FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Ensemble Methods for Lung Cancer Gene Mutation Prediction

Alexandra Costa Ventura

FOR JURY EVALUATION

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Tânia Pereira

Second Supervisor: António Cunha

Third Supervisor: Cláudia Freitas

Fourth Supervisor: Hélder P. Oliveira

July 30, 2021

Resumo

O cancro do pulmão apresenta uma elevada taxa de mortalidade. Um diagnóstico precoce e a escolha do tratamento mais adequado é extremamente importante para inverter esta tendência. No caso específico da terapia direcionada, uma genotipagem eficaz do tumor é fundamental já que este tipo de tratamento utiliza fármacos capazes de induzir a morte nas células cancerígenas. Biópsias são uma das formas de obter a informação relativa ao genoma do tumor, no entanto este método é extremamente invasivo e muitas vezes doloroso.

Imagens médicas são uma potencial alternativa às biópsias. Algumas características associadas a estas imagens mostraram-se capazes de identificar alterações genómicas dentro do ADN do tumor, um conceito denominado de radiogenómica. Além disso, os métodos *ensemble* mostram um grande potencial na superação de algumas barreiras dos modelos únicos usados anteriormente na previsão da mutação genética.

Esta dissertação tem por objetivo oferecer novos avanços no campo da radiogenomia estudando o uso de aprendizagem *ensemble* para prever o estado do gene mutado no cancro do pulmão. O modelo desenvolvido deverá ser capaz de identificar com maior exatidão o estado da mutação EGFR (*Epidermal Growth Factor Receptor*) no *dataset* utilizado.

O melhor resultado obtido corresponde ao modelo de previsão *ensemble* que utiliza como *baselearners*: *logistic regression (LR), linear support vector machine (SVM) and elastic net (EN)*; e uma rede neuronal como regra de combinação. Contudo este modelo que resultou numa AUC de $0.708(\pm 0.124)$ não superou os resultados obtidos pelos modelos quando usados sozinhos onde foi possivel obter AUCs de 0.712 (± 0.119) para LR, 0.711 (± 0.119) para SVM e 0.712 (± 0.120) para EN. Uma razão possível que explica este comportamento é a classificação incorreta dos mesmos exemplos pelos métodos o que resulta na indução destes mesmos erros na rede neuronal usada para o *ensemble*.

Abstract

Lung cancer has a huge mortality rate. A early diagnosis and the choice of the most appropriate treatment is extremely important to invert it. The particular case of target therapy, effective genotyping of the tumour is fundamental since this treatment use targeted drugs that can induce death in cancer cells. One way to get this information is trough biopsy but this method is extremely invasive and often painful.

Medical imaging is a potential alternative to biopsies. Some image features have shown to identify genomic alterations within tumour DNA, a field that is now called radiogenomics. Also, ensemble methods shows a big potential to overcome some barriers of the single models used previously in gene mutation prediction.

This dissertation aims to provide further advances in the radiogenomics field by studying the use of ensemble learning to predict the gene mutation status in lung cancer. The model developed intends to predict EGFR (Epidermal Growth Factor Receptor) mutation status in the used dataset.

The best result obtained correspond to the prediction ensemble model that use three baselearners: logistic regression (LR), linear support vector machine (SVM) and elastic net (EN); and a neural network (NN) as the combination rule. However, this model which results in an AUC of $0.708(\pm 0.124)$ did not outperform the single models of the first state with AUCs of $0.712(\pm 0.119)$ to LR, $0.711(\pm 0.119)$ to SVM and $0.712(\pm 0.120)$ to EN. A possible reason that could explain this behaviour is that the methods misclassified the same examples causing the NN used for ensemble to also be induced to make this same mistake.

Acknowledges

Em primeiro lugar um agradecimento especial aos meus orientadores que foram sempre incansáveis, Tânia Pereira, António Cunha, Hélder P. Oliveira, Cláudia Freitas pela dedicação e paciência. Um obrigado também ao Francisco Silva e à Joana Morgado pela disponibilidade e ajuda durante todo este projeto.

Ao meu irmão Lucas, que emprestou o nome ao projeto e que a mim me deu esta sorte grande de o poder ter como companheiro para a vida toda.

Aos meus pais, a quem devo tudo e que nunca falharam com apoio e carinho para chegar onde quisesse.

Aos meus amigos e restante família que sempre estiveram presentes e pelos quais serei sempre grata.

Por fim, uma palavra de apreço especial aos 16 de 2016 que acompanharam mais perto que ninguém esta (a)ventura. Aprendemos juntos aquilo que nenhum mestrado nos podia ensinar e vivemos aquilo a que nenhum texto escrito algum dia poderá fazer jus.

