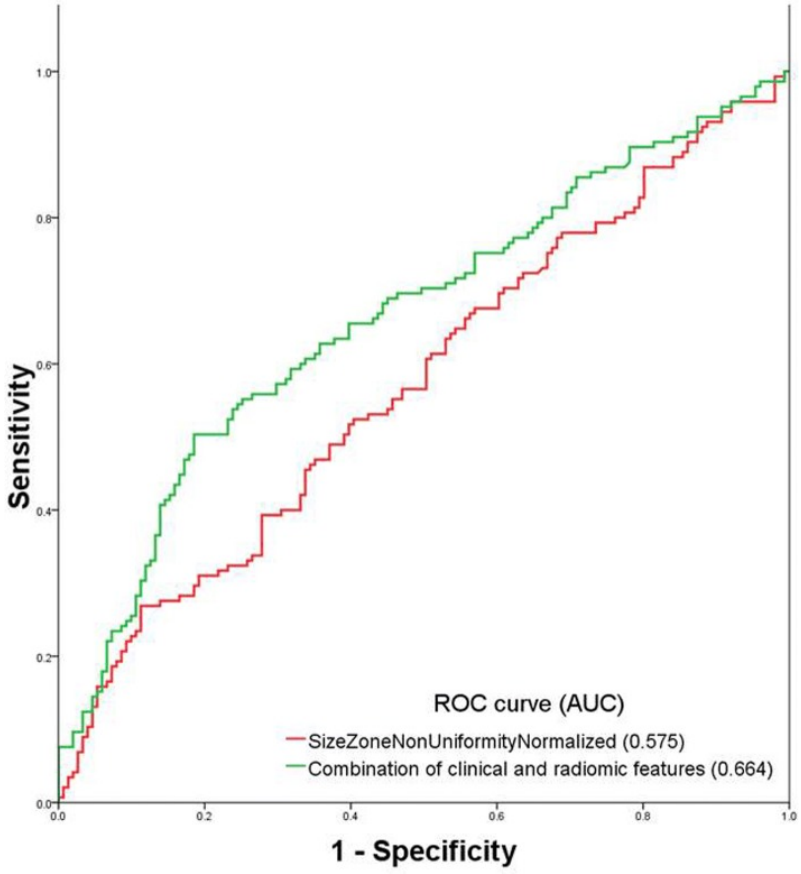


Figure 4.2: ROC curve of the model obtained using the GLSZM feature size zone non-uniformity normalised (red line) and the one obtained using the combination of radiomic and clinical features (gender and smoking status) (green line) to predict EGFR mutation status. Adapted from Mei et al. [34].

Ozkan et al. [35] used a rather small sample size to investigate the correlation between radiomic features and EGFR mutation status, having 46 patients with mutated EGFR and 20 patients with EGFR wildtype. In this study, only 5 grey-level texture features were analysed: angular second moment, contrast, correlation, entropy and inverse difference moment. Although they identified that contrast, correlation and difference moment are correlated with EGFR mutation status in Adenocarcinoma, the small sample size may introduce selection bias.

Having an identical purpose as the studies presented above, Digumarthy et al. [36] used only the slice in which the tumour had the maximal dimension to extract the radiomic features, rather the entire tumour volume, from a cohort of 69 patients with Adenocarcinoma. A statistical analysis showed that kurtosis and smoking status were significantly different between EGFR mutant and wildtype. However, the best performance was achieved using the total set of clinical and radiomic features in a multivariate LR analysis (AUC=0.863). Despite the satisfactory result achieved in this study, it would be important to validate the results with a larger sample size, in order to test the model generalisation ability in a more diverse group of patients, since the small sample of patients might have been affected by selection bias.

The previously described studies used concrete features that can reasonably be related to EGFR mutation status. However, there may exist abstract features which can also help to predict EGFR mutation status, yet they are complex to represent using hand-crafted feature engineering. Deep leaning plays an important role in these cases, by learning abstract associations between the input data and the desired label [46]. Wang et al. [37] used deep learning techniques, namely a Convolutional Neural Network (CNN) [46], to create a predictive model for EGFR mutation status, which shows improvements over models that use radiomic, semantic and clinical features. The data was obtained from 844 Adenocarcinoma patients (603 for training and 241 for validation), from two different hospitals. An ROI containing the entire tumour was manually selected and given as input to the network after resized to 64x64 pixels. A sequence of three consecutive slices was given as input to the network as a three-channel image, and the probability of being EGFR-mutated is obtained by averaging the probability of each set of 3 slices. An illustration of the CNN architecture is represented in Figure 4.3. The sub-network 1, represented in Figure 4.3, shares the same architecture of the first 20 layers of DenseNet [47] and it was trained using the weights of the model previously trained with 1.28 million natural images from the ImageNet dataset [48]. The sub-network 2 (see Figure 4.3) was trained using the 14926 CT images from the training set.

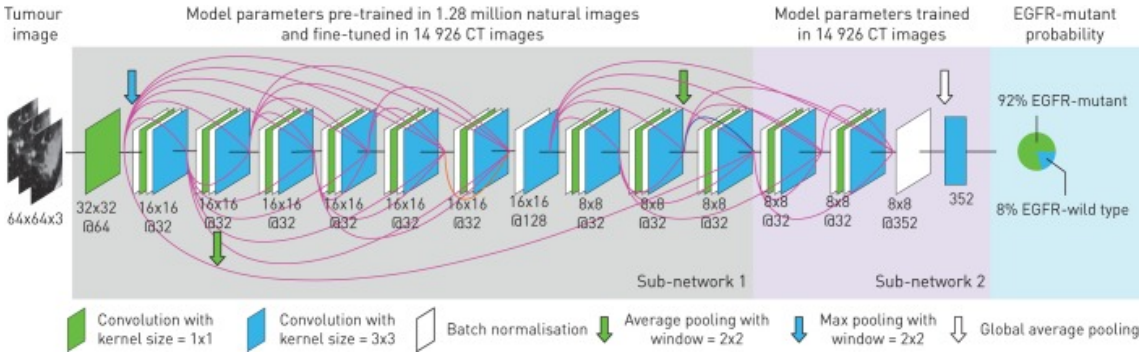


Figure 4.3: Architecture of the deep learning model. The sub-network 1 has the same structure as the first 20 layers of the DenseNet network, and it was pre-trained using 1.28 million images from Imagenet dataset. The sub-network 2 was trained with 14 926 CT slices. The output of the model is the probability of being EGFR-mutant. Adapted from Wang et al. [37].

In addition, it was developed a clinical model, a semantic model and a radiomics model. The clinical model used gender, age and tumour stage to predict EGFR mutation status, taking advantage of the Support Vector Machines (SVM) [49] predictive abilities and it achieved an AUC of 0.61. With superior performance, the semantic model achieved an AUC of 0.64 using 16 features in an LR model. Regarding the radiomics model, from 1108 extracted features a set of 8 features was selected using recursive feature elimination and later used in a Random Forest (RF) [50] with 100 trees, which achieved the same performance as the semantic model (AUC=0.64). The ROC curves respecting the developed models are represented in Figure 4.4.

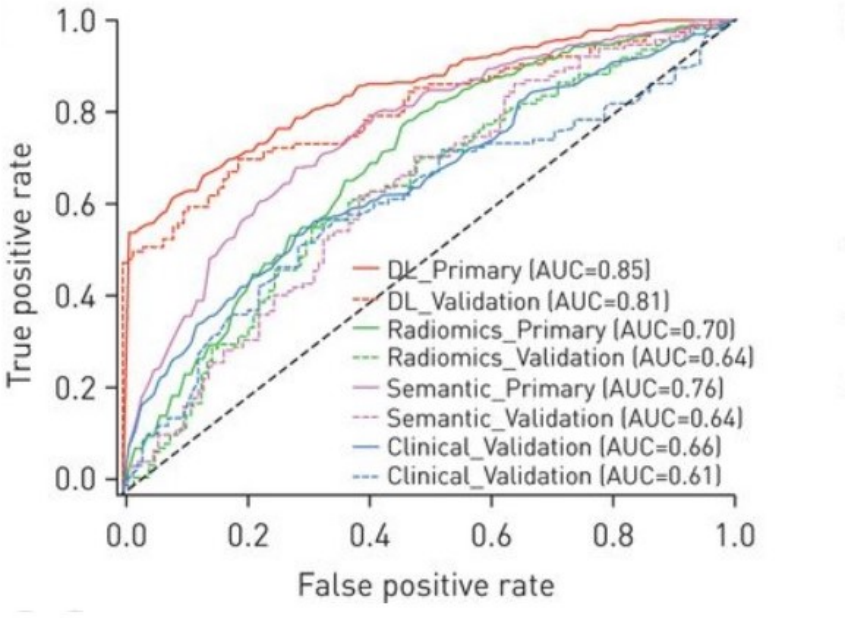


Figure 4.4: ROC curves of the deep learning model, radiomics model, semantic model and clinical model in the training (primary) and validation cohorts. (DL: deep learning). Adapted from Wang et al.[37].

As a secondary endpoint of this study, the same procedure was followed using patients from other histological subtypes besides Adenocarcinoma. This experience made the performance of the deep learning model decrease from an AUC of 0.81 to 0.77. This decrease may possibly be explained by the inclusion of data variability in the study, which makes it more difficult for the model to learn.

Li et al. [38] implemented a 3D CNN and a radiomic analysis using RF to detect EGFR mutations as well as to probe whether these two methods combined can more reliably identify mutated EGFR. In contrast to radiomic analysis, CNNs simply require minor pre-processing, and they uncover the relevant representations in a self-taught manner directly from the image. In this study, Li et al. used a cohort of 1010 patients with a balanced number of mutant and wildtype cases to build a radiomics-based model ($M_{radiomics}$) and a CNN-based model (M_{CNN}). These two models were combined, creating a third model which takes into account the prediction of each one of its sub-models. The weights that determine the contribution of each sub-model to the fused model are decided using a grid search approach. In addition, by making use of clinical data (smoking status and gender), a clinical features-based model was created using LR analysis and combined with the three previously mentioned models ($M_{radiomics}$, M_{CNN} and $M_{radiomics+CNN}$). The models were built using a training set of 810 patients and tested in a validation set of 200 patients. The architecture of the CNN-based model is presented in Figure 4.5.

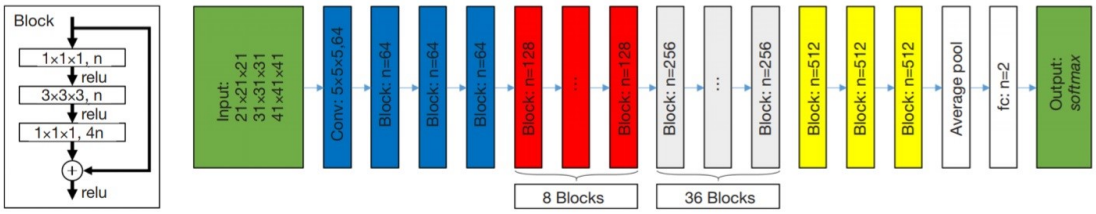


Figure 4.5: Architecture of the CNN-based model. The M_{CNN} comprised three residual CNNs with patches of 21x21x21 voxels (42 mm x 42 mm x 42 mm), 31x31x31 voxels (62 mm x 62 mm x 62 mm), 41x41x41 voxels (82 mm x 82 mm x 82 mm). Adapted from Li et al. [38].

The framework of this study is outlined in Figure 4.6.

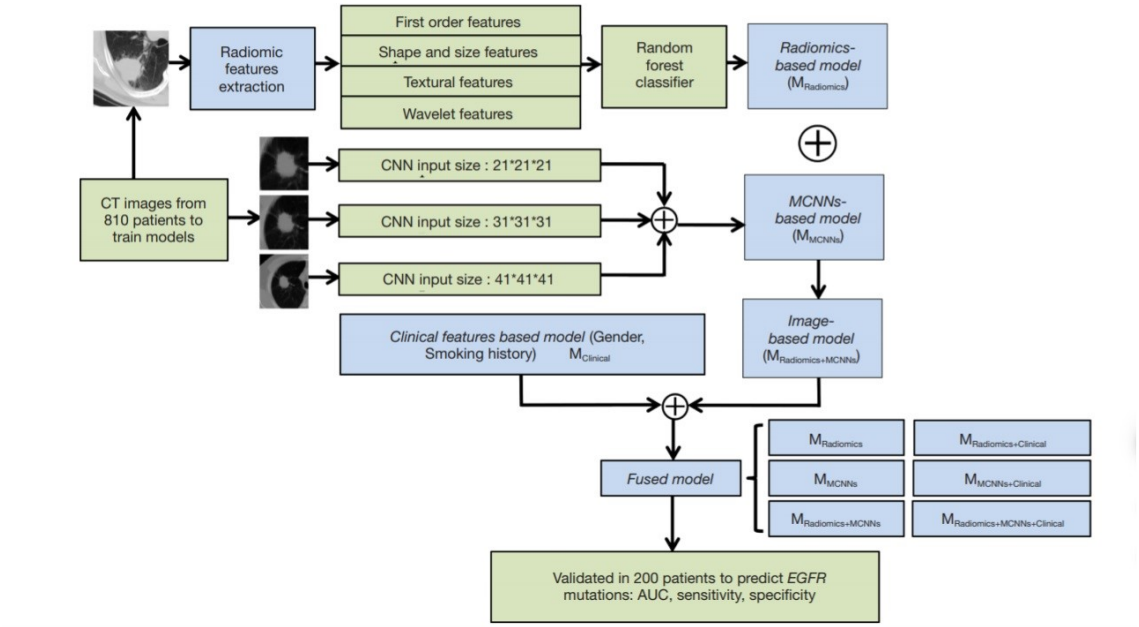


Figure 4.6: Scheme of the study's framework. $M_{radiomics}$ and M_{CNN} were developed extracting information from the CT images of the training set. The fusion of these two models generated $M_{radiomics+CNN}$. Clinical features were added to the previous models, creating new combinations that were then tested in the validation set. Adapted from Li et al. [38].

The ROC curves of all the developed models are presented in Figure 4.7. The AUC of $M_{radiomics}$, M_{CNN} and $M_{radiomics+CNN}$ were 0.740, 0.810 and 0.811. The M_{CNN} did not demonstrate significant inferiority when compared to $M_{radiomics+CNN}$. The highest AUC was achieved by $M_{radiomics+CNN+clinical}$ (AUC=0.834), although it is not significantly more efficient than M_{CNN} alone (AUC=0.810). From these results, one can conclude that radiomic features and CNNs can predict EGFR mutation status, though the addition of radiomic and clinical features is not significantly more efficient than using the CNN approach exclusively.

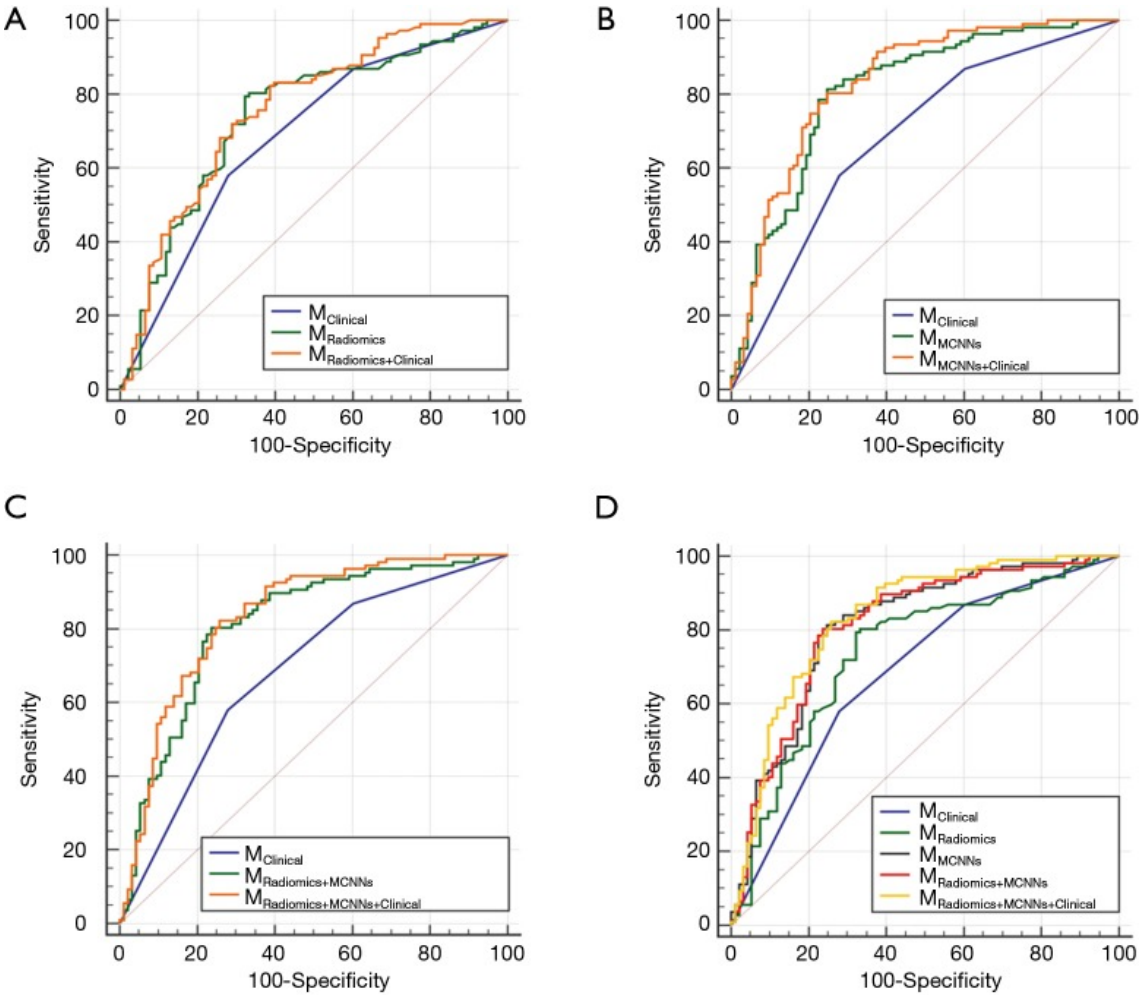


Figure 4.7: ROC curves of the developed models. (A) The AUC of $M_{clinical}$, $M_{radiomics}$ and $M_{radiomics+clinical}$ are 0.686, 0.740 and 0.758; (B) The AUC of M_{CNNs} and $M_{CNN+clinical}$ are 0.810 and 0.831; (C) The AUC of $M_{radiomics+CNNs}$ and $M_{radiomics+CNN+clinical}$ are 0.811 and 0.834; (D) Comparison between ROC curves of the models. Adapted from Li et al. [38].

Using semantic features, rather than radiomic features, Gevaert et al. [39] created a model to predict EGFR mutation status, which achieved an AUC of 0.89 in the test set. In this study, 89 semantic image features of NSCLC tumours were annotated by a thoracic oncologist. After a feature selection method has been applied, a Decision Tree [51] was built considering only four variables: emphysema, airway abnormality, the percentage of ground glass component and the type of tumour margin. In Figure 4.8, it is represented the obtained Decision Tree for predicting EGFR mutation status. According to the developed model, a wildtype is predicted in the presence of either emphysema or airway abnormalities, whereas the presence of a ground glass component implies a positive mutation status. The impact of clinical variables in the feature set was tested as a complementary endpoint in this study. Differently from the results achieved by Liu et al. [23], when clinical features were combined with the semantic features, the model performance decreased (AUC=0.82).

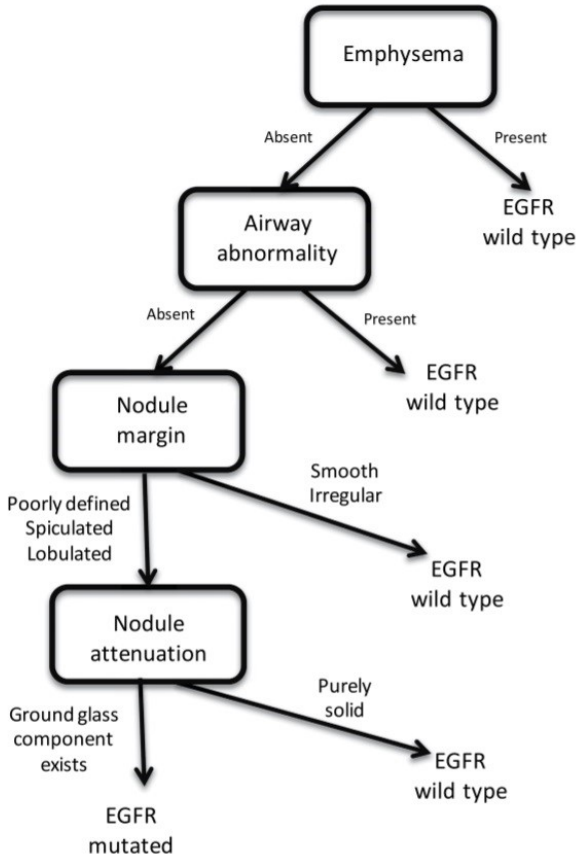


Figure 4.8: Decision Tree obtained for the prediction of EGFR mutation status using five semantic features. Adapted from Gevaert et al. [39].

4.1.2 KRAS

Information regarding KRAS mutation status is helpful for treatment planning [52]. It has proven to be correlated with response to chemotherapy since patients with mutated KRAS which undergo chemotherapy have revealed inferior responses and shorter survival compared to patients with no KRAS mutation [53, 54].

In the same study that was described in Section 4.1.1, Gevaert et al. [39] made an identical analysis of what was done regarding EGFR but this time concerning KRAS mutation status. A set of 89 qualitative features were annotated by a thoracic oncologist. However, no features reached the significance threshold of correlation with KRAS mutation and, consequently, the models built for KRAS were not considered useful (AUC=0.55). Two possible explanations for these results were identified. First, in the cohort used, KRAS mutations were less prevalent compared to EGFR mutations, 17% and 22% respectively. Second, there is the hypothesis that KRAS mutations do not have such a strong radiographic manifestation decoded by the semantic features used in this study compared to EGFR mutations.

In a former study, Weiss et al. [40] studied the differentiation of KRAS mutation status using a set of radiomic texture-based features extracted from a small sample set with 27 mutant cases

and 21 wildtype cases to build a Decision Tree. Kurtosis was the feature that most accurately differentiated KRAS mutated from wildtype tumours and consequently, it was used in the first tree node. Skewness and mean were assessed secondly, and the standard deviation was assessed at last. In case the misclassification rate in a branch was less than 15%, no further node was appended to that branch. The Decision Tree obtained, which is represented in Figure 4.9, achieved an accuracy of 89.6%.