A todos os outros que não contam destas menções, mas que ao longo destes últimos cinco anos cruzaram este meu caminho, um grande bem-haja por contribuírem de uma forma ou de outra para a história bonita que se escreveu.

Alexandra Costa Ventura

"Eu vou colar esses caquinhos
E fazer um Gaudí!"
- Capicua

Contents

List of Figures

List of Tables

Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
AI	Artificial Inteligence
ALK	Anaplastic Lymphoma Kinase
CT	Computed Tomography
CV	Cross-Validation
CNNs	Convolutional Neural Networks
DLR	Deep Learning Radiomics
DT	Decision Tree
EN	Elastic Network
GBDT	Gradient Boosting Decision Trees
EGFR	Epidermal Growth Factor Receptor
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
KNN	k-Nearest-Neighbour
KRAS	Kristen Rat Sarcoma Viral Oncogene Homolog
LASSO	Least Absolute Shrinkage and Selection Operator
LoG	Laplacian-of-Gaussian
LR	Logistic Regression
ML	Machine Learning
MR	Magnetic Resonance Imaging
MTV	Metabolic Tumour Volume
MV	Majority Voting
NGTDM	Neighbouring Gray Tone Difference Matrix
NN	Neural Network
NSCLC	Non-Small Cell Lung Cancer
PC	Principal Component
PCA	Principal Component Analysis
PET	Positron Emission Tomography
RF	Random Forest
RFE	Recursive Feature Elimination
ROI	Region of Interest
SCLC	Small Cell Lung Cancer
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Over-Sampling Technique
SUV	Standard Uptake Value
SVM	Support Vector Machine
TKI	Tyrosine Kinase Inhibitors
VOI	Volume of Interest

Chapter 1

Introduction

Lung cancer is the second most common cancer losing for breast cancer in women and prostate cancer in men. The chance of developing this cancer is about 1 in 15 for a man and about 1 in 17 for a woman including both smokers and non-smokers. It mainly occurs in older people, most people diagnosed are 65 or older with the average being about 70 years old[1].

The main cause pointed out to develop lung cancer is smoking tobacco; however, there is a minority of non-smokers who are diagnosed. Some of the causes that are pointed out, in this case, are exposure to certain chemicals such as radon, secondhand smoke, air pollution, among others[1].

It is possible to distinguish three main types of lung cancer: non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) (also called oat cell cancer) and lung carcinoid tumour. About 80% to 85% of all lung cancers are NSCLC and 10% to 15% are SCLC. Carcinoid tumours account for fewer than 5% of lung tumours. There are also other types of lung cancer like adenoid cystic carcinomas, lymphomas, and sarcomas but they are rare[1].

Lung cancer is the leading cause of cancer death, making up almost 25% of all cancer deaths. The advances in early detection have been helping decrease this number as there is a greater chance to find a successful treatment[1].

1.1 Motivation

In addition to the early detection of cancer, choosing the proper treatment is another significant factor to improve survival chances. There are many treatments for lung cancer, some better known like surgery, chemotherapy, radiation therapy, but there are also other options like targeted therapy, immunotherapy, stem cell or bone marrow transplant, hormone therapy, among others[1].

Knowing the tumour including its morphology, tissue type, genomic characteristics, etc. are important factors in this decision. In the particular case of target therapy, effective genotyping of the tumour is fundamental since this treatment use targeted drugs that can induce death in cancer cells. This treatment has high efficacy and fewer side effects than chemotherapy but this is only true for some mutated genes [1, 2].

[1] https://www.cancer.org/cancer/lung-cancer.html

The conventional way to collect information about tumour genotyping is by biopsies; however this is an extremely invasive and often painful and dangerous process that may even lead to the death of the patient. Furthermore, most tumours are not homogeneous, but rather composed of multiple clonal subpopulations of cancer cells. So, many times, the tissue obtained is not representative and the biopsy has to be repeated [3].

A promising alternative to the problems presented is to use medical images. Medical images have the advantages of being non-invasive, three dimensional and provide information regarding the entire tumour plus it is possible to extract from them a large number of quantitative features that allow us to build predictive models, linking image features to the genomic profiles of the tumour [3]. This ability to identify the presence of specific mutations from imaging features is called radiogenomics [4].

The predictive models used in radiogenomics are most of the times based on machine learning (ML) algorithms.

Lately, ensemble learning paradigms have emerged and proved to generate better results than single ML classifiers. Ensemble learning relies on the intuitive principle of combining several predictions to obtain a more accurate final result analogous to the way humans gather diverse opinions and combine them to make complex decisions [5, 6].