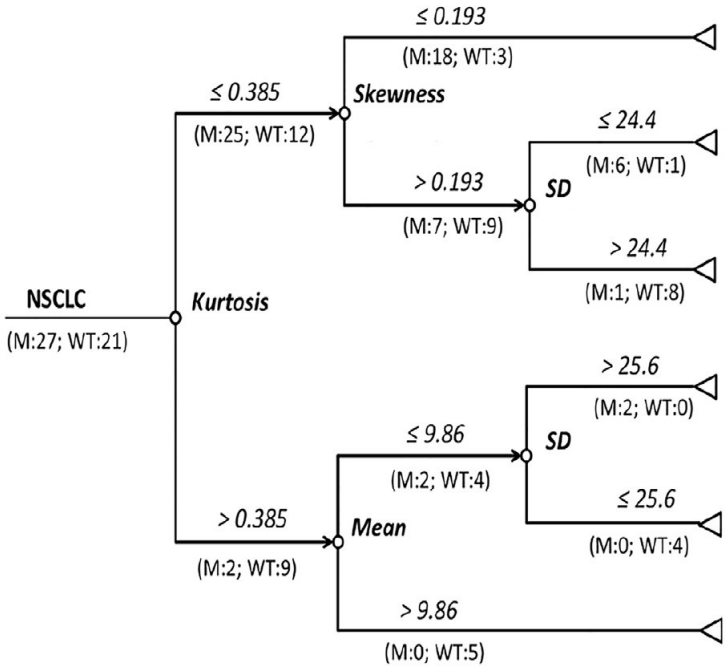


Figure 4.9: Decision Tree obtained for the prediction of KRAS mutation status. M: number of mutant cases; WT: number of wildtype cases; SD: standard deviation. Adapted from Weiss et al. [40].

4.2 Histological Subtype Prediciton

The radiomics literature includes studies which have focused their research on classifying a tumour as Adenocarcinoma or Squamous Cell Carcinoma, the two most common types of NSCLC. By identifying the subtype of NSCLC, the optimal treatment can more easily be defined [55]. A thorough search yielded a limited number of studies which investigated the association between radiomic features and the NSCLC histological subtype. Table 4.2 highlights important aspects of the studies which have been published regarding NSCLC histological subtype prediction.

Table 4.2: Overview of the literature regarding histological subtype prediction.

Publication	Description	Results	Relevant features
Parmar et al. [56]	Build radiomic signature to predict tumour histology.	11 clusters of features obtained by consensus clustering. LR model that uses medoids of the clusters to predict histological subtype (AUC=0.64).	Wavelet HHH statistical minimum, wavelet LHL GLRLM long run emphasis.
Wu et al.[57]	Compared 24 different feature selection and 3 classification methods for predicting histological sub-type.	The best result was achieved using Naive Bayes and ReliedFdistance (AUC=0.72).	Minimum, Wavelet: HLL GLRLM low run emphasis,HHL median, HLH cluster shade, HLL skewness.
Basu et al.[58]	Compared 2 different feature selection and 4 classification methods for predicting histological sub-type using 2D and 3D radiomic features.	The best result was achieved using 3D features in a Decision Tree without feature selection (accuracy=67%).	Not specified.
Digumarthy et al.[36]	Prediction of histological subtype using clinical features and radiomic features extracted from the slice in which the tumour has the largest axial dimension.	The best result was achieved using the total set of features (AUC= 0.92).	Clinical features: smoking status, gender. Radiomic features: longest diameter, entropy, standard deviation normalised.

The studies included in the literature overview in Table 4.2 are further described in detail.

Parmar et al. [56] applied a Consensus Clustering [59] approach, which is a resampling algorithm that estimates the number of clusters that best fit the data, with regard to reducing the feature space, using a cohort of 422 NSCLC patients. From a set of 440 radiomic features (including intensity, shape, texture and wavelet features), 11 clusters resulted from the consensus clustering approach. In Figure 4.10, it is represented the cluster consensus map, representing the segmentation of the features in 11 different radiomic feature clusters. An LR model was built using the 11 medoids, which achieved an AUC of 0.64, in a test set of 225 patients. In this case, a medoid is the feature in each cluster that has the highest pairwise correlation within the cluster. From the 11 cluster medoids, only two were associated with tumour histology: the wavelet HHH statistical minimum (AUC= 0.55) and wavelet LHL GLRLM long run emphasis (AUC=0.56).

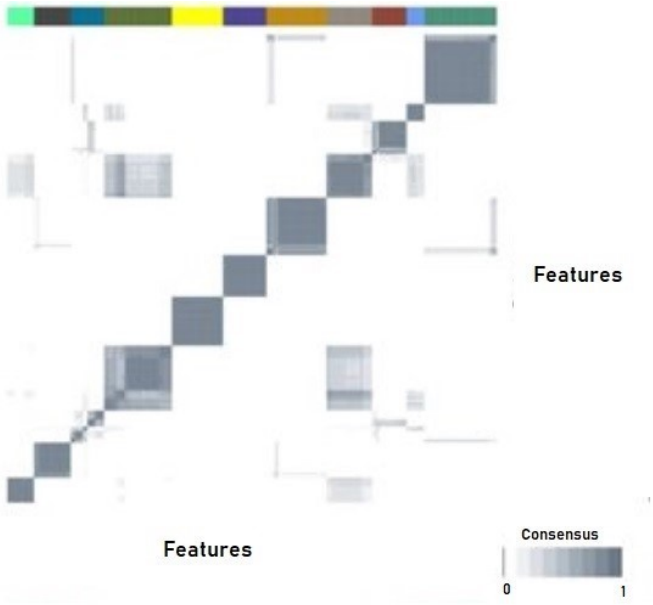


Figure 4.10: Cluster consensus map (11 clusters). Adapted from Parmar et al. [56].

With the purpose of predicting the histological subtype using radiomic features, Wu et al. [57] investigated 24 feature selection methods and 3 classification methods, using 440 features in a training cohort of 198 subjects and an independent test cohort of 152 subjects. A two-step feature selection procedure was addressed by combining the elimination of correlated features and then a univariate feature selection. For each feature selection method, the selection size varied from 5 to 45 (5,10,15,...,45) and, afterwards, the selected feature set was fit into the 3 investigated classifiers (RF, Naive Bayes (NB) [60] and K-Nearest Neighbours (K-NN) [61]). The best performance (AUC= 0.72) was obtained using ReliefF [62] as the feature selector and NB as the classifier. The radiomic features used in NB were the following: minimum, wavelet HLL GLRLM low run emphasis, wavelet HHL median, wavelet HLL skewness and wavelet GLCM cluster shade. It is noteworthy that ReliefF was the feature selector that ensured better performance for the three classifiers.

Basu et al. [58] conducted a similar approach of the one pursued by Wu et al. [57] in a small cohort of 74 subjects with NSCLC. However, this time both two-dimensional (2D) and 3D image features were hypothesised to be useful for classifying the tumours. The classifiers assessed in this study were Decision Tree, RF, Nearest Neighbour and SVM. ReliefF and Wrappers were used for feature selection. The model with the best performance achieved an accuracy of 67%, and it uses a Decision Tree without feature selection. From this study, it was concluded that in general, the results were superior using 3D features without feature selection. Certainly, one can not assume that feature selection is not required in radiomic studies since it depends on the size of the used feature set as well as on the features themselves.

Digumarthy et al. [36] attempted to predict the tumour histological subtype using only the image slice in which the tumour has the maximal dimension, instead of the entire tumour volume, in a cohort of patients with NSCLC. A group of 69 patients with Adenocarcinoma and 25 patients with Squamous Cell Carcinoma were used in this study. A Pearson correlation [24] analysis was performed to compare the relation between the features and the NSCLC subtype, and the most significant features were the smoking status, gender, longest diameter, entropy and normalised standard deviation. The best performance was achieved using the total set of features in a multivariate LR, with an AUC of 0.92. The fact that only one slice is used reduces the computational time significantly; however, plenty of information regarding the intra-tumour heterogeneity is being discarded when assuming that only one slice is able to predict the cancer subtype. To validate this result, it would be important to reproduce the work on a larger dataset.

4.3 Clinical Outcomes Prediction

In the field of radiomics, the development of prognostic predictive models has been an outstanding matter, with the main goal of correlating phenotypic features with clinical outcomes including, for instance, OS, distant metastasis and DFS. According to the National Cancer Institute [63], the OS is defined as the length of time between either the date of diagnosis or the start of treatment for the disease and the date that the patient is still alive. Also, in cancer terminology, DFS represents the length of time between the end of the primary treatment and the survival of the patient without any signs or symptoms of that cancer. The two measures defined above are commonly used to evaluate the performance of a new treatment. Table 4.3 points out important aspects of some published radiomic studies regarding clinical outcomes prediction.

Table 4.3: Overview of published studies regarding clinical outcomes prediction.

Publication	Description	Results	Relevant features
Weiss et al. [40]	Study the correlation between texture features, OS and DFS.	Kurtosis is a significant predictor of OS and DFS.	Kurtosis, skewness, mean and standard deviation.
Aerts et al. [64]	Build a prognostic radiomic signature.	Radiomic features selected on the basis of their stability and reproducibility have a strong prognostic power.	Kurtosis is a significant predictor of OS and DFS.
Parmar et al. [65]	Examine 14 feature selection methods and 12 classification methods, regarding their performance and stability to predict OS.	Wilcoxon test based feature selection method (AUC=0.65) and the classification method RF (AUC=0.66) were the methods with the prognostic performance.	Not specified.
Hawkins et al. [66]	Compared 4 different feature selection and 4 classification methods for predicting OS using a leave one-out approach.	The best result was achieved using a Decision Tree as classifier and Relief as feature selector (accuracy=77.5%).	Ratio free to attached, attachement to pleural wall, relative border to pleural wall, relative border to lung.
Coroller et al. [67]	Evaluate the ability of radiomic features to predict the occurrence of distant metastasis.	Radiomic features and clinical factors had a better performance compared to basic metrics, like volume and diameter.	Cluster prominence, standard deviation, entropy, overall stage, gender.

The studies previously mentioned in Table 4.3 are further described in detail.

Weiss et al. [40] demonstrated that texture features are linked to the patient's survival. In this study, kurtosis was identified as a significant predictor of OS and DFS, being a lower kurtosis value related to poorer survival. In addition, skewness, mean and standard deviation were proved to be associated with OS and DFS. The conclusions were taken based on a Kaplan-Meier survival analysis, considering the texture features, the patient's OS and DFS. Kaplan-Meier curves deal with time-to-event variables that have a beginning and an end anywhere along the timeline of the complete study [68].

Considering not only texture features, Aerts et al. [64] centred their research in building a radiomic signature with considerable prognostic power. Firstly, 440 radiomic features from four categories were extracted, describing (I) tumour intensity, (II) shape, (III) texture and (IV) multiscale wavelet. Then, the prognostic power of the most stable radiomic feature from each of the four feature groups defined was evaluated using a multivariate Cox Proportional-Hazards regression model [69]. This study showed that radiomic features selected on the basis of their stability and reproducibility were the ones with the strongest prognostic power. The resulting radiomic signature comprises (I) statistics energy, (II) shape compactness, (III) grey level non-uniformity, and (IV) grey level non-uniformity HLH.

As a recent area of research, the first prognostic-related radiomic studies that were published have focused on building a radiomic signature that is able to predict clinical outcomes, without stressing the comparison of different predictive modelling methods. In a more recent study, Parmar et al. [65] examined 14 feature selection methods and 12 classification methods with regard to their performance and stability for predicting the OS. In order to certify the unbiased evaluation of the different classification methods, the models used parameter configurations that were previously defined by Fernández-Delgado et al. [70] in a comparative study with 179 classifiers and 121 datasets. Parmar et al. concluded that the Wilcoxon test [71] feature selection method ensures the highest predictive performance with almost all the evaluated classifiers. Interestingly, the Wilcoxon test is a feature selection method that does not take into account the features redundancy when ranks the features. Additionally, RF was the classification method that achieved the highest performance (AUC=0.66) with the majority of the feature selection methods. One can argue that with a different parameter configuration, the performance of the classifiers would differ and the classification methods' ranking may not be the same.

Previously to this study, Hawkins et al. [66] compared 4 different feature selection and classification methods for prediction of OS in patients with NSCLC. The best result was achieved using a Decision Tree in a leave-one-out cross-validation (accuracy=77.5%) and was obtained after selecting 5 features per fold from 219 using Relief as feature selector. However, the results are not so reliable since the study only considered a cohort of 40 patients.

Coroller et al. [67] evaluated the ability of radiomic features to predict the occurrence of distant metastasis, as well as how their prognostic performance compares to the one obtained by basic metrics like volume and diameter. The prognostic performances were evaluated using the concordance index, which is a measure often used to evaluate how well a biomarker predicts the time

to an event [72]. The 15 features which had the highest concordance index were compared with tumour volume and diameter (basic metrics). Also, the same analysis was done regarding clinical features. As it is noted in Figure 4.11, the top 15 radiomic features had higher concordance indexes compared to basic metrics, along with the overall stage and gender that also demonstrated to be significant predictors of metastasis.

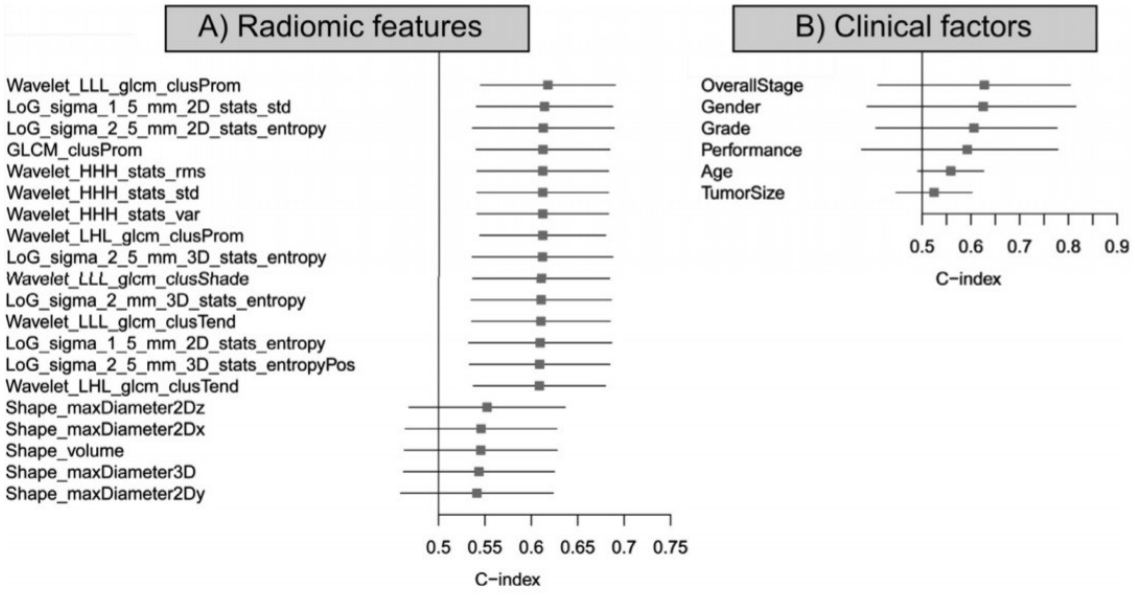


Figure 4.11: Forest plot of (A) the 15 best performing radiomic features for distant metastasis and the radiomic features equivalent to basic metrics (diameter and volume) and (B) the clinical features. C-index: concordance index. Adapted from Coroller et al. [67].

4.4 Public Datasets

A large number of radiomic studies that have been published used private datasets; however, there are also publicly available medical image data for radiomic experiments. The Cancer Imaging Archive [22] provides a collection of lung cancer-related datasets for public download, which comprises 22 datasets. While a significant number of datasets with medical images is available, only a few combine both medical images and clinical or molecular data. So far, there are only three complete and open access datasets for radiomics trials: (A) NSCLC-Radiogenomics [21], (B) NSCLC-Radiomics-Genomics [73] and (C) NSCLC- Radiomics [74]. Table 4.4 compares the central characteristics of the three datasets and, afterwards, each one of the mentioned datasets is briefly described.

Table 4.4: Comparison of some characteristics between the publicly available datasets: (A) NSCLC-Radiogenomics, (B) NSCLC-Radiomics-Genomics and (C) NSCLC- Radiomics.

Dataset	Modalities	Number of Subjects	Tumour Masks	Clinical Data	Mutation Status	Histological Subtype	OS
A	CT, PET/CT	211	✓	✓	✓	✓	✓
B	CT	89	×	✓	×	×	×
C	CT	422	✓	✓	×	✓	✓

NSCLC- Radiogenomics The NSCLC- Radiogenomics dataset comprises data collected between 2008 and 2012 from a cohort of 211 patients with NSCLC referred for surgical treatment. This dataset was developed with the aim of promoting either the study of the underlying relationship between tumour's molecular and medical imaging features and also to evaluate the prognostic power of medical image biomarkers. It includes CT and PET/CT images collected before surgical treatment, semantic annotations of the tumour using a controlled vocabulary, and 3D segmentation masks of the tumours. Also, medical images are paired with results of gene mutation status analysis, gene expression arrays and ribonucleic acid sequencing data, histological subtype results and survival outcomes. For the subjects which the dataset includes molecular information available, tissue samples from the excised tumours were analysed using ribonucleic acid sequencing technology, gene microarrays, or both, in addition to standard-of-care NSCLC mutational testing. Although the study includes 211 subjects, not all the data is available for all the subjects. Table 4.7 details the number of subjects available for each one of the data types mentioned above.

Table 4.5: Main collected data types and the corresponding number of subjects available in the NSCLC- Radiogenomics dataset.

Data Type	**Number of Subjects**
CT	211
Clinical Data	211
Tumour Masks	144
CT Semantic Annotations	190
PET/CT	201
EGFR Mutation Status	178
KRAS Mutation Status	170
ALK Mutation Status	158
Histological Subtype	207
Time to Death	61

It is important to mention that the data was acquired by different scanners using distinct scanning protocols depending on the institute and physician and, therefore, there was no attempt to standardise the acquisition and reconstruction protocols. More information about the NSCLC- Radiogenomics dataset can be obtained in the corresponding published article [14].

NSCLC- Radiomics- Genomics This collection includes images from 89 NSCLC patients that went through surgery. For these patients, the dataset provides pre-surgery CT scans, gene expression information, and clinical data, which includes the patient's genre, the tumour's histology subtype and location. The NSCLC- Radiomics- Genomics dataset is the Lung3 dataset referred in the study of Aerts et al. [64], described in the previous Section, used to build a prognostic radiomic signature.

Table 4.6: Main collected data types and the corresponding number of subjects available in the NSCLC-Radiomics-Genomics dataset.

Data Type	Number of Subjects
CT	89
Clinical Data	89
Histological Subtype	89

NSCLC- Radiomics The NSCLC- Radiomics collection includes CT scans, manual delineations of the 3D volume of the gross tumour performed by a radiation oncologist, and clinical data from a cohort of 422 subjects with NSCLC. This dataset is the one cited as Lung1 dataset in the study of Aerts et al. [64], described in the previous Section, used to build a radiomic prognostic signature. The CT images are paired with clinical data that provides the patient's age and gender, the tumour's stage and histological subtype, the survival time (in days from the start of treatment) and the death status.

Table 4.7: Main collected data types and the corresponding number of subjects available in the NSCLC- Radiomics dataset.

Data Type	Number of Subjects
CT	422
Clinical Data	422
Tumour Masks	318
Histological Subtype	203
Time to Death	422

4.5 Discussion

Taking into account the studies presented in this Chapter, one can conclude that although there is a significant number of works in the field of radiomics, there is still a margin for considerable progress. An important extension to some of the studies that were reviewed in this Chapter would be to perform hyper-parameter tuning to ensure the best performance of each of the used classifiers. In this way, it would be guaranteed that each classifier was configured with the set of hyper-parameters which maximises its performance.

One of the possible reasons for not existing a larger number of studies in this area can be the reduced number of public datasets that provide relevant data for radiomics-related studies.

The usage of different small cohorts of patients influences the different results and conclusions reached by the studies, serving as a barrier so that the results can effectively be put into practice in a clinical setting. Therefore, in the future, it would be important to create a public dataset with a significant number of subjects that would stimulate the research in this field.

4.6 Summary

Radiomics studies do not have only one problem to solve since one can use radiomics abilities to attempt to predict different valuable factors for clinical oncology. The existing studies can be divided into three categories according to the prediction goal: genes mutation status, histological subtype and clinical outcomes. In this chapter, studies published in these three categories were detailed and compared. Different methodologies were followed in the studies described; however, there are common conclusions one can take from the description presented in this Chapter:

- Radiomic features have the potential to make accurate predictions; however, when combined with clinical features, the performance increases;
- There is the hypothesis that KRAS mutation status does not express through medical images in the same extent as EGFR mutation status. This hypothesis is in accordance with the lower results in general as well as with the existence of fewer studies for KRAS mutation status prediction, compared to EGFR mutation status;
- Semantic features have shown potential to predict EGFR mutation status;
- The usage of deep learning techniques (i.e. CNNs) to predict EGFR mutation status outperforms the results obtained when radiomic features were used. Nevertheless, the usage of deep learning techniques requires a large amount of data, which to date does not exist publicly available.

Conclusively, even though radiomics is a recent field, some useful studies have been done regarding the analysis of the features extracted and the development of useful predictive models related to NSCLC tumours. However, there are still advances which can be accomplished such as improving the model's performance using different techniques and validating the previously obtained results using different datasets.

Chapter 5

Data Preparation

The present Chapter has as main focus the description of the different types of data this study uses. Furthermore, the processing steps required before building the predictive models are addressed, starting from the creation of the 3D binary masks until the extraction of the radiomic features from the tumour. Therefore, the Chapter is divided into four sections: (I) description of the available data (Section 5.1, (II) development of 3D binary tumour masks (Section 5.2), (III) CT images pre-processing (Section 5.3) and (IV) extraction of radiomic features (Section 5.4).

The CT images used in this study are stored in Digital Imaging and Communications in Medicine (DICOM) format, which is a standard for storing and transferring medical images in an electronic format, following a structured protocol. Its main purpose is to standardise the format of diagnostic images, following a set of rules that enable medical images and related information to be exchanged between equipment of different brands, either in hospitals, laboratories or clinics [75].

5.1 Data Description

This Section inspects the available data which meets the requirements for the proposed objectives: (I) to predict EGFR mutation status, (II) to predict KRAS mutation status and (III) to predict the tumour histological subtype (Adenocarcinoma or Squamous Cell Carcinoma). The development of ALK mutation status predictive models was not conducted in this study due to the small number of mutated cases available (wildtype: 108, mutant: 2). From the datasets described in Section 4.4 of Chapter 4, only subjects who meet the following terms will be used:

- Availability of CT image;
- Availability of tumour segmentation;
- Availability of target attribute (EGFR mutation status or KRAS mutation status or tumour histological subtype);
- Availability of semantic features.

Only the NSCLC-Radiogenomics [14, 21] and NSCLC-Radiomics [74] datasets comprise data that meets the requirements described above. Therefore, these two datasets and the respective data are described in detail below.

5.1.1 NSCLC- Radiogenomics Dataset

The NSCLC- Radiogenomics dataset is the only public dataset which comprises information regarding the mutation status of lung cancer-related genes (EGFR, KRAS and ALK). Despite including a cohort of 211 NSCLC subjects, only 116 were further considered in the presented radiomic study for EGFR mutation status prediction; 114 for KRAS mutation status prediction and 141 for tumour's histological subtype prediction. The scarce availability of tumour masks and target labels did not allow all subjects to be used. Since the NSCLC-Radiogenomics is a retrospectively collected dataset, the scanning protocol and scanning parameters were not standardised, thus slice thickness varied from 0.625 mm to 3 mm (median:1.5 mm), the X-ray current from 124 mA to 699 mA (mean: 220 mA) at 80-140 kVp (mean: 120 kVp). The subjects were in supine position with the arms at sides, while the scans were acquired from the top of the lung to the adrenal gland during a single breath [14].

The segmentation masks are stored as DICOM Segmentation Objects [76] and are represented as 3D binary images, where voxels belonging to the tumour ROI are represented by the value 1 and voxels outside the tumour ROI are represented by 0. In the cases where the segmentation mask image did not have the same dimensions as the corresponding CT image, the appropriate number of slices was added to the segmentation mask. In Figure 5.1 is presented a 2D and 3D representation of the overlap between the CT image and the tumour segmentation mask.

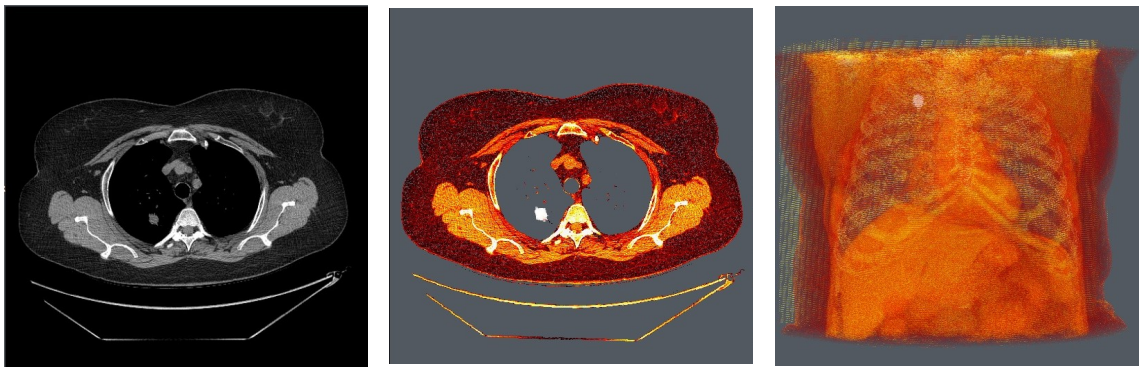


Figure 5.1: Representations of the tumour segmentation mask and corresponding CT image obtained with Aliza Medical Imaging DICOM Viewer [20]. Left: 2D CT slice with a visible tumour. Centre: 2D CT slice and corresponding tumour segmentation mask. Right: 3D visualisation of the overlap between tumour segmentation mask and CT image.

5.1.1.1 Semantic Features

The NSCLC-Radiogenomics dataset comprises a set of subjects whose tumour was analysed from the pre-treatment CT images using a controlled vocabulary. From these subjects, 158 are charac-

terised in terms of EGFR mutation status, 157 subjects characterised in terms of KRAS mutation status, and 187 characterised in terms of NSCLC histological subtype. The semantic terms are commonly used in radiology clinical practice and derive from descriptions in the radiology literature [14]. Definitions of some of the terms used in this description can be found in [77]. The template of semantic terms was developed by two academic thoracic radiologists exclusively for nodules since it is the most prevalent expression of lung cancer. Therefore, other manifestations of lung cancer besides nodules (e.g. central obstructive tumours) are not included in this analysis. The dataset includes 30 nodule and parenchymal features, which describe nodule's geometry, location, internal features and other related findings. However, the presence of some features is conditioned upon the presence/absence of other features. For instance, when emphysema is not present in the lung, the primary emphysema pattern is not collected. More information about the semantic annotations protocol can be found at [14].

5.1.2 NSCLC- Radiomics Dataset

The NSCLC- Radiomics dataset along with the NSCLC- Radiogenomics dataset will be used for conducting the experiments regarding the tumour's histological subtype prediction.

Contrarily to the NSCLC-Radiogenomics dataset, which comprises binary segmentation masks, the NSCLC-Radiomics dataset contains manual delineations of the tumour in each slice. These contours are commonly used by medical experts to point out a certain part of the slice by drawing a continuous line around the ROI and they are stored in DICOM-RT Structure Set format [76]. The radiotherapy structure set (RTSTRUCT) is one of the modules of the DICOM standard and it is mainly used to link patient structures and related data detailed on CT scanners [78]. To the best of found knowledge, no publication with information regarding the data acquisition protocol is available; however, the acquisition details of each CT image are embedded in DICOM metadata. Information regarding the dataset population and treatment is available at [64] (Supplementary Information).

All things considered, a total of 120 subjects in the present dataset cover the requirements to be used in the tumour's histological subtype prediction experiments, which includes being either Adenocarcinoma or Squamous cell carcinoma, having tumour contours and correspondent CT image.

5.2 Conversion from RTSTRUCT Contours to 3D Binary Mask

To extract the tumour ROI from the CT image it is essential to have a 3D binary mask which indicates the voxels that belong to the tumour, within the CT image. Since there is not a precise boundary that divides the tumour voxels from other tissues, the delineation is normally performed in a coarse manner.

In order to use the data available on the NSCLC-Radiomics dataset, it is necessary to convert the contours into a 3D binary mask. Based on the functions included in the *dicom-contour* [1] package, the procedure has the following outline:

Return the desired set of contours: A contour file can have multiple contours, which means that they may include contours drawn by different experts or contours from different structures (e.g. different tumours). Since these contours are manually drawn, they are prone to human error and in some cases more than one possible draw is available. Still, in the NSCLC-Radiomics dataset, each slice had only one contour available.

Find the combination between contour and CT slice: A crucial step in this procedure is to find the CT slice which agrees with each available contour. This association is possible due to the information available in the tags from the metadata dictionary, namely the Referenced SOP Instance UID (0008,1155) [2] available in the contour metadata which matches the SOP Instance attribute (0008,0018) [3] available on the metadata of the correspondent CT slice. In Figure 5.2 is presented an example from the NSCLC-Radiomics dataset that evidences the agreement between a contour draw and a CT slice.

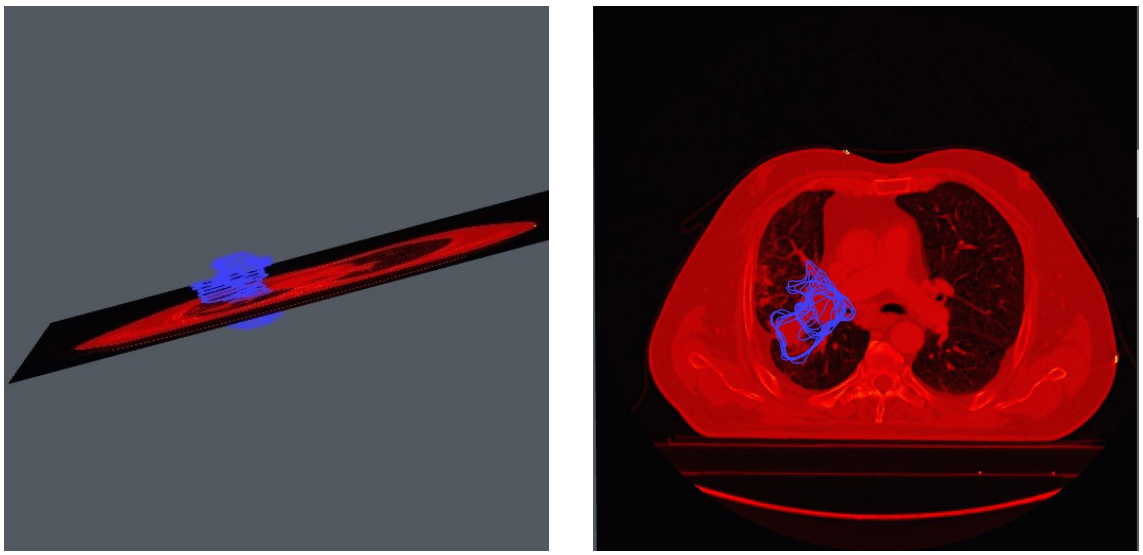


Figure 5.2: Overlap between a contour instance and the respective CT slice, obtained with Aliza Medical Imaging DICOM Viewer [20]. Left: Visualisation of the overlap from the side. Right: Visualisation of the overlap from the top.

Mapping from Cartesian coordinates to pixel positions: In computer vision, a continuous line is a collection of pixels, yet the software used to create the contours exports the pixels as (x, y, z) (height, width, depth) in millimetres [79]. Having this in mind, there is the urge to map from Cartesian coordinates to pixel positions. The spacing between two adjacent

[1] Available at: https://github.com/KeremTurgutlu/dicom-contour
[2] Keyword: ReferencedSOPInstanceUID
[3] Keyword: SOPInstanceUID.

pixels $(x_{spacing}, y_{spacing})$ (Pixel Spacing Attribute [4]) and the image origin which is the centre upper left point (x_{origin}, y_{origin}) (Image Position (Patient) Attribute [5]) are obtained for each CT slice associated with a contour, and the contour coordinates in millimetres (x_{mm}, y_{mm}) are converted to pixel positions (x, y) using the following Equation:

$$(x, y) = \left(\frac{x_{mm} - x_{origin}}{x_{spacing}}, \frac{y_{mm} - y_{origin}}{y_{spacing}} \right) \tag{5.1}$$

Figure 5.3 presents a representation which illustrates the meaning of origin and spacing.

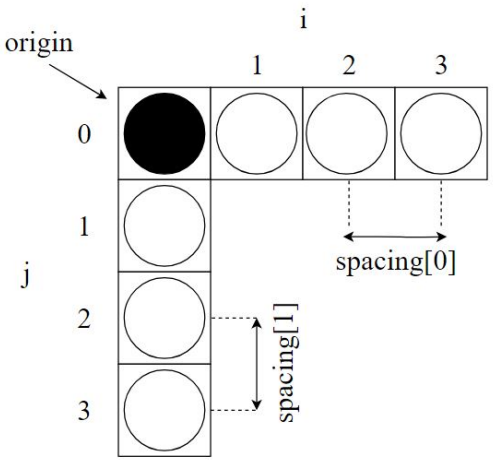


Figure 5.3: Representation which illustrates the meaning of origin and spacing.

Create 2D binary mask: Using the coordinates in units of pixels and the shape of the CT slice, a polygon is drawn and filled in the inside, thus creating a 2D binary mask with the same size of the CT slice.

Create mask voxel: The next step before joining the 2D masks is to order the slices correctly according to the Z coordinate available in the metadata. Once the slices are ordered, it is checked whether the slice is linked with a contour, and in case that is not verified a slice filled with zeros is added to the structure. Otherwise, the correspondent polygon mask is added. In the end, the algorithm returns a 3D binary mask with the same dimensions as the CT image. It is noteworthy that slices with contours are consecutive, which means that no type of interpolation was required.

5.3 CT Images Pre-Processing

Some properties of the images, such as the distance between slices, may differ from scan to scan, which consequently affects the learning ability of the algorithms. Therefore, before trying to

[4] For more information: dicom.innolitics.com/ciods/ct-image/image-plane/00280030
[5] For more information: dicom.innolitics.com/ciods/ct-image/image-plane/00200032

extract patterns from the images, the algorithms demand image pre-processing.

5.3.1 Hounsfield Units

The unit of measurement of a CT scan is the Hounsfield unit (HU), which is a measure of radiodensity. The HU scale considers the radiodensity of water, at standard temperature and pressure, is defined as 0 HU, and the radiodensity of air is defined as -1000 HU. The equation for computing the HU values based on radiodensity is shown in Equation 5.2, where μ represents the original linear attenuation coefficient of substance, μ_{water} represents the linear attenuation coefficient of water and μ_{air} the linear attenuation of air [80].

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{5.2}$$

By default, the values returned by the CT scanner are not in this unit. In such manner, the radiodensity values were converted to HU units, by multiplying the voxel value by the slope and adding the intercept related to the linear transformation, which values are stored in the metadata of the scans. By performing this transformation, it is ensured that from this point forward the values are in HU units.

In a typical transmission-based imaging technique, the emitted X-rays pass through the tissues, which affects the level of energy that reaches the detector, depending upon the type of tissue it meets along the path. The denser the tissue, the brighter the image will be. Therefore, tissues having low Hounsfield value will darken the film (e.g. lungs and fat), whereas tissues with high Hounsfield values provide white spots in the image, such as hard bones [80]. From Figure 5.4, where a histogram of voxel values and the respective CT slice is presented, one can clearly see which pixels are air and which are tissue.

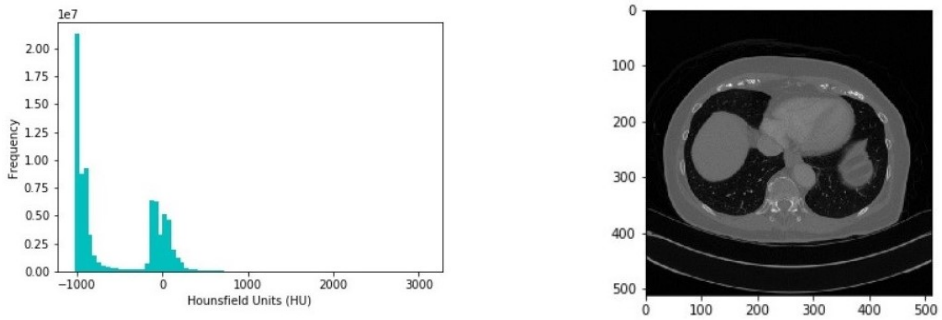


Figure 5.4: Histogram of voxel HU values (left) and respective CT slice (right).

5.3.2 Resampling

With the purpose of learning patterns from the data using an automatic analysis methodology, it is extremely important that a pixel is represented in the same way in the entire dataset. For instance, a scan may have a pixel spacing of [1.5, 0.25, 0.25], which means that neighbour slices are separated

by 1.5 mm and adjacent pixels by 0.25 mm in x and y directions; whereas, in a different scan the spacing may be [2.50, 0.10, 0.10]. A common practice to deal with this circumstance is to resample the full dataset to a defined isotropic resolution. Having this in mind, the entire dataset (including the tumour masks) were resampled to [1.00, 1.00, 1.00].

For a clearer understanding of the resampling process, an example will be given which considers that the original image has a pixel spacing of [1.25, 0.89, 0.89] and a shape of [269, 512, 512] and the new spacing is intended to be [1.00, 1.00, 1.00]. The resize factor is obtained by dividing the original spacing by the intended spacing $[1.25/1.00, 0.89/1.00, 0.89/1.00] = [1.25, 0.89, 0.89]$. Next, it is computed the shape of the image after the resampling, which is given by the round value of $imageshape \times resizefactor = [304 \times 1.25, 355 \times 0.89, 355 \times 0.89] = [336, 460, 460]$. The final resize factor might be slightly different from the one computed, due to the rounding performed. Thus, the real resize factor is computed by dividing the shape of the new image by the shape of the original image: $[336/269, 460/512, 460/512] = [1.25, 0.89, 0.89]$. By dividing the original spacing by the real resize factor, one can validate the final scan spacing: [1.25/1.25, 0.89/0.89, 0.89/0.89]=[1.00, 1.00, 1.00]. It is important to note that in this example only two decimal cases were considered; however, in this study, 8 decimal digits were held, which means that the final spacing may not always be [1.00, 1.00, 1.00]. Once the real resize factor was computed, an interpolation was performed where new points were created based on its nearest data points.

5.3.3 Normalisation

Images were normalised between -1000 HU and 400 HU, since -1000 HU is the radiodensity of air and values above 400 HU represent hard tissues not relevant for the task at hand [81]. Values under -1000 HU or above 400 HU were defined as -1000 HU and 400 HU, respectively.

Afterwards, the data was zero-centred by subtracting the mean Hounsfield value of the whole dataset from each voxel. The mean Hounsfield value obtained was -621 HU.

5.4 Extraction of Radiomic Features

From the 3D tumour of the pre-processed CT images, a set of 1218 radiomic features were extracted using the open-source package *Pyradiomics* implemented in *Python* [45]. The extraction of radiomic data using *Pyradiomics* comprises four main steps: (I) loading of images and segmentation masks; (II) application of filters; (III) computation of features using the different feature classes and (IV) returning extracted features. It is important to note that *Pyradiomics* does not support DICOM files as input, hence both CT images and segmentation masks were converted to Nrrd ("nearly raw raster data") format.

Once the files are loaded into *Pyradiomics*, the conformity between the pair image and binary mask is verified, and an error message is returned in case the mask includes areas outside of the image physical-space.

Features were computed both on the original image and on images obtained after application of wavelet and Laplacian of Gaussian (LoG) filters. A wavelet transform decouples textural information by decomposing the original image in low and high frequencies. A 3D undecimated wavelet transform was applied to each CT image, which decomposed the original image into 8 different images. Considering that L is a low-pass filter and H a high-pass filter, the original image X is decomposed into 8 new images after the wavelet decomposition: X_{LLL}, X_{LLH}, X_{LHH}, X_{LHL}, X_{HHH}, X_{HHL}, X_{HLL}, X_{HLH}. For instance, X_{LHL} is obtained after applying a low-pass filter along the x-dimension, a high-pass filter along the y-dimension and a low-pass filter along the z-dimension. The remaining images are built in a similar way, applying their respective sequence of low or high-pass filters in x, y and z-direction [82]. It must be noted that since the employed wavelet decomposition is undecimated, each decomposition is shift-invariant and shares the same size as the original image, which means the original tumour segmentation mask can be also used with the images obtained after the decomposition [64] (Supplementary Information). A schematic of the 3D wavelet transform applied to each CT image is presented in Figure 5.5.

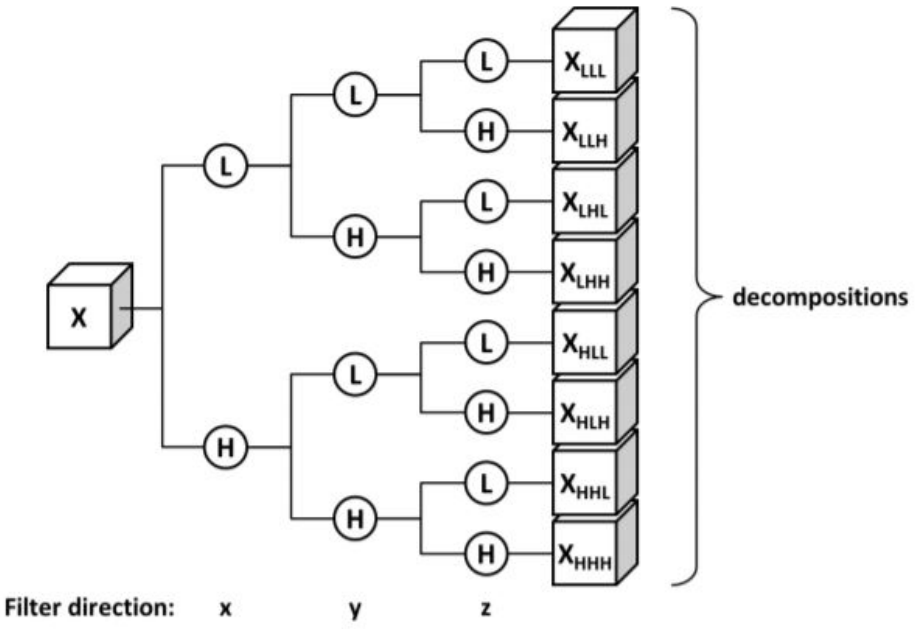


Figure 5.5: Schematic of the 3D wavelet transform applied to each CT image. The original image (X) is decomposed into 8 images, by applying directional low-pass and high-pass filters. Adapted from Aerts et al. [64] (Supplementary Information).

Concerning the LoG, five filters with different sigma values were applied (sigma=1.0 mm, 2.0 mm, 3.0 mm, 4.0 mm, 5.0 mm), with the intention of improving texture analysis by detection of multi-scale edges and ridges [83]. On balance, considering the original image and the resulting images after filter application, there were 14 different images to extract features for each sample.

The *Pyradiomics* package enabled the extraction of six classes of features: shape-based features (14 features), first-order features (18 features), GLCM features (22 features), GLRLM fea-

tures (16 features), GLSZM features (16 features) and GLDM features (14 features). Shape features include different descriptors of the size and shape of the ROI. Having in mind that shape descriptors are independent of intensity values, only the tumour segmentation masks were required for its computation. First-order features are related to the voxel intensities within the ROI using basic metrics (e.g. mean and standard deviation). GLCM features describe the second-order joint probability function of the ROI. GLRLM features define the length of successive pixels that share the same grey level intensity. GLSZM features quantify the grey level zones in an image, which represent the number of connected voxels that have the same grey level value. At last, GLDM features quantify grey level dependencies in an image. The list of features extracted is presented in Appendix A. A straightforward overview of the steps involved with extracting radiomic features using *Pyradiomics* is presented in Figure 5.6.

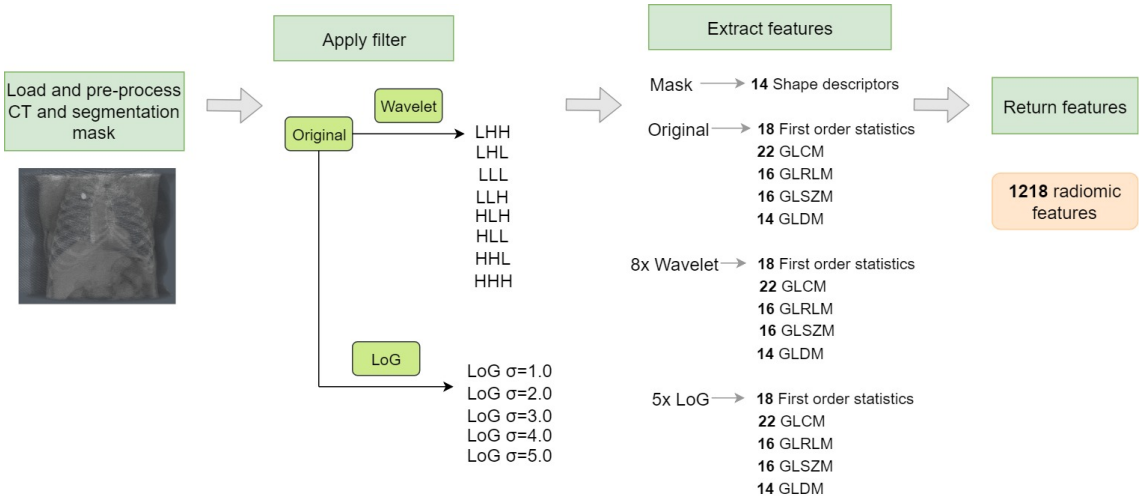


Figure 5.6: Overview of the process of feature extraction via *Pyradiomics*. First, medical images and segmentation masks are loaded into the software. Then, after filters have been applied to the original image, features are extracted from the resultant images.

5.5 Summary

This Chapter reviewed the data available to accomplish the proposed task. The available number of patients for each proposed task is presented in Table 5.1.

Table 5.1: Number of patients available to accomplish the proposed tasks.

Task	**Number of patients**
Semantic data	
EGFR mutation status	158
KRAS mutation status	157
Histological subtype	187
Radiomic data	
EGFR mutation status	116
KRAS mutation status	114
Histological subtype	261*

*Sum of patients from the two datasets.

Furthermore, it was described the methodology followed to convert the RTSTRUCT files available in the NSCLC-Radiomics dataset into 3D binary masks, in order to afterwards extract the radiomic features. Next, the pre-processing procedure of the CT images was described in detail, which included: conversion to HU units, resampling, normalisation and zero-centring. Finally, the process of extraction of radiomic features was described, in which 1218 features were returned.