1.2 Objectives

This dissertation aims to provide further advances in the radiogenomics field by studying the use of ensemble methods combined with some machine learning algorithms to predict the gene mutation status in lung cancer. The main objective is to develop a model, that takes advantage of ensemble learning, that must identify more accurately EGFR (Epidermal Growth Factor Receptor) mutation status in the used dataset.

1.3 Contributions

This dissertation presents the following contributions:

- Development of predictive models of EGFR mutation status in lung cancer patients using CT images.
- An investigation about the impact of using ensemble methods in the prediction of EGFR mutation status using different combinations of base learners and aggregation rules. analisys

1.4 Document Structure

This document is divided six chapters. The present chapter introduces the theme and explain the motivation and objectives of this dissertation. Chapter 2 present some concepts needed to better understand the problem presented, including a more deep description of the lung cancer, precision

medicine and mutated genes, biopsy, medical image and radiogenomics. Chapter 3 covers some relevant studies done about using image features, clinical data and sometimes even some extra tumoural diseases and characteristics to predict gene mutation status using medical images and is briefly described some basic concepts about ensemble learning and the advantages of using it. In Chapter 4 is presented the implementation for the mutant prediction from choosing the dataset, processing the data and implementing the model itself. The Chapter 5 correspond to the analysis and discussion of the results obtained. Finally, chapter 6 comprises the conclusions reached with this dissertation.

Chapter 2

Problem Contextualisation: Lung Cancer

This chapter aims to provide a brief contextualisation of the problem presented. Firstly a description of the various types and subtypes of lung cancer is given in Section 2.1, then there will be a short overview of precision medicine and mutated genes (Section 2.2). Two diagnostic methods will be analysed, biopsies and medical imaging (Sections 2.3 and 2.4 respectively), their advantages and disadvantages. Finally, it will be introduced the topics of radiomics and radiogenomics in Section 2.5 as part of the solution of using medical imaging and its features (analysed in more detail in Subsection 2.5.1) to trace the genetic profiles of tumours.

2.1 Lung Cancer

Lung cancer is a genetic disease that starts when cells in the body begin to grow out of control due to the accumulation of multiple genetic mutations and epigenetic alterations.

According to information taken from the American Cancer Society website[1], some types and subtypes of lung tumours can be distinguished [7]:

- **Non-small cell lung cancer (NSCLC):** that includes three subtypes that are group together because of their similar treatment and prognoses.
  - **Adenocarcinomas:** starts in mucus-secreting cells and tend to be found in the outer parts of the lung. It is the most common type of lung cancer in non-smokers but it still occurs mainly in current and former smokers. It affects more women than men and is more common in younger people than other types of cancer. Adenocarcinoma is typically found before it has spread.
  - **Squamous cell carcinoma:** as the name suggests start in squamous cells and is usually found in the central part of the lungs. It is often linked to a history of smoking.

[1] https://www.cancer.org/cancer/lung-cancer.html

  - **Large cell (undifferentiated) carcinoma:** can appear in any part of the lung and tends to grow and spread quickly.

- **Small cell lung cancer (SCLC):** or oat cell cancer tends to grow and spread faster than NSCLC and for this reason is typically already widespread when diagnosed. Tends to respond well to chemotherapy and radiation therapy.

- **Lung carcinoid tumours:** represents less than 5% of lung tumours and usually grows slowly.

- **Other lung tumours:** adenoid cystic carcinomas, lymphomas, and sarcomas, as well as benign lung tumours such as hamartomas. All of these tumours are rare and treated differently from the more common lung cancers.

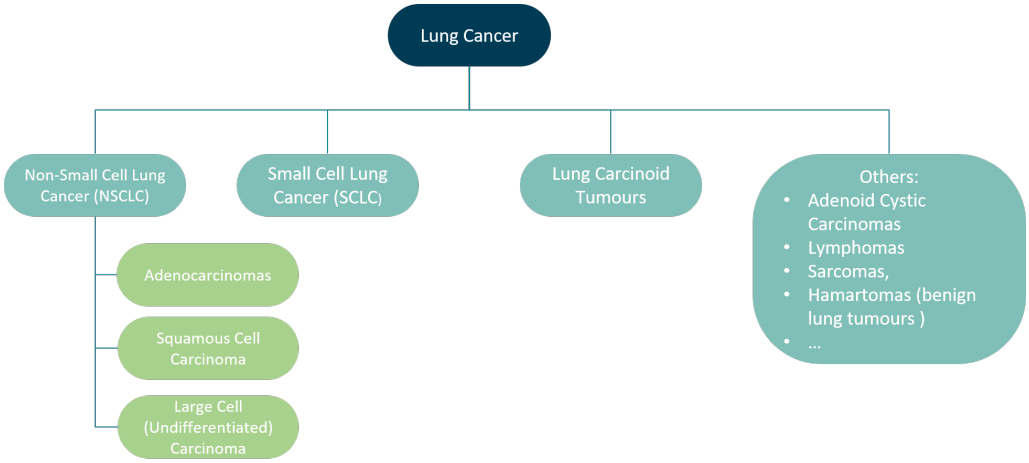


Figure 2.1: Scheme representing the different types and subtypes of Lung Cancer.