Chapter 6

Mutation Status Prediction

The present chapter describes experiments that were conducted in order to achieve predictive models for EGFR and KRAS mutation status, along with understanding in what extent radiomic, clinical and semantic features are associated with mutations in these genes. Two individual studies were held using semantic features obtained from observation of CT images by experienced radiologists, and radiomic features extracted from CT images using *Pyradiomics*. Additionally, it was hypothesised whether the semantic feature emphysema could be predicted using radiomic features. Therefore, Section 6.1 focus on the experiments carried using semantic features to predict EGFR and KRAS mutation status, whereas, Section 6.2 has as main focus the selection of features related to EGFR and KRAS mutations, as well as the development of predictive models. At last, the experiment conducted regarding the prediction of the presence of emphysema is presented in Section 6.3. The data used in each study was divided into a training set (80%) and a test set (20%), ensuring class proportions are similar between the two sets. It is important to mention that the set of subjects used for training and testing were kept constant for all the performed experiments inside a Section.

6.1 Semantic Features

This study aimed to investigate whether clinical and semantic features are advantageous mutation status predictors and build predictive models using two algorithms: RF and Multi-layer Perceptron (MLP) networks. Hyper-parameters were chosen applying grid-search with 5-fold cross-validation to the training data and selecting the set of hyper-parameters of the model with the highest F-measure. In Figure 6.1 is presented a scheme which explains the data split process.

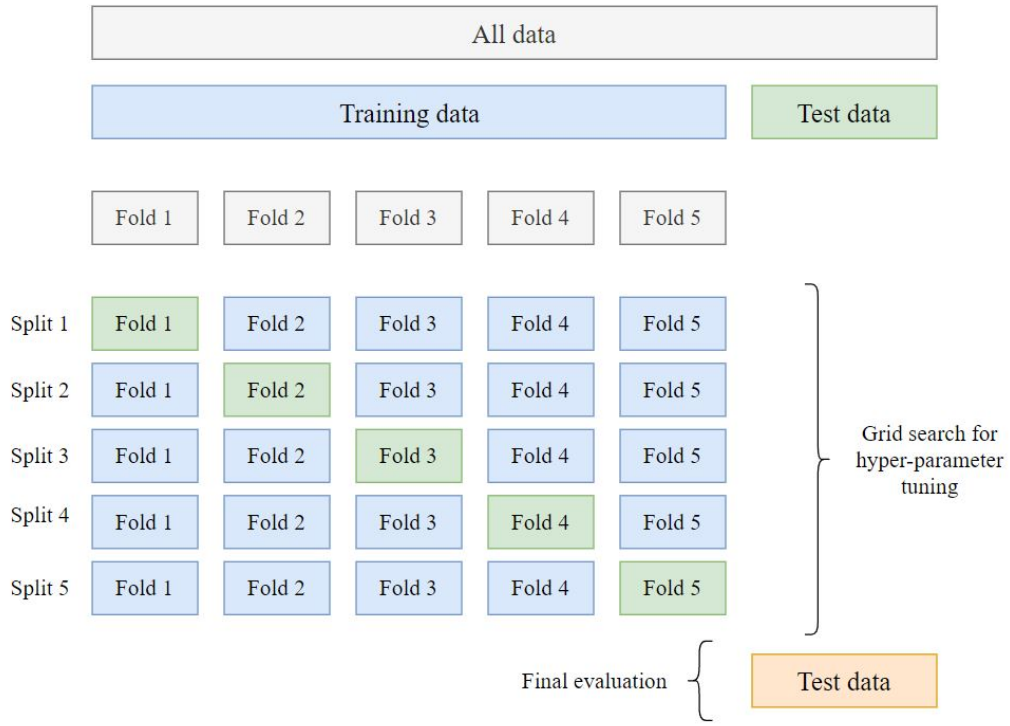


Figure 6.1: Five-fold cross-validation grid-search for hyper-parameter tuning. The data is divided into a training set and a test set. The training set is then divided into five folds, which combine into five splits. For each combination of hyper-parameters in the grid search the model is trained using each one of the possible splits where each fold is once the validation set. The model with the set of hyper-parameters which achieved the best performance is then evaluated in the test set.

6.1.1 Dataset

The study used the subset of subjects from the NSCLC-Radiogenomics dataset whose tumour was characterised by semantic features. It included 158 NSCLC patients tested for EGFR mutation status and 157 NSCLC patients tested for KRAS mutation status.

From the 30 semantic features available in the NSCLC-Radiogenomics dataset, some were discarded due to a large number of not applicable values (e.g. the fibrosis type field in a patient that has fibrosis absent), thus, a subset of 18 qualitative features was used in this study. The used set includes nodule and parenchymal features, which describe the nodules geometry, location, internal features and other related findings. Additionally, the patient's gender and smoking status were considered due to its significant association with mutation status prevalence, confirmed in recent studies [84–86]. From this point forward, gender and smoking status are designated as **clinical features** and the semantic features extracted from the images as **imaging features**. Table 6.1 shows detailed information regarding the data distribution and the nomenclature used to classify the tumours.

Table 6.1: Distribution of the semantic data used for KRAS and EGFR mutation status prediction experiments.

Feature	KRAS (%)	EGFR (%)
Clinical Features		
Gender		
Female	35.0	36.1
Male	65.0	63.9
Smoking Status		
Former	61.1	62.0
Non-smoker	24.2	22.8
Current	14.6	15.2
Imaging Features		
Axial Location		
Central	21.7	20.9
Peripheral	78.3	79.1
Nodule Attenuation		
Solid	56.1	57.6
Partially solid	40.8	39.2
Ground glass	3.2	3.2
Nodule Margins Primary Pattern		
Irregular	17.2	15.2
Lobulated	30.6	32.2
Spiculated	25.5	26.6
Poorly defined	16.6	15.8
Smooth	10.2	10.1
Nodule Margins Secondary Pattern		
Irregular	37.6	39.9
Lobulated	31.2	31.0
Spiculated	3.8	3.2
Poorly defined	16.6	15.2
Smooth	10.8	10.8
Nodule shape		
Complex	58.0	57.0
Round	29.3	29.7
Oval	12.1	12.7
Polygonal	0.6	0.6
Nodule calcification		
Peripheral	6.4	7.0
None	93.6	93.0
Nodule's findings		
Pleural retraction	10.2	10.1
Vascular convergence	20.4	20.3
Septal thickening	18.5	21.5
None	9.6	8.9
Attachment to pleura	7.6	7.6
Entering airway	21.7	20.3
Attachement to vessel	9.6	8.9
Bronchovascular bundle	2.5	2.5
Satellite Nodules in Primary Lesion Lobe (>4mm noncalcified)		
Absent	77.1	75.9
Solid	7.6	8.2
Partially-solid	5.1	5.1
Non-solid	10.2	10.8

Feature	KRAS (%)	EGFR (%)
Imaging Features		
Nodules in Non-Lesion Lobe Same Lung (>4mm noncalcified)		
Absent	82.8	81.6
Solid	5.7	6.3
Partially-solid	1.9	2.5
Non-solid	9.6	9.5
Nodules in Contralateral Lung (>4mm noncalcified)		
Absent	76.4	75.3
Solid	5.1	5.7
Partially-solid	7.6	7.6
Non-solid	10.8	11.4
Centrilobular Nodules		
Absent	87.3	88.6
Present	12.7	11.4
Emphysema		
absent	47.8	48.7
present	52.2	51.3
Fibrosis		
absent	89.8	88.6
present	10.2	11.4
Nodule Internal Features		
Not applicable	50.3	51.3
Cavitated	8.9	8.2
Internal air bronchogram sign	28.0	27.8
Reticulation	7.0	7.0
Necrosis	3.2	3.2
Nodule cysts	2.5	2.5
Lung Parenchyma Features		
Bronchial wall thickening	33.8	32.9
Tree-in-bud sign	1.9	2.5
Airway ectasia	9.6	9.5
Mosaic oligemia	5.1	5.1
Bronchial prominence	0.6	1.3
Bronchiectasis	3.8	4.4
Normal	45.2	44.3
Primary Emphysema Pattern		
Paraseptal	12.7	12.0
Centrilobular	37.6	38.0
Panacinar	1.9	1.3
Not applicable	47.8	48.7
Secondary Emphysema Pattern		
Paraseptal	17.2	17.1
Centrilobular	8.3	7.6
Panacinar	0.6	0.6
Not applicable	73.9	74.7
Nodule Periphery		
Emphysema	26.1	26.6
Normal	70.1	69.6
Fibrosis	3.8	3.8

The dataset comprises percentages of 26% and 25% mutated cases for EGFR and KRAS, respectively. Before feeding the data into the model, features were binarised following a one-hot encoding strategy [87] (see Figure 6.2). After that, the number of features increased from 20 to 73.

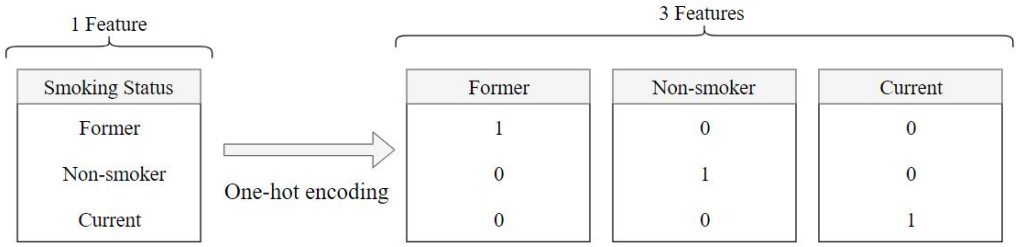


Figure 6.2: One-hot encoding a categorical feature.

6.1.2 Synthetic Minority Over-sampling Technique- Nominal and Continuous

In general, machine learning algorithms assume the distribution of classes is similar, which in this case would result in a model biased towards the negative class. However, in this study, the correct classification of both classes is equally important, since the classification of a patient with the wrong mutation status could lead to the administration of a less suitable treatment and, consequently, to shorter progression-free survival.

To overcome the class imbalance, it was conducted a Synthetic Minority Over-sampling Technique-Nominal and Continuous (SMOTE-NC) approach [88], in which the 5 nearest neighbours that belong to the minority class are found for each sample of the minority class, and then random points are generated within the line that connects the point and the neighbour (see Figure 6.3). In comparison to traditional over-sampling, SMOTE-NC has the advantage of building a more general decision region of the minority class [88]. After applying SMOTE-NC, the training set contains the same number of mutated and wildtype samples.

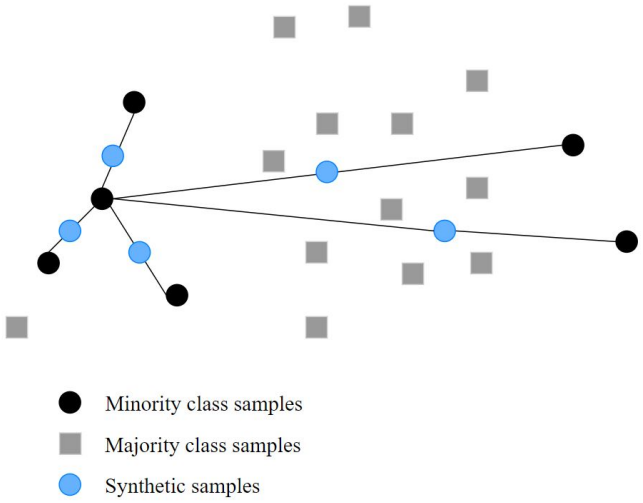


Figure 6.3: Generation of new samples from the minority class using SMOTE-NC.

6.1.3 Random Forest

As an algorithm based on ensemble learning, RF makes the predictions taking advantage of a group of models, instead of a single model. RF samples both observations and features in order to build independent Decision Trees and then the result is computed based on the majority of votes for classification problems (see Figure 6.4) and based on the average for regression problems. Bearing in mind that a Decision Tree is an unstable algorithm, by averaging the results of all Decision Trees, the variance component of the model will be minimised, which approximates the ensemble to an ideal model.

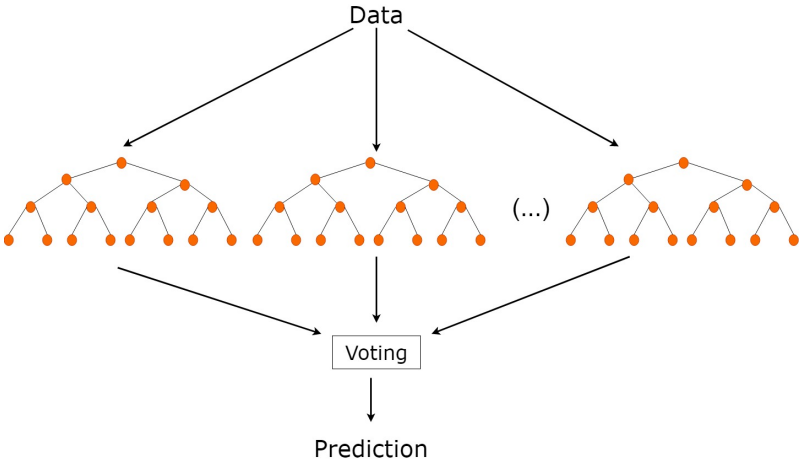


Figure 6.4: Random Forest classifier.

Additionally, RF has the ability to recognise the importance of the features. It perceives feature importance by beholding how much the tree nodes, which used that feature, decrease error over all the trees in the forest. When searching for a split while building the trees, different criteria can be used to decide which is the best split. Gini criterion, one of the most used criteria, finds features that contribute towards impurity reduction by summing the impurity decrease (weighted by the number of samples it splits) for each feature over all the nodes in all the trees [89]. Therefore, the smaller the impurity decrease, the less important is the feature. Further information regarding RF can be found in [50]. Due to the ability of RF to recognise the importance of the features for the task at hand, it was conducted an analysis to search for the most valuable ones.

6.1.4 Multi-layer Perceptron

Rosenblatt [90] proposed a perceptron network which is able to learn binary classification tasks empirically. However, the perceptron learning algorithm is confined to linearly separable classification tasks. The MLP surpasses this limitation using multiple layers of perceptrons and taking advantage of the backpropagation algorithm to solve non-linearly separable problems [24]. The MLP can be seen as a deep, artificial neural network. It is composed by an input layer which receives the input, an output layer which makes the prediction about the input, and in the middle

of those two, an arbitrary number of hidden layers which are the real computational engine of the MLP. Figure 6.5 illustrates a regular MLP network.

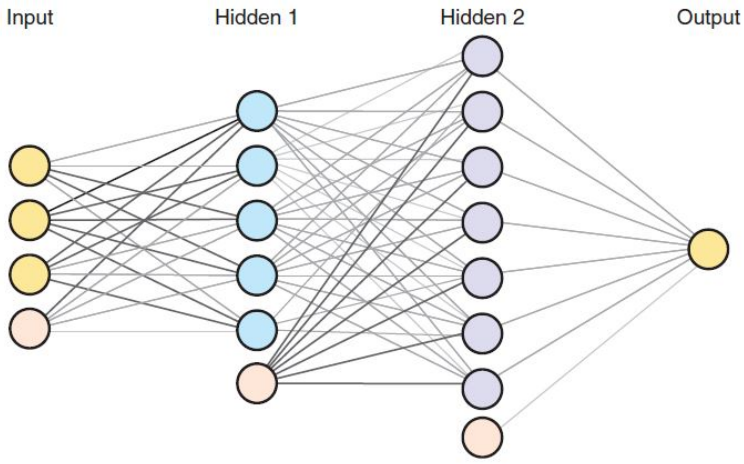


Figure 6.5: Multi-layer perceptron network with two hidden layers. Adapted from [24].

6.1.5 Results

The same experimental set-up was followed regarding EGFR and KRAS mutations. Interestingly, the average results were quite different, in a sense that it was possible to achieve models that reliably predict EGFR mutations but not KRAS mutation presence. From the experiments conducted, the predictive model for KRAS mutation with the best performance had an AUC of 0.56, having the RF as the classifier. Therefore, the following results are exclusively regarding EGFR.

6.1.5.1 EGFR Mutation Status Prediction

Three experiments were conducted in order to achieve the greatest set of features to predict EGFR mutation status: using clinical features, imaging features and both imaging and clinical features. Clinical features were only attempted with MLP, since the RF is not an ideal model for a set of 2 features, due to the lack of feature combinations. The range of values used in the grid search for RF and MLP for each hyper-parameter is presented in Table 6.2 and Table 6.3, respectively.

Table 6.2: Set of hyper-parameter values used in the grid search for RF.

Hyper-parameter	**Values**
Maximum depth	3, 5, 10, 15, 20, 25, 30
Maximum features	$\log_2 n$, $\sqrt{n}$, n (n: number of features)
Minimum samples split	2, 3, 5, 7, 9
Minimum samples leaf	1, 3, 5, 7
Number of estimators	30, 50, 70, 100, 300, 500, 700
Bootstrap	True, False

Table 6.3: Set of hyper-parameter values used in the grid search for MLP.

Hyper-parameter	Values
Alpha	0, 0.00001, 0.0001, 0.001, 0.01
Hidden layers	1, 2, 4, 6, 8, 10, 12
Learning rate	0.0001, 0.001, 0.01, 0.1
Momentum	0.0001, 0.001, 0.01, 0,1, 0.5, 0.9

The hyper-parameters of the models which achieved the highest F-measure in the 5-fold cross-validation for RF and MLP are included in Table 6.4 and Table 6.5, respectively, for each experiment.

Table 6.4: Hyper-parameters of the RF model which achieved the highest F-measure in the 5-fold cross-validation.

Experiment	Maximum depth	Maximum features	Minimum samples split	Minimum samples leaf	Number of estimators	Bootstrap
Imaging features	5	$\log_2 n^*$	2	3	300	False
Imaging + Clinical features	15	$\log_2 n^*$	7	1	30	False

*n: number of features

Table 6.5: Hyper-parameters of the MLP model which achieved the highest F-measure in the 5-fold cross-validation.

Experiment	Alpha	Hidden layers	Learning rate	Momentum	Activation	Optimiser
Clinical features	0.0	2	0.001	0.5	ReLU[a]	SGD[b]
Imaging features	0.0	10	0.0001	0.9	ReLU[a]	SGD[b]
Imaging + Clinical features	0.0	8	0.001	0.0001	ReLU[a]	SGD[b]

[a]Rectified Linear Unit
[b]Stochastic Gradient Descent

Table 6.6 shows the results obtained on the test set by the RF and MLP models which provided the best results in the training set, in the three performed experiments.

Table 6.6: Results obtained in the different experiments regarding EGFR mutation status prediction.

Experiment	Recall	Precision	F-measure	AUC
Clinical features				
MLP	0.56	0.50	0.53	0.68
Imaging features				
RF	**1.00**	**0.80**	**0.89**	**0.96**
MLP	**1.00**	0.73	0.84	0.94
Imaging + Clinical features				
RF	**1.00**	**0.80**	**0.89**	**0.96**
MLP	**1.00**	**0.80**	**0.89**	**0.96**

On average, a better performance was achieved when both clinical and imaging features were used. The RF model achieved an equal performance using both clinical and imaging features or just imaging features; however, MLP reached a higher performance (AUC=0.96) when clinical features were added to the feature set.

Focusing on the MLP experiments, clinical features achieved a satisfactory performance when predicting EGFR mutation status (AUC=0.68), whereas, when using imaging data a reliable predictive model is obtained (AUC=0.94). Further, when imaging and clinical data were combined, it was created a model which further increases the imaging data performance (AUC=0.96). The RF model achieved the same performance in the two experiments potentially due to its intrinsic feature subsampling.

The ROC curves for the RF and MLP models using imaging and clinical features are presented in Figure 6.6. The confusion matrix obtained by the two models using imaging and clinical features was identical, and it is presented in Figure 6.7.

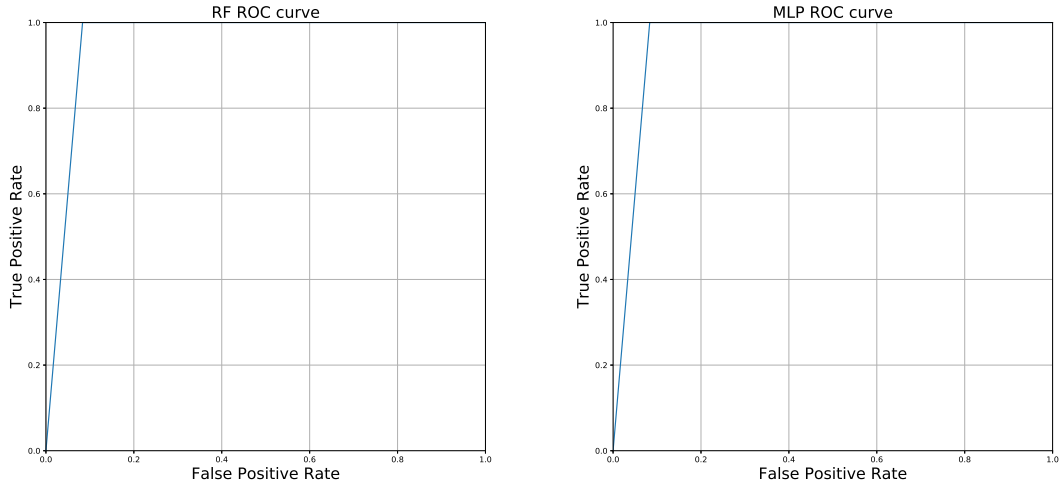


Figure 6.6: ROC curve of the RF model (left) and MLP model (right) using clinical and imaging features.

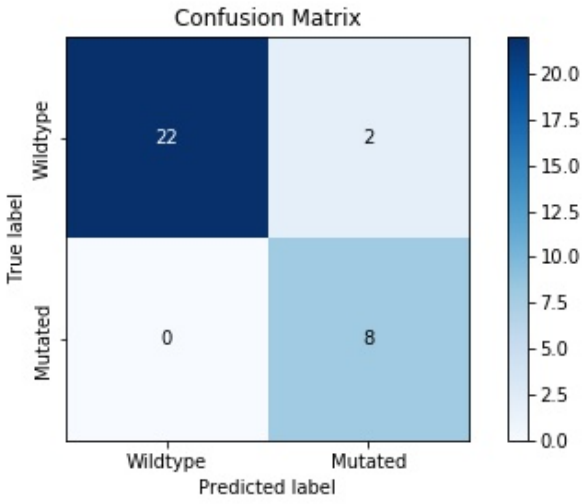


Figure 6.7: Confusion matrix of the RF and MLP models using clinical and imaging features.

Feature importance Taking advantage of the RF ability to recognise features importance, it was conducted an analysis in order to find the most valuable predictors amongst the used set of features. The RF model outputs a normalised score for each feature which sums to one.

Since the one-hot encoding approach was used to convert from categorical data to binary vectors, there is an importance score associated with each feature value and not a single score to the feature itself. Therefore, the feature score was considered to be the sum of the scores of each of its values. For instance, if having emphysema present had a score of 0.10 and emphysema absent a score of 0.08, the feature emphysema has a score of 0.18. Figure 6.8 shows the importance scores as a result of this analysis. Emphysema is plainly the feature with the highest score (0.18) followed by lung parenchymal features (0.16).

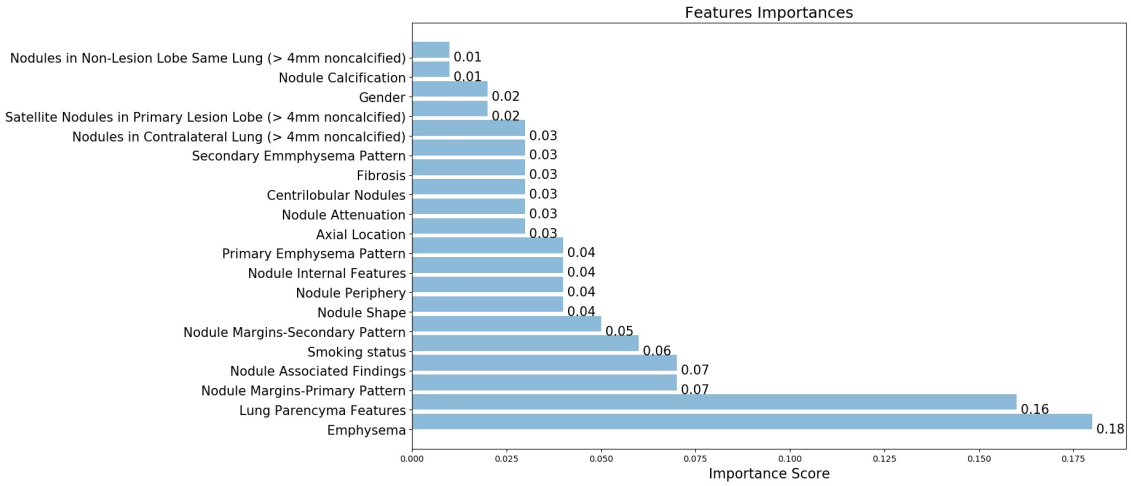


Figure 6.8: Features importance scores.

6.1.6 Discussion

As far as it was concerned, this work is the second attempt to create predictive models for EGFR and KRAS mutation status using qualitative radiographic image features. Even though our model outperforms the model developed by Gevaert et al. [39] (AUC=0.89), the present work did not use the same dataset as the first attempt and, consequently, results cannot be directly compared.

The results of this study suggest that an image signature exists that accurately predicts EGFR mutation status but not KRAS. Considering the fact that class proportions are similar in EGFR (26% mutated, 74 % wildtype) and KRAS (25 % mutated, 75% wildtype), the most reasonable hypothesis for this result is that KRAS mutations are not evident through semantic features in the same extent as EGFR mutations, which appear to have particular patterns. Gevaert et al. [39], which was also able to create a predictive model for EGFR mutations but not for KRAS, also hypothesised that the results might result from different class proportions between EGFR and KRAS; however, in the present study class proportions are similar, which enhances the hypothesis that KRAS mutations do not manifest through qualitative imaging features.

When clinical and imaging data were combined, the performance slightly increased, on average. Furthermore, it was concluded that from the total set of features used in this work, emphysema and lung parenchymal features were the ones which presented a higher correlation with EGFR mutation status. These results have shown that, although the limited amount of data, images and clinical data combined are potential predictors of EGFR mutation status.

6.2 Radiomic Features

This study had the main purpose of investigating the ability of clinical features and radiomic features extracted directly from the image, to predict the mutation status of EGFR and KRAS. Considering the vast number of extracted features (1218 features) and the short sample size, a feature selection procedure was followed in order to remove redundant and unnecessary features for the problem in mind. Thereafter, exclusively using the significant features, predictive models were developed taking advantage of appropriate machine learning algorithms concerning the number of selected features.

6.2.1 Dataset

The study was conducted using the subset of subjects from the NSCLC-Radiogenomics dataset which satisfied two requirements: tumour segmentation available and mutation status of EGFR and/or KRAS. Considering these conditions, there are 116 subjects available for studying EGFR mutation status and 114 subjects for studying KRAS mutation status.

Each subject was characterised by the 1218 tumoral radiomic features described in Section 5.4 and two clinical features (gender and smoking status). The inclusion of clinical features in this analysis was decided due to its association with EGFR and KRAS mutation status assessed in

previous studies [84–86] and also confirmed in the study of Section 6.1. Table 6.7 presents detailed information regarding the data distribution of the clinical features used in this study for each case.

Table 6.7: Clinical data distribution for both KRAS and EGFR-related experiments using radiomic experiments.

Feature	KRAS (%)	EGFR (%)
Gender		
Female	23.7	24.1
Male	76.3	75.9
Smoking Status		
Former	65.8	66.4
Non-smoker	14.0	13.8
Current	20.2	19.8

The used dataset contains percentages of 20% and 23% mutated cases for EGFR and KRAS, respectively.

6.2.2 Data Pre-Processing

Categorical features, namely gender and smoking status, were binarised following a one-hot encoding procedure, which increased the total number of features from 1220 to 1223. Also, radiomic features were standardised, by removing the mean and scaling to unit variance as shown in Equation 6.1:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{6.1}$$

where x is the feature value, μ and σ are the mean and the standard deviation of the training samples and x_{new} is the feature value after the standardisation. Centring and scaling are performed independently on each feature by computing the related statistics on the samples of the training set. Mean and standard deviation are then stored in order to apply the transform method on the test set. Standardisation is a usual procedure before feeding the data into many machine learning algorithms, since they may perform poorly if the individual features are not approximated with some extent to normally distributed data [24].

6.2.3 Feature Selection

The *curse of dimensionality* states that as the number of features increases, the data volume becomes more sparse, with lack of objects in various regions of the feature space. Thus, the number of objects required to induce high predictive performance models increases exponentially with the number of features. To overcome this problem, there are two main alternatives one may follow: to increase the number of objects or to reduce the number of features [24]. Since, in this case, it was not possible to increase the number of objects, it was conducted a two-step feature selection

procedure which returns the most relevant features by virtue of independence with other features and importance on the data. By removing unnecessary features, not only the generalisation performance of the model increases, but also the model interpretability and training speed.

The procedure includes two prime steps: removal of highly correlated features and removal of low importance features in accordance with a Gradient Boosting Machine (GBM) [91] model. The followed approach was motivated by the *feature selector*[1] tool.

Removal of highly correlated features: The existence of highly correlated features in the dataset leads to the decrease of the generalisation performance on the test set, due to the high variance. In this study, a correlation matrix is built, where the Pearson correlation coefficient is represented for each pair of radiomic features. Afterwards, one feature is removed for each pair that has a coefficient greater than 0.95. Experiments were attempted with correlation threshold values in a range from 0.90 to 0.99, and the set of features returned after the feature selection process remained the same. Thus, it was decided to use 0.95 as the correlation threshold value.

The Pearson coefficient evaluates the linear correlation between two features. In case the features increase proportionally, the Pearson correlation coefficient has the value 1, whereas, if they decrease proportionally its value is -1. A value of 0 is obtained when the points create a cloud without any increasing or decreasing trend. Positive values suggest the presence of a positive trend between the two features; hence as it becomes similar to a straight line, the Pearson correlation value becomes closer to 1. Likewise, negative values mean the existence of a negative trend, and as it becomes closer to -1, the trend becomes closer to a straight line [24].

Removal of low importance features: Taking advantage of the ability of tree-based machine learning models to recognise features importance for a given task, this step aims to find features that have almost zero importance based on a GBM model, from the LightGBM library [92, 93]. Differently from RF, boosting algorithms build trees one at a time and each new tree helps to correct errors made by the previously trained tree [24]. The number of estimators and the learning rate were chosen applying grid-search with 5-fold cross-validation to the training data and selecting the set of hyper-parameters used in the model which achieved the highest F-measure.

Similarly to RF, GBM randomly subsamples objects and features. Hence, features importance are averaged over 10 training runs, with regard to reducing variance. It is important to note that once the averaged features importances are computed, they are normalised to add up to one. Additionally, the model is trained using early stopping with a validation set of 15% of the training data to prevent overfitting. The training process stops once the F-measure on the validation set does not improve after 100 were built. Subsequently, only the features needed to reach 99% of the cumulative feature importance are retained to be used in the classification step.

[1] Available at: https://github.com/WillKoehrsen/feature-selector

6.2.4 Classification

Using the set of features preserved after the feature selection step, a data augmentation procedure was held using SMOTE or SMOTE-NC in case there were nominal variables in the feature set (i.e. gender or smoking status). The main difference between SMOTE and SMOTE-NC is that while SMOTE-NC deals with nominal and continuous data, SMOTE only handles continuous features [88].

Different machine learning algorithms were used to develop predictive models with the intention of finding the ones that best solve the classification problem. This time, since the dimensional space was substantially reduced, machine learning algorithms efficient in low-dimensional spaces were preferred, i.e. NB, LR, K-NN and SVM. For all the algorithms, except NB (does not require hyper-parameter tuning), hyper-parameters were selected from the model with the highest F-measure obtained from a grid-search with 5-fold cross-validation.

6.2.5 Results

An identical experimental set-up was followed concerning EGFR and KRAS mutation status. However, for a clearer understanding, the results achieved regarding KRAS and EGFR are described separately, in Section 6.2.5.1 and Section 6.2.5.2, respectively.

6.2.5.1 KRAS Mutation Status Prediction

From the total set of 1223 features, a total of 838 features were identified as highly correlated. After removing those features, the feature set still included 385 features. The correlation matrices obtained before and after the removal of highly correlated features are presented in Appendix B.1. Later, the set of 385 features was used to perform the hyper-parameter tuning of the GBM. The ranges of values used in the grid-search are presented in Table 6.8.

Table 6.8: Set of hyper-parameter values used in the grid search for GBM.

Hyper-parameter	Values
Number of estimators	200, 400, 600, 800, 1000, 1200
Learning rate	0.01, 0.02, 0.050, 0.1, 0.15

The GBM model which achieved the highest F-measure used 400 trees and a learning rate of 0.1. It identified that 380 features were not contributing to 99% of cumulative importance; hence, only 5 features remained as KRAS valuable predictors. In Figure 6.9 is represented how the cumulative importance varied with the number of features. Also, in Figure 6.10 is presented the importance scores of the features with a score different from zero according to the 10 training runs of the GBM model. From these two figures, one can conclude that only the top 5 features are valuable for predicting KRAS mutation status.

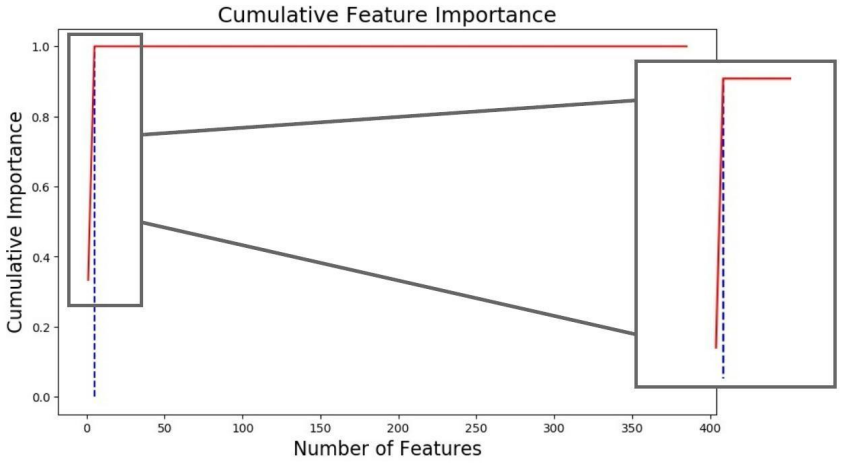


Figure 6.9: Cumulative importance versus the number of features regarding KRAS mutation status prediction. The vertical line highlights 99% of cumulative importance.

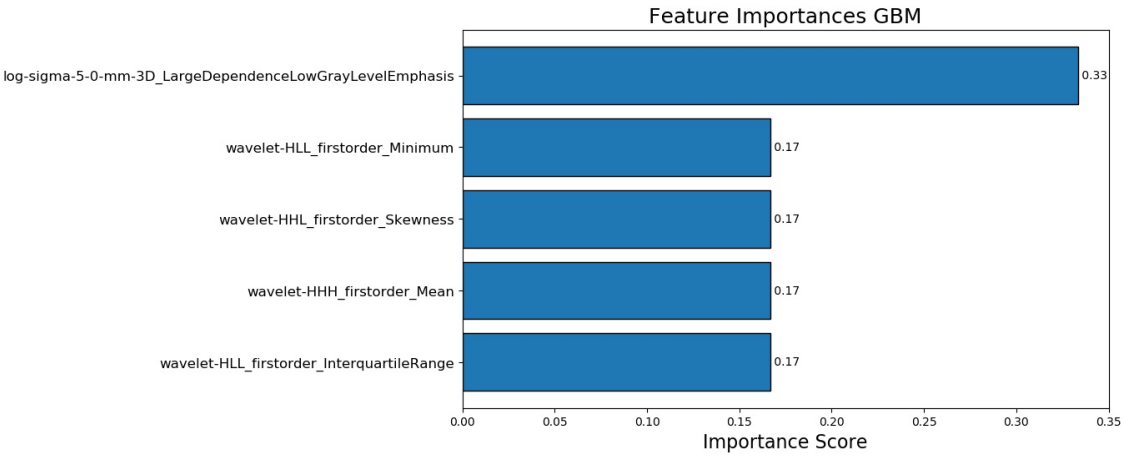


Figure 6.10: Features whose score is different from zero and respective score regarding KRAS mutation status prediction.

As a secondary endpoint, it was attempted to create a model which uses the total set of features (1223 features), in order to evaluate whether the feature selector procedure helps to indeed improve the model's generalisation ability. Thus, when using the total set of features the RF algorithm was preferred, due to its ability to deal with a large number of features.

Using exclusively the set of 5 features, SMOTE was applied to the training set, which increased the number of objects from 91 to 140, creating the fair number of mutant cases to level the number of wildtype cases. Regarding the RF experiment, SMOTE-NC was elected, since it includes the two categorical variables (i.e gender and smoking status). Hyper-parameter tuning was performed in the algorithms, when applicable, using the ranges of values presented in Table 6.9.

Table 6.9: Set of hyper-parameter values used in the grid-search for LR, K-NN, SVM and RF.

Algorithm	Hyper-parameter	Values
LR	**C**	0.0001, 0.001, 0.01, 0.1, 1, 10
K-NN	**Number of neighbours**	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
SVM	**Kernel**	RBF [a], linear
	Gamma	0.0001, 0.001, 0.01, 0,1, 0.5, 1
	C	0.1, 1, 5, 10 ,15, 20, 25
RF	**Maximum depth**	3, 5, 10, 15
	Maximum features	$\log_2 n$, $\sqrt{n}$, n (n: number of features)
	Minimum samples split	2, 3, 5
	Minimum samples leaf	1, 2, 3, 5
	Number of estimators	15, 25, 50, 75, 100
	Bootstrap	True, False

[a]Radial Basis Function

The hyper-parameter values of the models which achieved the highest F-measure in the cross-validation are presented in Table 6.10, regarding each used algorithm.

Table 6.10: Set of hyper-parameter values of the models which achieved the highest F-measure in the grid-search, regarding KRAS mutation status prediction.

Algorithm	Hyper-parameter	Value
LR	**C**	1
K-NN	**Number of neighbours**	1
SVM	**Kernel**	RBF [a]
	Gamma	0.5
	C	20
RF	**Maximum depth**	10
	Maximum features	$\log_2 n$ [b]
	Minimum samples split	3
	Minimum samples leaf	1
	Number of estimators	25
	Bootstrap	True

[a]Radial Basis Function
[b]n: number of features

Table 6.11 shows the results obtained by the previously mentioned models on the test set.

Table 6.11: Results obtained on the test set by the different models regarding KRAS mutation status prediction.

Model	Recall	Precision	F-measure	AUC
NB	**0.80**	0.27	0.40	0.59
LR	0.60	0.27	0.38	0.58
K-NN	0.60	**0.33**	**0.43**	**0.63**
SVM	0.40	0.22	0.29	0.51
RF*	0.00	0.00	0.00	0.42

*Using the total set of features.

The model which achieved the highest overall performance was the K-NN, with an AUC of 0.63. The NB and LR models reached a mediocre performance (NB: AUC=0.59 and LR: AUC=0.58), whereas SVM did not manage to attain a satisfactory result (AUC=0.51) since its class separation ability is inexistent. As expected, when all the extracted radiomic features are used, the developed model suffers from overfitting. By attempting to extract patterns in a huge dimensional space (1223 dimensions), the model ends up over-adjusting to the training set (F-measure: 0.89), weakening the possibility of generalisation when introduced to new objects (F-measure: 0.00).

The ROC curve and confusion matrix obtained by the model with the best performance (K-NN) are presented in Figure 6.11.

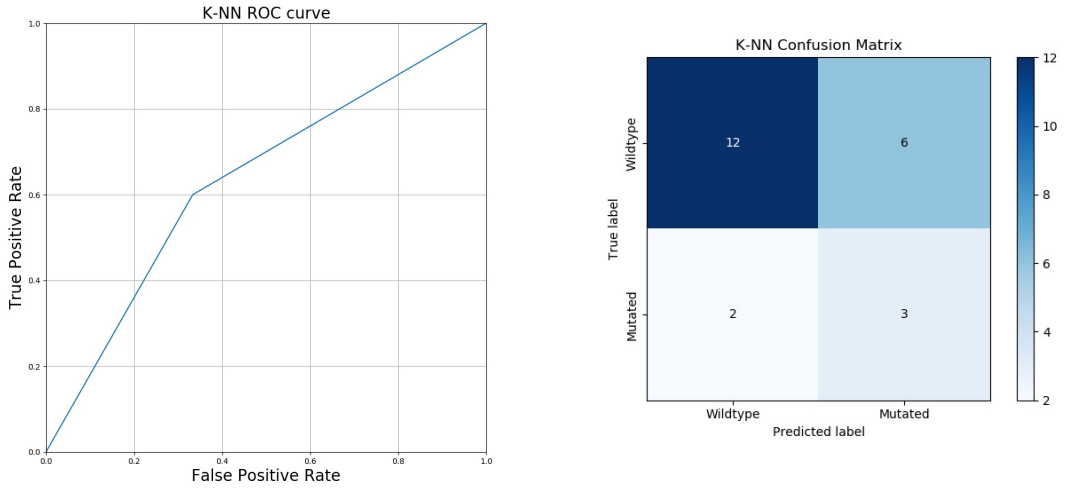


Figure 6.11: ROC curve (left) and confusion matrix (right) obtained by the model with the highest performance for predicting KRAS mutation status.

6.2.5.2 EGFR Mutation Status Prediction

Concerning the EGFR mutation status, from the total set of a 1223 features, 829 were removed after being identified as highly correlated with other features. As a result, the feature set included 394 features after the first feature selection step. The correlation matrices obtained before and after the removal of highly correlated features are presented in Appendix B.2.

The grid-search for the GBM model was performed using the same ranges of values as in the KRAS mutation status experiment, which are presented in Table 6.8. The GBM model which achieved the highest F-measure used 1000 trees and a learning rate of 0.02. Interestingly, only 2 features were needed to reach 99% of cumulative importance: minimum and skewness. In Figure 6.12 is represented how the cumulative importance increases with the number of features, whereas in Figure 6.13 is presented the importance scores of the the features whose importance is different from zero.

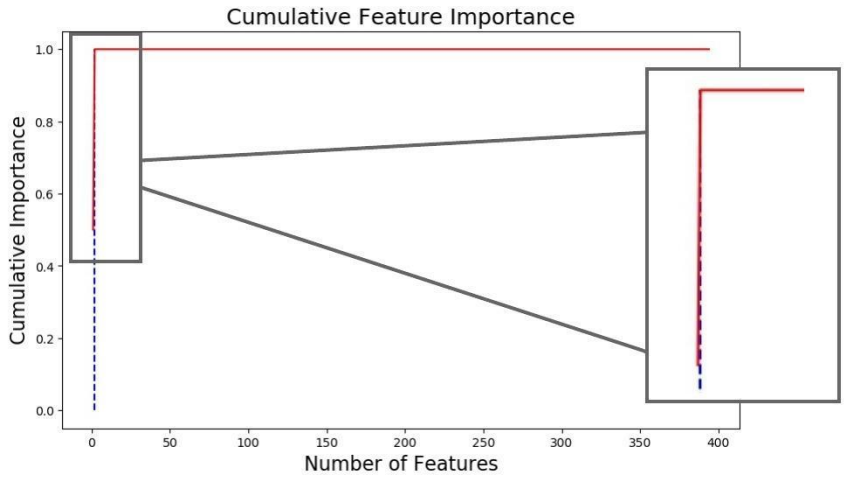


Figure 6.12: Cumulative importance versus the number of features regarding EGFR mutation status prediction. The vertical line highlights 99% of cumulative importance.

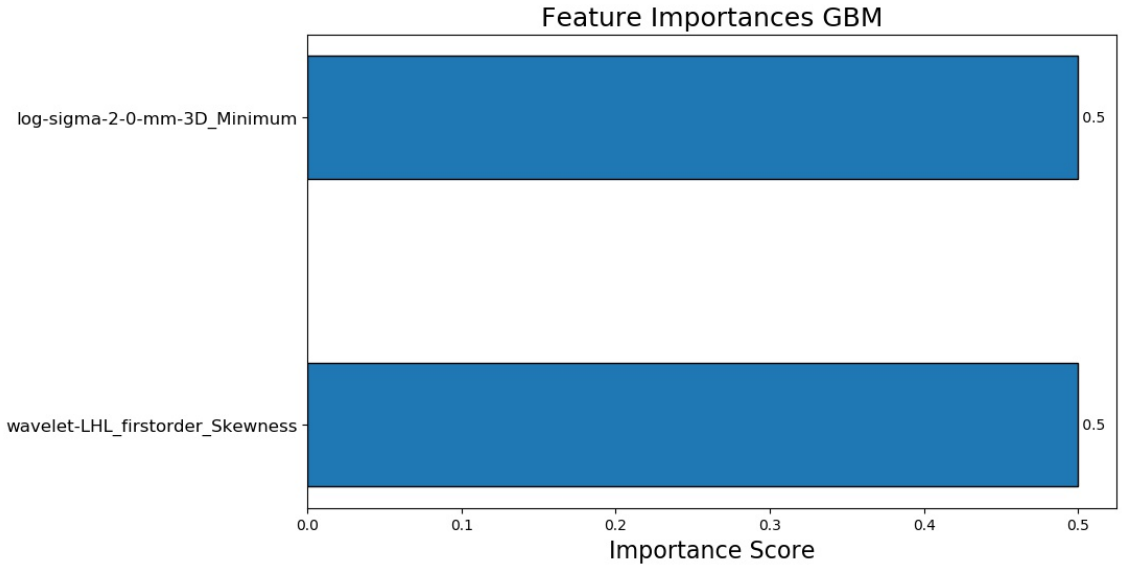


Figure 6.13: Features whose score is different from zero and respective score regarding EGFR mutation status prediction.

In order to inspect whether the boundary which divides EGFR mutant from wildtype can be easily seen through the two selected features, a scatter plot was produced highlighting in different colours the mutant and the wildtype objects. The same plot was generated after applying the SMOTE, which increased the number of objects from 92 to 148. Both scatter plots are presented in Figure 6.14.

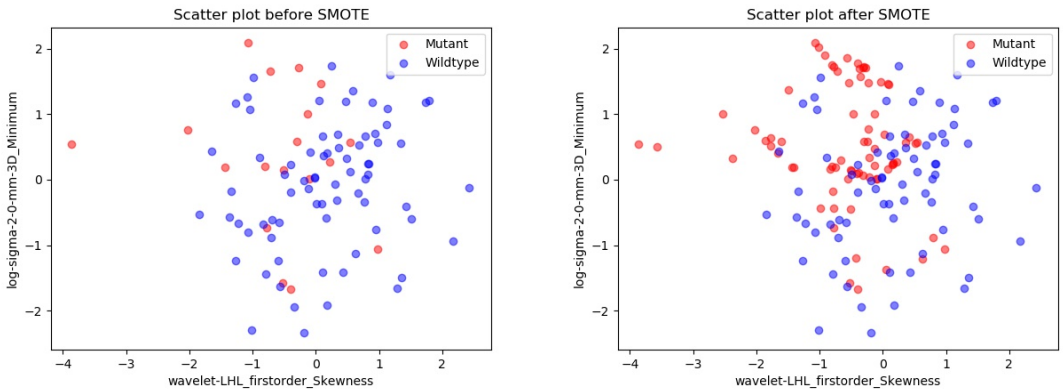


Figure 6.14: Scatter plot of the two selected features for EGFR mutation status prediction before (left) and after (right) applying SMOTE.

From focusing on the plots presented in Figure 6.14, it is not clear to find the partition between classes. Though, machine learning algorithms hold the ability to find the best decision region boundary. Hyper-parameter tuning was performed using the same grid of values as in the KRAS mutation status prediction experiment (Table 6.9) and, afterwards, the generalisation ability of the models was evaluated on the test set. The values of the hyper-parameters of the models which achieved the highest F-measure are presented in Table 6.12.

Table 6.12: Set of hyper-parameter values of the models which achieved the highest F-measure in the grid-search, regarding EGFR mutation status prediction.

Algorithm	Hyper-parameter	Value
LR	**C**	0.01
K-NN	**Number of neighbours**	9
SVM	**Kernel**	RBF [a]
	Gamma	0.1
	C	20
RF	**Maximum depth**	5
	Maximum features	n [b]
	Minimum samples split	2
	Minimum samples leaf	1
	Number of estimators	50
	Bootstrap	True

[a] Radial Basis Function
[b] n: number of features

The results obtained on the test set for the different models are presented in Table 6.13.

Table 6.13: Results obtained on the test set by the different models regarding EGFR mutation status prediction.

Model	Recall	Precision	F-measure	AUC
NB	**0.80**	0.36	0.50	0.72
LR	**0.80**	0.33	0.47	0.69
K-NN	0.60	0.43	0.50	0.69
SVM	**0.80**	**0.50**	**0.62**	**0.79**
RF*	0.00	0.00	0.00	0.47

*Using the total set of features.

The SVM model was the one which achieved the highest performance in the test set, with an AUC of 0.79. The remaining models, despite having lower results, also presented a satisfactory performance. Similarly to the experience with KRAS, when all the features were used, the model suffered from overfitting, with an F-measure of 0.93 in the training set and 0.00 in the test set.

The ROC curve and confusion matrix obtained using the SVM model are presented in Figure 6.15, whereas Figure 6.16 presents a representation of the SVM decision boundary.

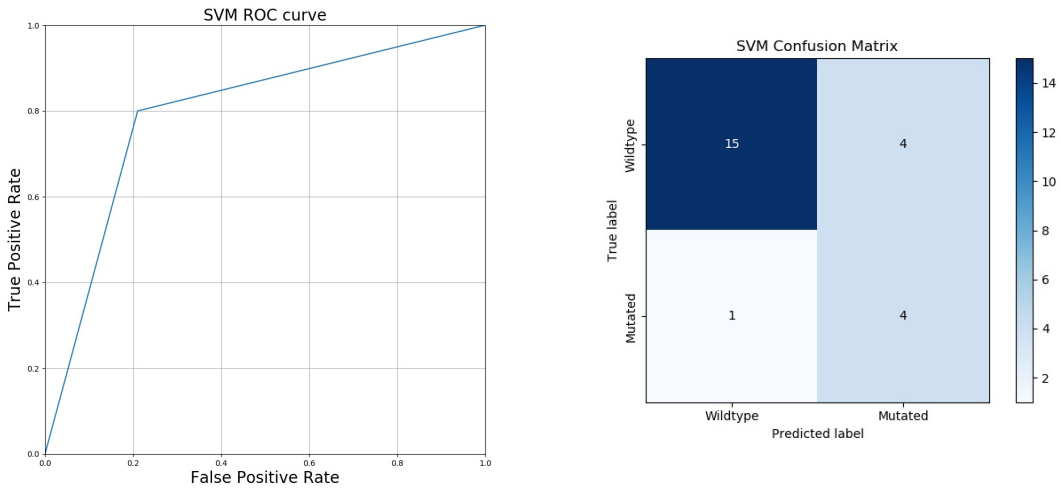


Figure 6.15: ROC curve (left) and confusion matrix (right) obtained by the model with the highest performance for predicting EGFR mutation status.

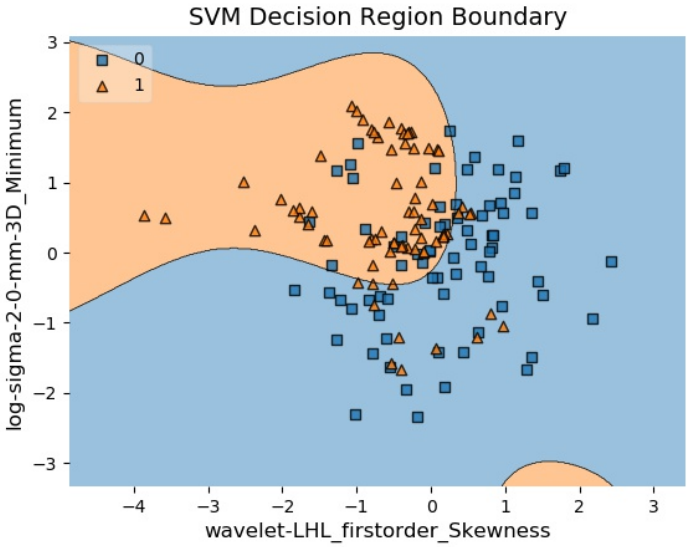


Figure 6.16: SVM decision region boundary between EGFR mutant and wildtype. Triangles represent the mutant cases and squares represent the wildtype cases.

6.2.6 Discussion

Unlike semantic features, radiomic features do not require human intervention for its extraction. However, the large number of radiomic features extracted demanded a feature selection step, which took into account features linearly correlated and its importance according to a GBM model. The use of this procedure to reduce the feature space dimensionality had a significantly positive impact since when the total set of features was used, the developed model suffered from overfitting. In this way, one can conclude that the feature selection procedure, not only reduced the dimensional space and consequently improved the mutation status prediction, but also decreased the computational time and increased the model interpretability.

It is important to bear in mind that it can not be fully assumed that the features returned by feature selection step are the most related to the mutation status since features that were linearly correlated were previously removed. Thereby, one can infer with some certainty about the radiomic features (e.g. minimum and skewness) related to the mutation status, but not about the type of filter applied (e.g. log-sigma-2-0-mm-3D-Minimum, Wavelet-LHL-firstorder-Skewness), since in a pair of correlated features, the one that is removed is chosen arbitrarily.

This study reached a result which supports that KRAS mutation status manifests through CT images to some extent. Although it does not enable a prediction with proper confidence (AUC=0.63), this time, the classification is not random as it was when using semantic features (Section 6.1). The fact that the K-NN model was the one which better differentiated between mutant and wildtype KRAS is surprising since it is a rather simple model. However, a recent study [94] confirmed that SMOTE can more easily attenuate the bias towards the classification in the majority class when K-NN is used as classifier. The literature review yield only one study that

created a model to predict KRAS mutation status. Weiss et al. [40] used accuracy as the evaluation measure, which does not allow to directly compare the results. In this study, the accuracy is not a good measure of evaluation since classes are not balanced on the test set. The accuracy in this study's best model is 65%; however, if the model predicted every object as wildtype the accuracy would be 78%. Certainly, the model developed in this study is preferable over the one that predicts everything as wildtype, since the number of FP and FN is of great importance in this context. Nevertheless, skewness and mean were identified as important features for predicting the mutation status of KRAS in both studies.

Furthermore, it was developed a model that reliably predicts EGFR mutation status (AUC=0.79), using the SVM algorithm as the classifier. Curiously, only two features presented significant importance when predicting EGFR mutation status, i.e. minimum and skewness. According to the literature revision, this is the first study which achieved notable results using a small set of two features. From the literature review presented in Chapter 4, two studies developed models which exceeded this model performance: Digumarthy et al. [36] created a model which achieved an AUC of 0.86 using the information from a single CT slice from each patient, and Li et al. [38] achieved a performance of 0.83 when using a combination of radiomic and CNN-based features. In spite of the great result, by using a single CT slice to characterise a tumour, Digumarthy et al. [36] risk missing plenty of information. Hence, to indeed validate the model, it would be important to increase the cohort's size and variance. Moreover, the methodology followed by Li et al. [38] is only feasible when a large number of patients is available, which is not the case of this study.

6.3 Predict Emphysema Presence Using Radiomic Features

Contrarily to semantic features, radiomic features do not demand human intervention. Thus, by looking at the outstanding performance of semantic features to predict EGFR mutation status (AUC=0.96), it was raised the hypothesis of whether radiomic features are able to predict semantic features. In particular, since emphysema was the semantic feature which mostly contributed to the EGFR mutation status prediction, this study attempted to create a model which predicts the presence or absence of emphysema receiving the radiomic features as input.

6.3.1 Dataset and Methodology

The exact same procedure and feature set used to predict EGFR and KRAS mutation status was used in this experiment. Summarising, redundant features with a Pearson correlation coefficient greater than 0.95 were removed as well as low importance features according to a GBM model. Afterwards, the oversampling technique SMOTE was applied and machine learning algorithms were used to create predictive models.

This study included 143 subjects, in which 84 (59%) had emphysema present and 59 (41%) emphysema absent. In similarity with the studies presented in the previous sections, the clinical features gender and smoking status were also included in the feature set. Table 6.14 presents information regarding the distribution of clinical data used in this experiment.

Table 6.14: Clinical data distribution for emphysema prediction experiments.

Feature	%
Gender	
Female	25%
Male	75%
Smoking Status	
Former	64%
Non-smoker	15%
Current	21%

6.3.2 Results

From the total set of 1223 features, 821 were identified as highly correlated, thus, a set of 402 features were used to train the GBM model. The correlation matrices obtained before and after the removal of highly correlated features are presented in Appendix B.3. The grid-search for the GBM model was performed using the same ranges of values as presented in Table 6.8. The model which achieved the highest F-measure used 800 estimators and a learning rate of 0.01. A set of 356 features were identified as low important for emphysema presence prediction and, consequently, only 46 features were used to create the predictive model. In Figure 6.17 is represented how the cumulative importance increases with the number of features and in Figure 6.18 is presented the importance scores of the top 10 most important features, according to the GBM model. The list of the 46 features used to predict emphysema presence is presented in Appendix C.

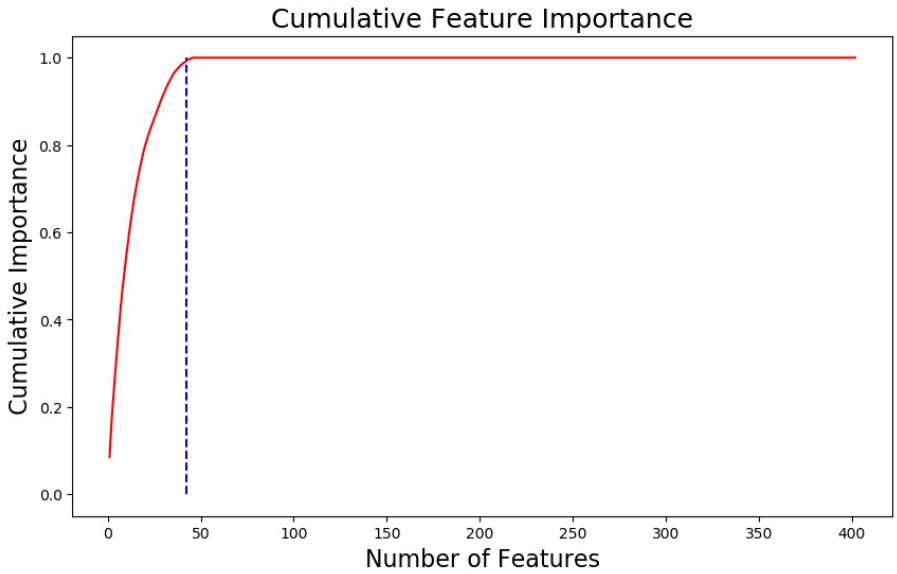


Figure 6.17: Cumulative importance versus the number of features regarding emphysema presence prediction. The vertical line highlights 99% of cumulative importance.

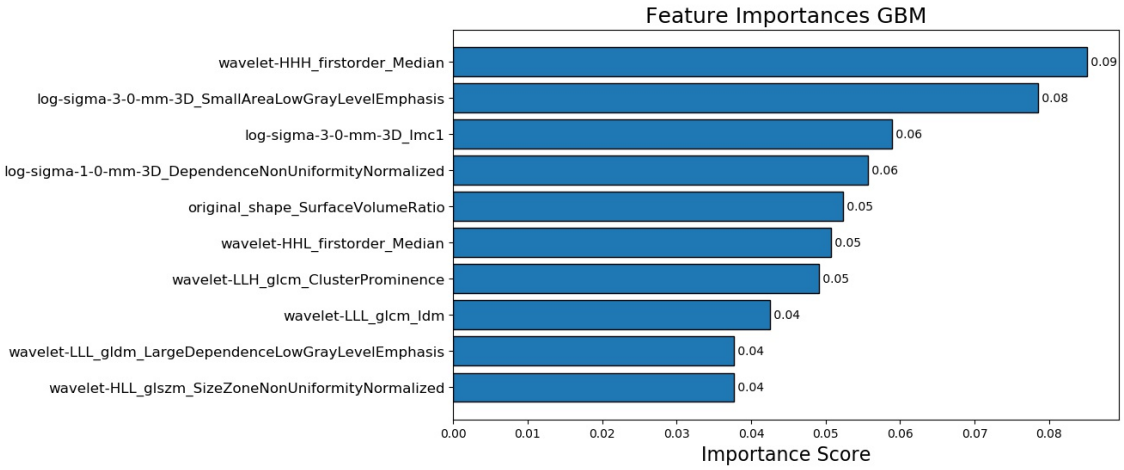


Figure 6.18: Top 10 most important features and respective importance scores regarding emphysema presence prediction.

The RF was the chosen classifier for the prediction of emphysema presence, due to the total number of features which remained after feature selection. Hyper-parameter tuning was performed using the values presented in Table 6.15 in the grid-search.

Table 6.15: Set of hyper-parameter values used in the grid search for RF, regarding emphysema presence prediction.

Hyper-parameter	Values
Maximum depth	3, 5, 10, 15
Maximum features	$\log_2 n$, $\sqrt{n}$, n (n: number of features)
Minimum samples split	2, 3, 5, 7, 9, 11
Minimum samples leaf	1, 3, 5, 7
Number of estimators	10, 20, 50, 100, 200
Bootstrap	True, False

The model which achieved the highest F-measure in the 5-fold cross validation used the following hyper-parameters: Maximum depth: 5, Maximum features: $\log_2 n$, Minimum samples leaf: 1, Minimum samples split: 9, Number of estimators: 20, Bootstrap: True. The results obtained by the model on the test set are presented in Table 6.16.

Table 6.16: Results obtained in the test set regarding emphysema presence prediction.

Experiment	Recall	Precision	F-measure	AUC
RF	0.94	0.67	0.78	0.64

The ROC curve and confusion matrix of the model are presented in Figure 6.19.

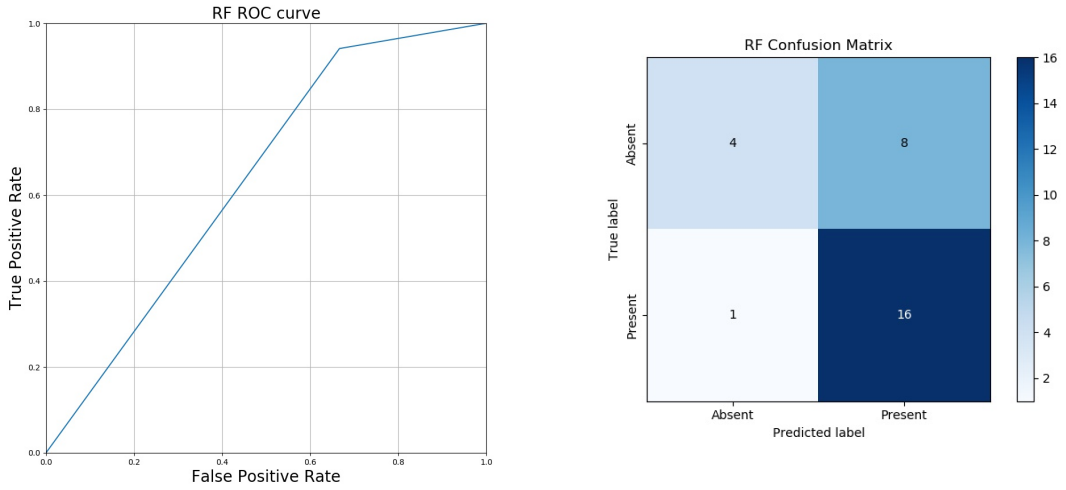


Figure 6.19: ROC curve and confusion matrix obtained by the model created to predict emphysema presence.

6.3.3 Discussion

From the obtained results, one can assume that there is a chance of predicting the presence of emphysema using radiomic features. Nevertheless, this result is limited to this dataset and, consequently to support with security this association; it is vital to validate the model with an independent dataset.

A literature review yielded no study with this purpose. Therefore, this result opens interesting opportunities in the field of radiomics. Considering that there is a way which does not require human intervention to predict the presence of features that non-invasively and reliably predict EGFR mutation status, the scientific community can more quickly find a solution that is effectively used in clinical practice.

6.4 Summary

In this Chapter, two different studies were held with the intention of creating predictive models of EGFR and KRAS mutation status. While the study described in Section 6.1 used semantic features annotated by radiologists, the study presented in Section 6.2 used radiomic features directly extracted from the CT images. The two conducted studies can not be directly compared in terms of power to predict the mutation status since the number of available subjects is different. In such a way, one can not assume that semantic features are more powerful predictors than radiomic features since the number of available subjects with radiomic features is shorter. Nevertheless, there are some conclusions which can be taken from this Chapter, regarding the used dataset:

- It is possible to achieve a reliable predictive model (AUC=0.96) to predict EGFR mutation status using a combination of semantic and clinical features.

- The combination of semantic and clinical features does not express KRAS mutation status with the same extent as EGFR mutation status.
- Using clinical and radiomic features it is possible to create a model which predicts EGFR mutation status with confidence (AUC=0.79).
- Using radiomic and clinical features, it is possible to create a model which differentiates at some extent between mutant and wildtype KRAS.

In the light of the pleasant performance of semantic features to predict EGFR mutation status, it was a conducted an experiment in Section 6.3, which aimed to predict the most important semantic feature, according to the experiment in Section 6.1 - emphysema- using radiomic and clinical features. The satisfactory result obtained (AUC=0.64) opens new opportunities to radiomics.

Chapter 7

Tumour Histological Subtype Prediction

This Chapter presents experiments that were performed with the purpose of reasoning whether clinical and radiomic features are related to tumour histological subtype as well as to develop a model capable of reliably differentiate between an Adenocarcinoma and a Squamous Cell Carcinoma. The radiomic features which were extracted directly from the CT images using *Pyradiomics* were evaluated in terms of their ability to differentiate between Adenocarcinoma and Squamous Cell Carcinoma and used to create a predictive model. Therefore, the dataset used in this study is described and analysed in Section 7.1, the methodology pursued and respective results are presented in Section 7.2 and Section 7.3, respectively. Lastly, a discussion of the results is presented in Section 7.4.

Experiments using semantic features are not described in this Chapter, due to the impoverished results obtained when predicting the NSCLC histological subtype. One of the plausible reasons for the weak performance may be the strong class imbalance of the dataset since from the 188 patients available only 16% were Squamous Cell Carcinomas.

7.1 Dataset

The study was conducted using the subset of subjects from the NSCLC-Radiogenomics [14, 21] and NSCLC-Radiomics [74], which fulfil the two requirements: tumour segmentation available and classified as either Adenocarcinoma or Squamous Cell Carcinoma. In conformity with these two conditions are 141 subjects from the NSCLC-Radiogenomics dataset and 120 subjects from the NSCLC-Rdiomics dataset. In order to inspect whether there is a great difference between the dimensions of the tumours of the two datasets, which could affect the decision boundary of the model, it was performed an analysis regarding the average tumour dimensions of each one of the cohorts. The results of the analysis are presented in Table 7.1.

Table 7.1: Statistics of the dimensions of the tumours in the two used datasets and in the total cohort, in millimetres.

Dataset	Depth (μ, σ)	Width (μ, σ)	Height (μ, σ)
NSCLC-Radiomics	60.8 (28.6)	53.9 (24.4)	52.2 (23.2)
NSCLC-Radiogenomics	35.4 (20.7)	35.6 (18.6)	35.9 (19.3)
Both datasets	46.9 (27.7)	43.9 (23.3)	43.3 (22.6)

From the results presented in Table 7.1, one can reach the following conclusions:

- On average, the NSCLC-Radiomics dataset includes larger tumours compared to the NSCLC-Radiogenomics dataset.
- The averages in each of the 3 dimensions (depth, width, height) are similar among each other. This conclusion is more evident in the NSCLC-Radiogenomics dataset than in the NSCLC-Radiomics dataset. This fact leads to that in the total dataset the average size in the 3 dimensions is similar.

This time, only the gender was used as a clinical feature since the NSCLC-Radiomics dataset does not include information regarding the smoking status. The total set comprises 198 male patients (76%) and 65 female patients (24%). Regarding the predictive label, the dataset includes 118 Squamous Cell Carcinomas (45%) and 143 Adenocarcinomas (55%). From this point forward the Squamous Cell Carcinoma were tagged as 0 and the Adenocarcinomas as 1. In addition to the clinical feature gender, the set of radiomic features described in Section 5.4 was extracted from the set of CT images, which resulted in a total set of 1219 features. The dataset was then split into a training set (80%) and a test set (20%), ensuring class proportions are similar among the two sets.

7.2 Methodology

A one-hot encoding procedure was applied to the categorical feature gender, which increased the feature set from 1219 features to 1220 features. Each radiomic feature was standardised, by removing the mean and scaling to unit variance. The standardisation statistics were computed on the training set, and then the same transform method was applied to the samples on the test set. Given that the ratio between the number of features and the number of cases is enormous, the exact same feature selection procedure described in Section 6.2.3 was held in this study, which included the removal of highly correlated and with low importance features. Using the set of features preserved after the feature selection step, a SMOTE-NC or SMOTE technique was applied depending on whether the feature set included the categorical feature (gender), with a view to balancing the number of patients belonging to each class in the training set. An appropriate machine learning algorithm according to the number of selected features was trained to predict whether the NSCLC subtype is Adenocarcinoma or Squamous Cell Carcinoma. The model's hyper-parameters were

chosen performing a 5-fold cross-validation grid-search and selecting the hyper-parameter values from the model which achieved the highest F-measure.

7.3 Results

From the set of 1220 features, 851 were identified as highly correlated. Therefore, after the removal of the correlated features, the feature set included 369 features. The correlation matrices generated before and after the removal of the highly correlated features are presented in Appendix B.4.

Using the set of 369 features, a 5-fold cross-validation grid-search was performed for the GBM in order to return the set of hyper-parameters of the model which achieved the highest F-measure in the training set. The values used in the grid search for GBM are the ones presented in Table 6.8. Thereby, the GBM model which achieved the highest F-measure used 600 trees and a learning rate of 0.01. Following, the model was used to investigate which features from the set of 369 features were not contributing to 99% of cumulative importance to predict the histological subtype. All in all, only the top 54 features were valuable for differentiating between Adenocarcinoma and Squamous Cell Carcinoma. The list of the 54 features used to build the predictive model is included in Appendix D. In Figure 7.1 is represented how the cumulative importance changes with the number of features. Plus, in Figure 7.2 is presented the importance scores of the top 10 most important features according to the 10 training runs of the GBM model.

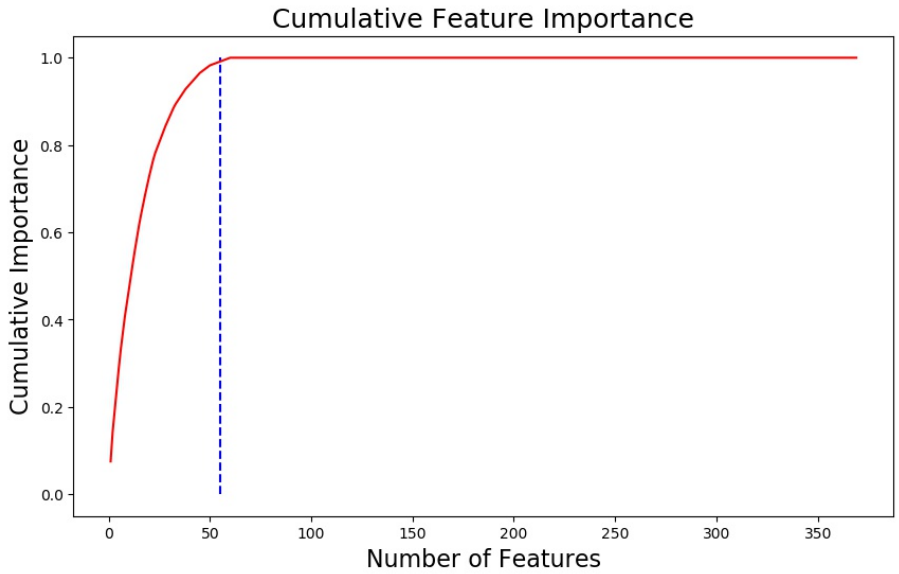


Figure 7.1: Cumulative importance versus the number of features regarding histological subtype prediction. The vertical line highlights 99% of cumulative importance.

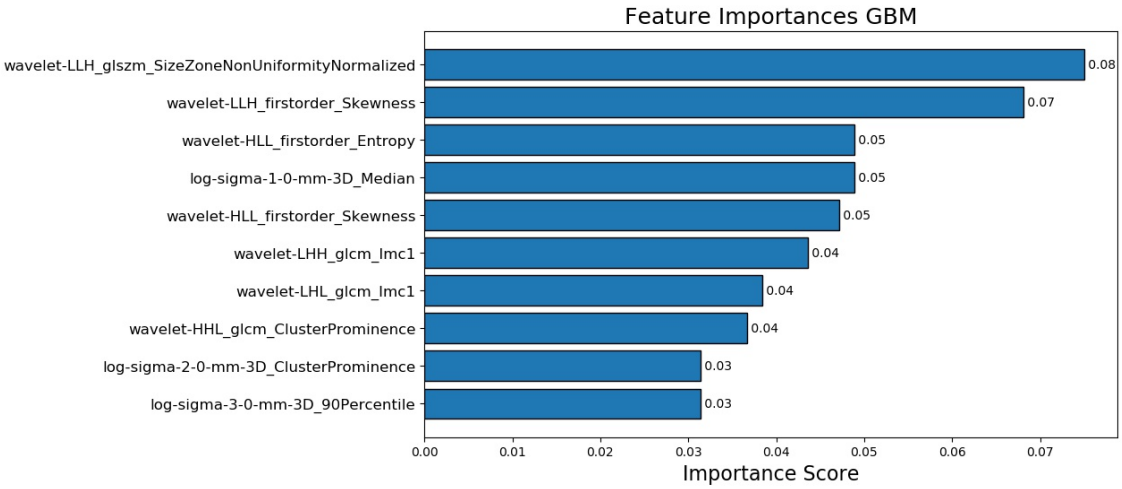


Figure 7.2: Top 10 most important features and respective importance scores regarding histological subtype prediction.

The SMOTE-NC was applied to the set of 54 features since one of the preserved features was the female gender. Considering the number of selected features, the RF classifier was chosen to build the predictive model. An experiment using the total set of 1220 features was also conducted in order to compare the results and validate the performance of the feature selection methodology. Likewise, the SMOTE-NC was also applied to the total set of features.

Hyper-parameters for both models were selected from a grid-search with 5-fold cross-validation, using the set of values presented in Table 7.2.

Table 7.2: Set of hyper-parameter values used in the grid search, regarding histological subtype prediction.

Hyper-parameter	**Values**
Maximum depth	3, 5, 10, 15, 20
Maximum features	$\log_2 n$, $\sqrt{n}$, n (n: number of features)
Minimum samples split	2, 3, 5
Minimum samples leaf	1, 2, 3, 5
Number of estimators	25, 50, 75, 100, 125
Bootstrap	True, False

The hyper-parameters of the models which achieved the highest F-measure in the 5-fold cross-validation are included in Table 7.3, for each experiment.

Table 7.3: Hyper-parameters of the models which achieved the highest F-measure in the 5-fold cross-validation, regarding histological subtype prediction. RF: model which uses the total set of 1220 features; FS+RF: model which uses the features returned after the feature selection step.

Experiment	**Maximum depth**	**Maximum features**	**Minimum samples split**	**Minimum samples leaf**	**Number of estimators**	**Bootstrap**
RF	5	$\log_2 n^*$	2	3	50	True
FS+RF	15	$\log_2 n^*$	3	1	100	False

*n: number of features

The results obtained by each model in the test set are presented in Table 7.4. Given that in this experiment, the test set is relatively balanced (55% vs. 45%), the metric accuracy was also found useful to be included in the results.

Table 7.4: Results obtained by the two models regarding NSCLC histological subtype prediction. RF: model which uses the total set of 1220 features; FS+RF: model which uses the features returned after the feature selection step.

Experiment	**Recall**	**Precision**	**F-measure**	**AUC**	**Accuracy**
RF	0.72	0.72	0.72	0.70	0.70
FS+RF	0.72	0.72	0.72	0.70	0.70

The ROC curve of each one of the two models is presented in Figure 7.3. The confusion matrix obtained by the two models was identical, and it is presented in Figure 7.4.

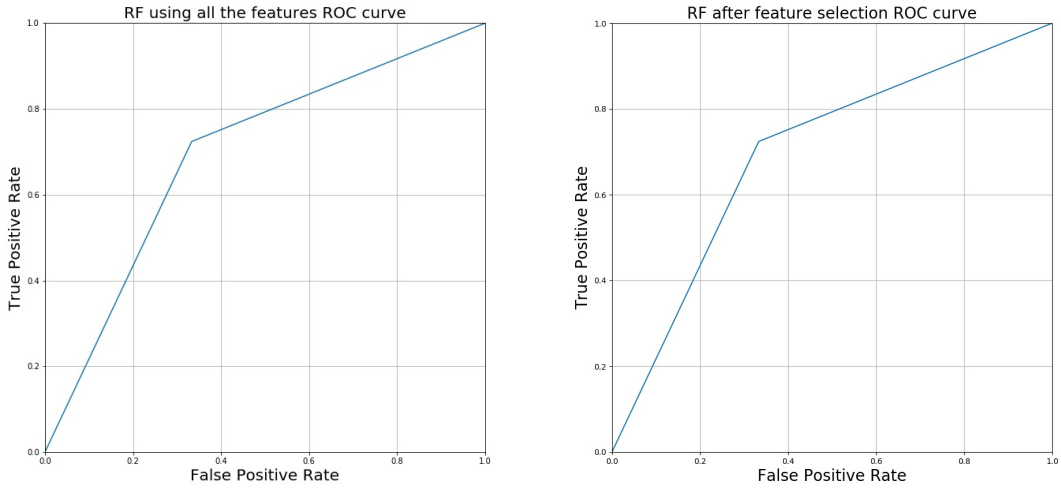


Figure 7.3: ROC curves of the RF models created to predict NSCLC histological subtype. Left: model which uses all the features. Right: model which uses the features returned by the feature selection process.

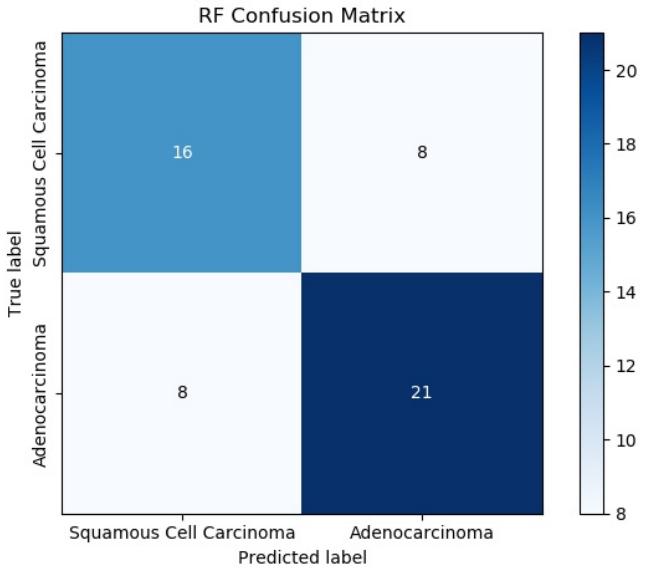


Figure 7.4: Confusion matrix of the RF models created to predict NSCLC histological subtype.

7.4 Discussion

Considering the obtained results, one can conclude that there is an association between radiomic features and the differentiation between Adenocarcinoma and Squamous Cell Carcinoma since the models can distinguish with certainty between these two subtypes of NSCLC.

Interestingly, the results were precisely the same in the two experiments. Unlike the results obtained in Chapter 6, in which the RF using all the features was not able to generalise, in this case, it was possible to generalise even using a quite extensive set of features. There are two hypotheses that may justify this result: the larger number of patients confronts in some extent the curse of dimensionality, and the differentiation between Adenocarcinoma and Squamous Cell Carcinoma is more evident through radiomic features than the EGFR and KRAS mutation status. Anyhow, the model which used feature selection is preferable since it significantly reduced training time and increased the interpretability of the model.

From the literature review, two studies with the same purpose created models which exceeded the performance achieved by the model created in this study: Wu et al.[57] compared 24 feature selection and 3 classification methods and achieved a similar performance (AUC=0.72) as the one reported in this study; whereas Digumarthy et al.[36] achieved the great performance of AUC=0.92, using radiomic features extracted from the slice in which the tumour has the largest axial dimension. Despite the great result of Digumarthy et al., by using only one slice, one risks to waste valuable information. Therefore, it would be advantageous to validate the performance using a larger cohort.

7.5 Summary

In this Chapter, it was described experiments conducted with the intention of developing a predictive model which differentiates between Adenocarcinoma and Squamous Cell Carcinoma, using a set of radiomic features and the clinical feature gender as input. Having in mind the high-dimensionality of the feature set, features went through a two-step feature selection procedure which removed features based on redundancy and low importance. Both the reduced feature set and the complete feature set were used two create predictive models, using the RF as the classifier. The two models achieved the same promising performance (AUC=0.70), which suggests that radiomic features and the clinical feature gender are effective predictors of the NSCLC histological subtype.

Chapter 8

Conclusions

Advances in genomics have proven that tumours are populated by different minor subclones of malignant cells that drive tumour development and progression. The tumoural spatial and temporal heterogeneity has been a barrier for the customisation of cancer care to the individual patient, since the current clinical practice to monitor a tumour is by invasive biopsy, which is incapable of seizing the full state of a tumour. This work presented predictive models which in a non-invasive way characterise a tumour, in order to avoid some biopsies without limiting the number of times a tumour is monitored.

Medical imaging is a non-invasive technique that can present a complete picture of the tumour phenotype and its environment at a macroscopic level. In this work, different models were created in which a tumour is described by semantic features, visually assessed by radiologists, by clinical features, and/or by radiomic features, extracted using the open-source package *Pyradiomics*. The association between the tumour features and valuable genetic factors, namely EGFR and KRAS mutation status and NSCLC histological subtype, were assessed and predictive models were created.

This study resulted in predictive models of EGFR and KRAS mutation status, emphysema presence and NSCLC histological subtype prediction, using a vast set of features never used in the literature obtained by an open-source software. In this work, it was concluded that both semantic features and radiomic features are powerful predictors of EGFR mutation status, with performances of AUC=0.96 and AUC=0.79, respectively. Although a model capable of predicting KRAS mutation status using radiomic features was achieved (AUC=0.63), it did not attain the same predictive power as the model which predicts EGFR mutation status (AUC=0.79). Furthermore, it was not found any pattern between semantic features and KRAS mutation status, which raises the hypothesis that KRAS mutations are not evident through CT images in the same extent as EGFR mutations. The outstanding performance of semantic features to predict EGFR mutation status (AUC=0.96) raised the ambition of predicting the most important semantic feature-emphysema- using radiomic features. The promising result (AUC=0.64) supports the chance of predicting the presence of emphysema using radiomic features. Moreover, it was created a model which differentiates between Adenocarcinoma and Squamous Cell Carcinoma (AUC=0.70), using

radiomic features.

Although the models which were created achieved promising results, the sample size was not sufficiently large to achieve solid conclusions. The main conclusion of this study is that the studied genetic factors can indeed express through CT images. Thereby, in the future, the models' performance ought to be validated in independent cohorts, and the proposed methodology should be repeated using a larger sample size. A key issue to be investigated in the future is whether models created using deep learning techniques exceed the results presented in this study. It would also be valuable to statistically ascertain the relation between each feature and the target variable in order to understand which features are stronger predictors and check whether it matches with the results achieved in this study. Furthermore, it is important to emphasise the value of creating standard protocols, so that the results obtained by different studies can be compared and these models can be in the near future used in clinical settings.

Appendix A

List of Extracted Features

Shape features:

Mesh Volume

Voxel Volume

Surface Area

Surface Area to Volume ratio

Sphericity

Maximum 2D Diameter (Slice)

Maximum 2D Diameter (Column)

Maximum 2D Diameter (Row)

Maximum 3D Diameter

Major Axis Length

Minor Axis Length

Least Axis Length

Elongation

Flatness

First-order statistics features:

10th Percentile

90th Percentile

Energy

Entropy

Interquartile Range

Kurtosis

Maximum

Mean Absolute Deviation

Mean

Median

Minimum

Range

Robust Mean Absolute Deviation

Root Mean Squared

Skewness

Total Energy

Uniformity

Variance

GLCM features:

Autocorrelation

Joint Average

Cluster Prominence

Cluster Shade

Cluster Tendency

Contrast

Correlation

Difference Average

Difference Entropy

Difference Variance

Joint Energy

Joint Entropy

Informational Measure of Correlation (IMC) 1

Informational Measure of Correlation (IMC) 2

Inverse Difference Moment (IDM)

Inverse Difference Moment Normalised

Inverse Difference (ID)

Inverse Difference Normalised (IDN)

Inverse Variance

Maximum Probability

Sum Entropy

Sum of Squares

GLRLM features:

Grey Level Non-Uniformity

Grey Level Non-Uniformity Normalised

Grey Level Variance

High Grey Level Run Emphasis

Long Run Emphasis

Long Run High Grey Level Emphasis

Long Run Low Grey Level Emphasis

Low Grey Level Run Emphasis

Run Entropy

Run Length Non-Uniformity

Run Length Non-Uniformity Normalised

Run Percentage

Run Variance

Short Run Emphasis

Short Run High Grey Level Emphasis

Short Run Low Grey Level Emphasis

GLSZM features:

Grey Level Non-Uniformity

Grey Level Non-Uniformity Normalised

Grey Level Variance

High Grey Level Zone Emphasis

Large Area Emphasis

Large Area High Grey Level Emphasis

Large Area Low Grey Level Emphasis

Low Grey Level Zone Emphasis

Size Zone Non-Uniformity

Size Zone Non-Uniformity Normalised

Small Area Emphasis

Small Area High Grey Level Emphasis

Small Area Low Grey Level Emphasis

Zone Entropy

Zone Percentage

Zone Variance

GLDM features:

Dependence Entropy

Dependence Non-Uniformity

Dependence Non-Uniformity Normalised

Dependence Variance

Grey Level Non-Uniformity

Grey Level Variance

High Grey Level Emphasis

Large Dependence Emphasis

Large Dependence High Grey Level Emphasis

Large Dependence Low Grey Level Emphasis

Low Grey Level Emphasis

Small Dependence Emphasis

Small Dependence High Grey Level Emphasis

Small Dependence Low Grey Level Emphasis

Appendix B

Correlation matrices

The correlation matrices obtained before and after the removal of highly correlated features are presented below, regarding each one of the performed studies.

B.1 KRAS Mutation Status Prediction

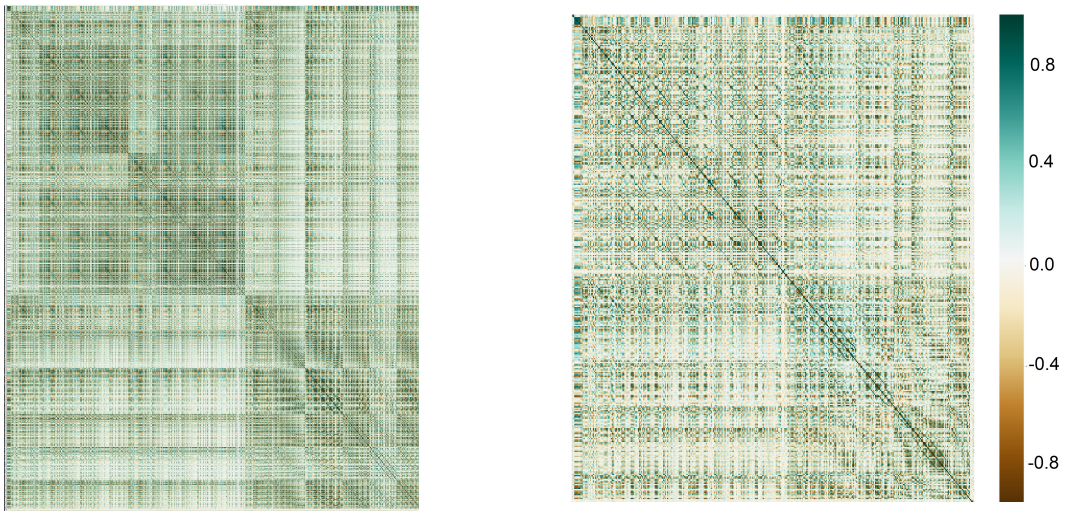


Figure B.1: Correlation matrices obtained before (left) and after (right) the removal of highly correlated features regarding KRAS mutation status prediction.

B.2 EGFR Mutation Status Prediction

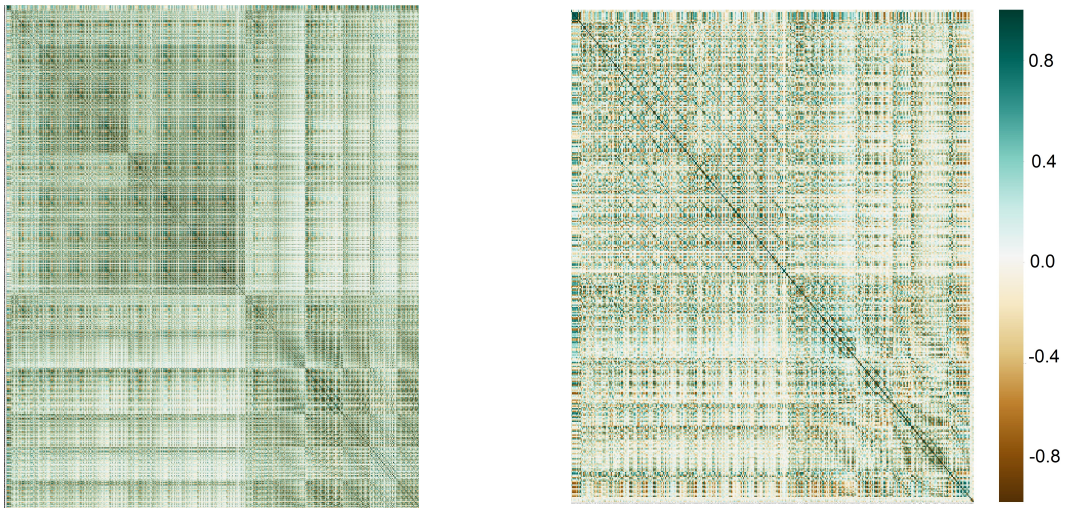


Figure B.2: Correlation matrices obtained before (left) and after (right) the removal of highly correlated features regarding EGFR mutation status prediction.

B.3 Emphysema Presence Prediction

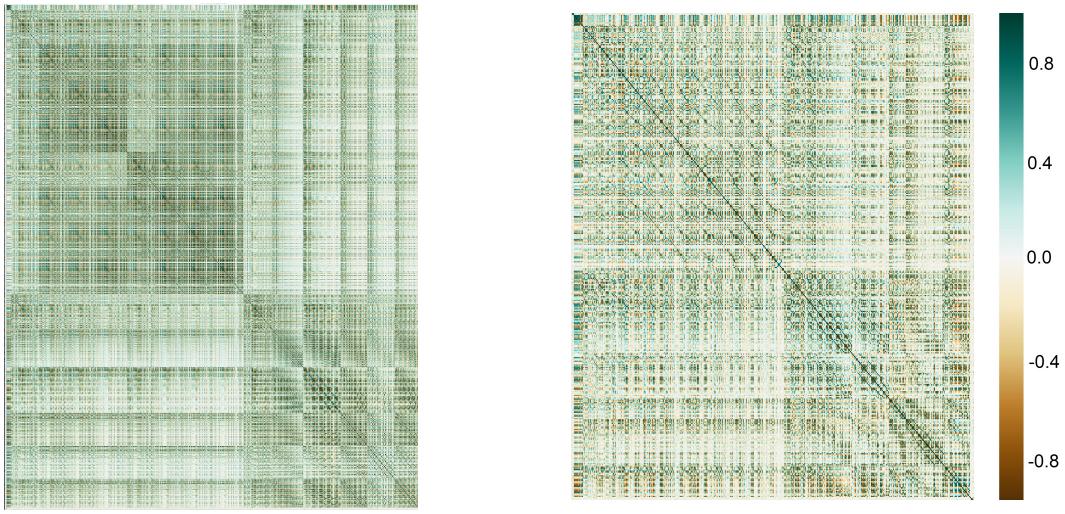


Figure B.3: Correlation matrices obtained before (left) and after (right) the removal of highly correlated features regarding emphysema presence prediction.

B.4 NSCLC Histological Subtype Prediction

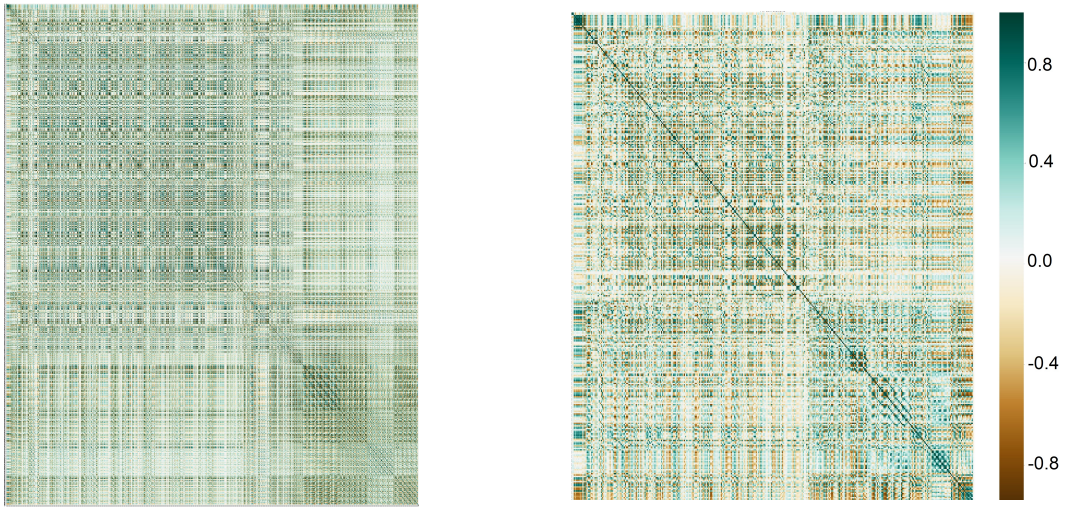


Figure B.4: Correlation matrices obtained before (left) and after (right) the removal of highly correlated features regarding histological subtype prediction.

Appendix C

Set of Features Used for Emphysema Prediction

Original (without filter):

Surface to Volume Ratio

Mean

GLCM Cluster Prominence

GLRLM Run Entropy

GLDM Dependence Entropy

GLDM Large Dependence Low Grey Level Emphasis

Wavelet filter:

LLH GLCM ClusterProminence

LLH GLDM Large Dependence High Grey Level Emphasis

LHL GLCM Cluster Shade

LHH Mean

LHH Median

LHH Skewness

HLL Energy

HLL Kurtosis

HLL GLSZM Size Zone Non-Uniformity Normalised

HLL GLDM Small Dependence Low Grey Level Emphasis

HLH Mean

HLH GLCM Cluster Shade

HLH GLCM IMC2

HHL Median

HHL Skewness

HHH Median

HHH GLCM Autocorrelation

LLL Minimum

LLL Range

LLL GLCM IDM

LLL GLRLM Long Run Low Grey Level Emphasis

LLL GLSZM Size Zone Non-Uniformity Normalised

LLL GLDM Large Dependence Low Grey Level Emphasis

LLL GLDM Small Dependence High Grey Level Emphasis

LoG filter:

Sigma 1.0 mm Skewness

Sigma 2.0 mm 90th Percentile

Sigma 2.0 mm Interquartile Range

Sigma 3.0 mm Minimum

Sigma 3.0 mm Range

Sigma 1.0 mm IMC1

Sigma 2.0 mm Inverse Variance

Sigma 3.0 mm IMC1

Sigma 4.0 mm Cluster Shade

Sigma 5.0 mm Cluster Shade

Sigma 3.0 mm Small Area Low Grey Level Emphasis

Sigma 1.0 mm Dependence Non-Uniformity Normalised

Sigma 3.0 mm Small Dependence Low Grey Level Emphasis

Sigma 4.0 mm Large Dependence Low Grey Level Emphasis

Sigma 5.0 mm Large Dependence Low Grey Level Emphasis

Clinical features:

Gender Female

Appendix D

Set of Features Used for NSCLC Histological Subtype Prediction

Original (without filter):

Elongation

10Percentile

GLSZM Grey Level Non-Uniformity Normalised

GLSZM High Grey Level Zone Emphasis

Wavelet filter:

LLH Entropy

LLH Skewness

LLH GLCM Cluster Shade

LLH GLCM IMC2

LLH GLSZM Size Zone Non-Uniformity Normalised

LHL GLCM IMC1

LHL GLSZM Size Zone Non-Uniformity Normalised

LHH Skewness

LHH GLCM IMC1

LHH GLRLM Long Run Emphasis

LHH GLDM Large Dependence High Grey Level Emphasis

HLL 90th Percentile

HLL Entropy

HLL Mean

HLL Skewness

HLL GLCM Cluster Shade

HLL GLRLM Short Run Low Grey Level Emphasis

HLL GLSZM Grey Level Non-Uniformity Normalised

HLL GLDM Dependence Variance

HLL GLDM Large Dependence Low Grey Level Emphasis

HLH GLCM Joint Energy

HLH GLSZM Large Area Emphasis

HLH GLDM Dependence Entropy

HLH GLDM Small Dependence High Grey Level Emphasis

HHL Median

HHL GLCM Cluster Prominence

HHL GLRLM Long Run Low Grey Level Emphasis

HHL GLSZM Large Area Emphasis

HHH Median

HHH GLSZM Zone Entropy

LLL GLDM Small Dependence High Grey Level Emphasis

LoG filter:

Sigma 1.0 mm 90th Percentile

Sigma 1.0 mm Mean

Sigma 1.0 mm Median

Sigma 2.0 mm 90th Percentile

Sigma 3.0 mm 90th Percentile

Sigma 1.0 mm Correlation

Sigma 2.0 mm Cluster Prominence

Sigma 2.0 mm Cluster Tendency

Sigma 3.0 mm Cluster Shade

Sigma 1.0 mm Size Zone Non-Uniformity Normalised

Sigma 1.0 mm Zone Entropy

Sigma 2.0 mm Size Zone Non-Uniformity Normalised

Sigma 3.0 mm High Grey Level Zone Emphasis

Sigma 3.0 mm Size Zone Non-Uniformity Normalised

Sigma 4.0 mm Grey Level Variance

Sigma 4.0 mm High Grey Level Zone Emphasis

Sigma 5.0 mm Small Area Emphasis

Sigma 3.0 mm Dependence Non-Uniformity Normalised

Clinical features:

Gender Female

References

[1] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386, 2015.

[2] Brambilla E. Burke A.P. Marx A. Nicholson A. G. Travis, W.D. *Pathology Genetics Tumours of the Lung, Pleura, Thymus and Heart*. IARC Press, 2003.

[3] Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90, 2011.

[4] Maryska LG Janssen-Heijnen and Jan-Willem W Coebergh. Trends in incidence and prognosis of the histological subtypes of lung cancer in north america, australia, new zealand and europe. *Lung cancer*, 31(2-3):123–137, 2001.

[5] Madeleine Scrivener, Evelyn EC de Jong, Janna E van Timmeren, Thierry Pieters, Benoît Ghaye, and Xavier Geets. Radiomics applied to lung cancer: a review. *Translational Cancer Research*, 5(4):398–409, 2016.

[6] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[7] American Cancer Society. Lung cancer, 2018. URL https://www.cancer.org/cancer/lung-cancer.html. [Online; accessed 01-November-2018].

[8] Geewon Lee, Ho Yun Lee, Hyunjin Park, Mark L Schiebler, Edwin JR van Beek, Yoshiharu Ohno, Joon Beom Seo, and Ann Leung. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. *European journal of radiology*, 86:297–307, 2017.

[9] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[10] Takayuki Kosaka, Yasushi Yatabe, Hideki Endoh, Kimihide Yoshida, Toyoaki Hida, Masahiro Tsuboi, Hirohito Tada, Hiroyuki Kuwano, and Tetsuya Mitsudomi. Analysis of epidermal growth factor receptor gene mutation in patients with non–small cell lung cancer and acquired resistance to gefitinib. *Clinical Cancer Research*, 12(19):5764–5769, 2006.

[11] Jeffrey A Engelman and Jeffrey Settleman. Acquired resistance to tyrosine kinase inhibitors during cancer therapy. *Current opinion in genetics & development*, 18(1):73–79, 2008.

[12] Alexa B Turke, Kreshnik Zejnullahu, Yi-Long Wu, Youngchul Song, Dora Dias-Santagata, Eugene Lifshits, Luca Toschi, Andrew Rogers, Tony Mok, Lecia Sequist, et al. Preexistence and clonal selection of met amplification in egfr mutant nsclc. *Cancer cell*, 17(1):77–88, 2010.

[13] Hugo JWL Aerts. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA oncology*, 2(12):1636–1642, 2016.

[14] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*, 5:180202, 2018.

[15] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, D Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[16] Maria Werner-Wasik, Ying Xiao, Edward Pequignot, Walter J Curran, and Walter Hauck. Assessment of lung cancer response after nonoperative therapy: tumor diameter, bidimensional product, and volume. a serial ct scan-based study. *International Journal of Radiation Oncology* Biology* Physics*, 51(1):56–61, 2001.

[17] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David Fenstermacher, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012.

[18] Ella A Kazerooni and Barry H Gross. *Cardiopulmonary imaging*, volume 4. Lippincott Williams & Wilkins, 2004.

[19] Thorsten M Buzug. Computed tomography. In *Springer Handbook of Medical Technology*, pages 311–342. Springer, 2011.

[20] Aliza Medical Imaging & DICOM Viewer, 2018. URL https://sites.google.com/site/alizaviewer. [Online; accessed 27-December-2018].

[21] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Weiruo Zhang, Ann Leung, Michael Kadoch, Joseph Shrager, Andrew Quon, Daniel Rubin, Sylvia Plevritis, and Sandy Napel. Data for NSCLC Radiogenomics Collection, 2017. URL https://wiki.cancerimagingarchive.net/x/W4G1AQ. type: dataset. [Online; accessed 12-December-2018].

[22] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, Dec 2013. ISSN 1618-727X. doi: 10.1007/s10278-013-9622-7.

[23] Ying Liu, Jongphil Kim, Yoganand Balagurunathan, Qian Li, Alberto L Garcia, Olya Stringfield, Zhaoxiang Ye, and Robert J Gillies. Radiomic features are associated with egfr mutation status in lung adenocarcinomas. *Clinical lung cancer*, 17(5):441–448, 2016.

[24] João Moreira, André Carlos Ponce de Leon Ferreira, and Tomáš Horváth. *A General Introduction to Data Analytics*. Wiley Online Library, 2018.

[25] Pratap Dangeti. *Statistics for Machine Learning: Techniques for Exploring Supervised, Unsupervised, and Reinforcement Learning Models with Python and R*. Packt Publishing, 2017. ISBN 1788295757, 9781788295758.

[26] Ryan Wilson and Anand Devaraj. Radiomics of pulmonary nodules and lung cancer. *Translational lung cancer research*, 6(1):86, 2017.

[27] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, et al. Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*, 2016.

[28] Kyongtae T Bae, Pornpim Fuangtharnthip, Srinivasa R Prasad, Bonnie N Joe, and Jay P Heiken. Adrenal masses: Ct characterization with histogram analysis method. *Radiology*, 228(3):735–742, 2003.

[29] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[30] Gillian Ellison, Guanshan Zhu, Alexandros Moulis, Simon Dearden, Georgina Speake, and Rose McCormack. Egfr mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples. *Journal of clinical pathology*, 66(2):79–89, 2013.

[31] Markus D Siegelin and Alain C Borczuk. Epidermal growth factor receptor mutations in lung adenocarcinoma. *Laboratory investigation*, 94(2):129, 2014.

[32] Gregory J Riely, Jenifer Marks, and William Pao. Kras mutations in non–small cell lung cancer. *Proceedings of the American Thoracic Society*, 6(2):201–205, 2009.

[33] Brandon Golding, Anita Luu, Robert Jones, and Alicia M Viloria-Petit. The function and therapeutic targeting of anaplastic lymphoma kinase (alk) in non-small cell lung cancer (nsclc). *Molecular cancer*, 17(1):52, 2018.

[34] Dongdong Mei, Yan Luo, Yan Wang, and Jingshan Gong. Ct texture analysis of lung adenocarcinoma: can radiomic features be surrogate biomarkers for egfr mutation statuses. *Cancer Imaging*, 18(1):52, 2018.

[35] Efe Ozkan, Anna West, Jeffrey A Dedelow, Benjamin F Chu, Weiqiang Zhao, Vedat O Yildiz, Gregory A Otterson, Konstantin Shilo, Subha Ghosh, Mark King, et al. Ct gray-level texture analysis as a quantitative imaging biomarker of epidermal growth factor receptor mutation status in adenocarcinoma of the lung. *American Journal of Roentgenology*, 205(5):1016–1025, 2015.

[36] Subba R Digumarthy, Atul M Padole, Roberto Lo Gullo, Lecia V Sequist, and Mannudeep K Kalra. Can ct radiomic analysis in nsclc predict histology and egfr mutation status? *Medicine*, 98(1), 2019.

[37] Shuo Wang, Jingyun Shi, Zhaoxiang Ye, Di Dong, Dongdong Yu, Mu Zhou, Ying Liu, Olivier Gevaert, Kun Wang, Yongbei Zhu, et al. Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning. *European Respiratory Journal*, 53(3), 2019.

[38] Xiao-Yang Li, Jun-Feng Xiong, Tian-Ying Jia, Tian-Le Shen, Run-Ping Hou, Jun Zhao, and Xiao-Long Fu. Detection of epithelial growth factor receptor (egfr) mutations on ct images of patients with lung adenocarcinoma using radiomics and/or multi-level residual convolutionary neural networks. *Journal of thoracic disease*, 10(12):6624, 2018.

[39] Olivier Gevaert, Sebastian Echegaray, Amanda Khuong, Chuong D Hoang, Joseph B Shrager, Kirstin C Jensen, Gerald J Berry, H Henry Guo, Charles Lau, Sylvia K Plevritis, et al. Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific reports*, 7:41674, 2017.

[40] Glen J Weiss, Balaji Ganeshan, Kenneth A Miles, David H Campbell, Philip Y Cheung, Samuel Frank, and Ronald L Korn. Noninvasive image texture analysis differentiates k-ras mutation from pan-wildtype nsclc and is prognostic. *PloS one*, 9(7):e100244, 2014.

[41] Anita Midha, Simon Dearden, and Rose McCormack. Egfr mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutmapii). *American journal of cancer research*, 5(9):2892, 2015.

[42] Gregory J Riely, William Pao, DuyKhanh Pham, Allan R Li, Naiyer Rizvi, Ennapadam S Venkatraman, Maureen F Zakowski, Mark G Kris, Marc Ladanyi, and Vincent A Miller. Clinical course of patients with non–small cell lung cancer and epidermal growth factor receptor exon 19 and exon 21 mutations treated with gefitinib or erlotinib. *Clinical cancer research*, 12(3):839–844, 2006.

[43] Makoto Maemondo, Akira Inoue, Kunihiko Kobayashi, Shunichi Sugawara, Satoshi Oizumi, Hiroshi Isobe, Akihiko Gemma, Masao Harada, Hirohisa Yoshizawa, Ichiro Kinoshita, et al. Gefitinib or chemotherapy for non–small-cell lung cancer with mutated egfr. *New England Journal of Medicine*, 362(25):2380–2388, 2010.

[44] Tony S Mok, Yi-Long Wu, Sumitra Thongprasert, Chih-Hsin Yang, Da-Tong Chu, Nagahiro Saijo, Patrapim Sunpaweravong, Baohui Han, Benjamin Margono, Yukito Ichinose, et al. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine*, 361(10):947–957, 2009.

[45] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[46] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[47] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[50] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

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

[52] Marta Román, Iosune Baraibar, Inés López, Ernest Nadal, Christian Rolfo, Silvestre Vicent, and Ignacio Gil-Bazo. Kras oncogene in non-small cell lung cancer: clinical perspectives on the treatment of an old target. *Molecular cancer*, 17(1):33, 2018.

[53] Megan L Hames, Heidi Chen, Wade Iams, Jonathan Aston, Christine M Lovly, and Leora Horn. Correlation between kras mutation status and response to chemotherapy in patients with advanced non-small cell lung cancer. *Lung Cancer*, 92:29–34, 2016.

[54] Hongxuan Zhou, Yun Dai, Liqun Zhu, Chun Wang, Xiaodong Fei, Qin Pan, Juxiang Chen, Xianqing Shi, Yanfeng Yang, Xiaoxing Tao, et al. Poor response to platinum-based chemotherapy is associated with kras mutation and concomitant low expression of brac1 and tyms in nsclc. *Journal of International Medical Research*, 44(1):89–98, 2016.

[55] Christian Manegold. Treatment algorithm in 2014 for advanced non-small cell lung cancer: therapy selection by tumour histology and molecular biology. *Advances in medical sciences*, 59(2):308–313, 2014.

[56] Chintan Parmar, Ralph TH Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek Rietveld, Michelle M Rietbergen, Benjamin Haibe-Kains, Philippe Lambin, and Hugo JWL Aerts. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific reports*, 5:11044, 2015.

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

[58] Satrajit Basu, Lawrence O Hall, Dmitry B Goldgof, Yuhua Gu, Virendra Kumar, Jung Choi, Robert J Gillies, and Robert A Gatenby. Developing a classifier model for lung tumors in ct-scan images. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1306–1312. IEEE, 2011.

[59] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

[60] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[61] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 2016.

[62] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier, 1992.

[63] Comprehensive Cancer Information, 1980. URL https://www.cancer.gov/. [Online; accessed 30-November-2018].

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

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

[66] Samuel H Hawkins, John N Korecki, Yoganand Balagurunathan, Yuhua Gu, Virendra Kumar, Satrajit Basu, Lawrence O Hall, Dmitry B Goldgof, Robert A Gatenby, and Robert J Gillies. Predicting outcomes of nonsmall cell lung cancer using ct image features. *IEEE access*, 2: 1418–1426, 2014.

[67] Thibaud P Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Gretchen Hermann, Philippe Lambin, Benjamin Haibe-Kains, Raymond H Mak, and Hugo JWL Aerts. Ct-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350, 2015.

[68] Jason T Rich, J Gail Neely, Randal C Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W Wang. A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3):331–336, 2010.

[69] David Roxbee Cox. *Analysis of survival data*. Routledge, 2018.

[70] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[71] Marie Kraska-Miller. *Nonparametric statistics for social and behavioral sciences*. CRC Press, 2013.

[72] Adam R Brentnall and Jack Cuzick. Use of the concordance index for predictors of censored survival data. *Statistical methods in medical research*, 27(8):2359–2373, 2018.

[73] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Data From NSCLC-Radiomics-Genomics, 2015. URL https://wiki.cancerimagingarchive.net/x/GAL1. type: dataset. [Online; accessed 12-December-2018].

[74] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Data From NSCLC-Radiomics, 2015. URL https://wiki.cancerimagingarchive.net/x/FgL1. type: dataset. [Online; accessed 12-December-2018].

[75] Peter Mildenberger, Marco Eichelberg, and Eric Martin. Introduction to the dicom standard. *European radiology*, 12(4):920–927, 2002.

[76] Charles E Kahn, John A Carrino, Michael J Flynn, Donald J Peck, and Steven C Horii. Dicom and radiology: past, present, and future. *Journal of the American College of Radiology*, 4 (9):652–657, 2007.

[77] David M Hansell, Alexander A Bankier, Heber MacMahon, Theresa C McLoud, Nestor L Muller, and Jacques Remy. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722, 2008.

[78] Subrahmanyam Gorthi, CM Bach, and Jean-Philippe Thiran. Exporting contours to dicom-rt structure set. *Insight Journal*, (1):1–18, 2009.

[79] Contour Data Attribute – DICOM Standard Browser, . URL https://dicom.innolitics.com/ciods/rt-dose/roi-contour/30060039/30060040/30060050. [Online; accessed 03-June-2019].

[80] Anil Kalra. Developing fe human models from medical images. In *Basic Finite Element Method as Applied to Injury Biomechanics*, pages 389–415. Elsevier, 2018.

[81] Stephan A Bolliger, Lars Oesterhelweg, Danny Spendlove, Steffen Ross, and Michael J Thali. Is differentiation of frequently encountered foreign bodies in corpses possible by hounsfield density measurement? *Journal of forensic sciences*, 54(5):1119–1122, 2009.

[82] Ales Prochazka, Lucie Grafova, Oldrich Vysata, and Neurocenter Caregroup. Three-dimensional wavelet transform in multi-dimensional biomedical volume processing. In *Proc. of the IASTED International Conference on Graphics and Virtual Reality, Cambridge*, pages 263–268, 2011.

[83] Sergei V Fotin, Anthony P Reeves, Alberto M Biancardi, David F Yankelevitz, and Claudia I Henschke. A multiscale laplacian of gaussian filtering approach to automated pulmonary nodule detection from whole-lung low-dose ct scans. In *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260, page 72601Q. International Society for Optics and Photonics, 2009.

[84] Eirini Papadopoulou, Nikolaos Tsoulos, Angeliki Tsirigoti, Angela Apessos, Konstantinos Agiannitopoulos, Vasiliki Metaxa-Mariatou, Konstantinos Zarogoulidis, Pavlos Zarogoulidis, Dimitrios Kasarakis, Stylianos Kakolyris, et al. Determination of egfr and kras mutational status in greek non-small-cell lung cancer patients. *Oncology letters*, 10(4):2176–2184, 2015.

[85] Anna M Varghese, Camelia S Sima, Jamie E Chaft, Melissa L Johnson, Gregory J Riely, Marc Ladanyi, and Mark G Kris. Lungs don't forget: comparison of the kras and egfr mutation profile and survival of collegiate smokers and never smokers with advanced lung cancers. *Journal of Thoracic Oncology*, 8(1):123–125, 2013.

[86] Snjezana Dogan, Ronglai Shen, Daphne C Ang, Melissa L Johnson, Sandra P D'Angelo, Paul K Paik, Edyta B Brzostowski, Gregory J Riely, Mark G Kris, Maureen F Zakowski, et al. Molecular epidemiology of egfr and kras mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related kras-mutant cancers. *Clinical cancer research*, 18(22):6169–6177, 2012.

[87] Andreas C Müller, Sarah Guido, et al. *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.", 2016.

[88] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[89] Lior Rokach and Oded Maimon. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005.

[90] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[91] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[92] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[93] Welcome to LightGBM's documentation! — LightGBM 2.2.4 documentation, . URL https://lightgbm.readthedocs.io/en/latest/index.html. [Online; accessed 13-May-2019].

[94] Lara Lusa et al. Evaluation of smote for high-dimensional class-imbalanced microarray data. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 89–94. IEEE, 2012.