There is also the possibility of some cancers that start in other organs spread to the lungs, but these are not considered lung cancers. The cancer treatment is based on where it started (the primary cancer site).

2.2 Precision Medicine and Mutated Genes

Given the high lethally of lung cancer, the need to find more treatments and more effective has arisen. The option of using genetic profiling to direct the treatment emerged and precision medicine using mutation-targeting strategies has shown increasingly successful [8].

Epidermal Growth Factor Receptor (EGFR), Kristen Rat Sarcoma Viral Oncogene Homolog (KRAS) and Anaplastic Lymphoma Kinase (ALK) are the most frequently mutated gene in lung cancer [9]. One of the best known target therapy is using tyrosine kinase inhibitors (TKIs) like gefitinib, erlotinib, and afatinib that prevent cell survival and uncontrolled proliferation processes,

caused by mutated EGFR activation [8]. Crizotinib was the first drug approved for NSCLC harbouring ALK rearrangements[9]. But there are some genes, like KRAS, that were considered non-druggable targets for their poor response to this type of treatment [10].

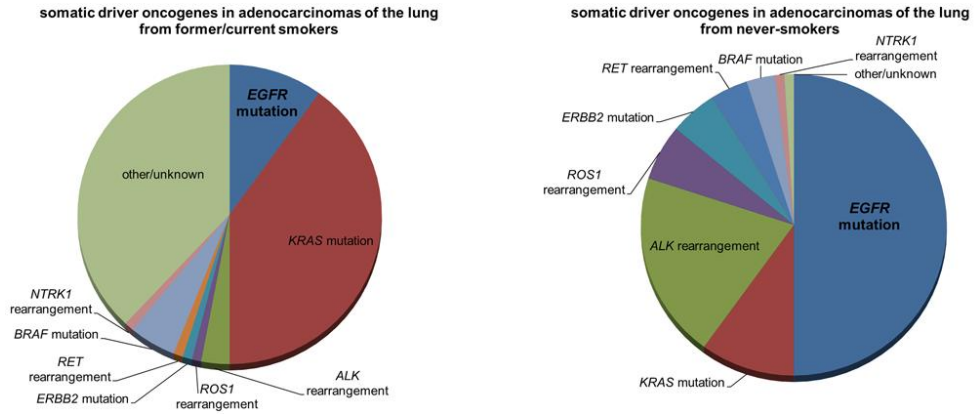


Figure 2.2: Pie charts showing the distribution of driver oncogene mutations in lung adenocarcinomas from former/current smokers (left chart) and from never-smokers (right chart). From Jorge et al. [9].

Despite the positive results target therapy still have some barriers. Kill cancer cells can represent initial success prolonging patient survival for several months but there are still challenges like drug resistance [1, 8].

2.3 Biopsy

Traditionally the way of characterising the tumour genotype is to collect a sample tissue doing a biopsy.

Biopsies are known to be very intrusive and often associated with pain and discomfort for the patient. In addition, they can be very complicated processes due to the location of the tumour, which can make the collection of tissue very difficult or even impossible. Besides, even if the tissue collection is successful the amount may not be sufficient or quite representative because of the heterogeneity associated with the tumour and the evolution of the disease. This heterogeneity in the tumour is associated with the different degrees of proliferation and/or differentiation of the cells, which form minor subclones of malignant cells.

Sequential or multiple biopsies are not a solution to achieve representative results for all the above reasons added to the time-consuming and financial costs [10, 11].

2.4 Medical Images

Medical imaging has been fundamental not only in diagnosis but also in staging, treatment planning, postoperative surveillance, and response evaluation in the routine management of lung cancer [7]. Medical images have the advantages of being non-invasive, three dimensional and provide

information regarding the entire tumour [3].

Computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI) are some of medical images used [7]. CT is the one that stands out since CT is currently the primary mean for screening and monitoring lung cancer. Screening scans typically use low dose CT images, while diagnostic scans are more often high quality and with contrast enhancement [4].

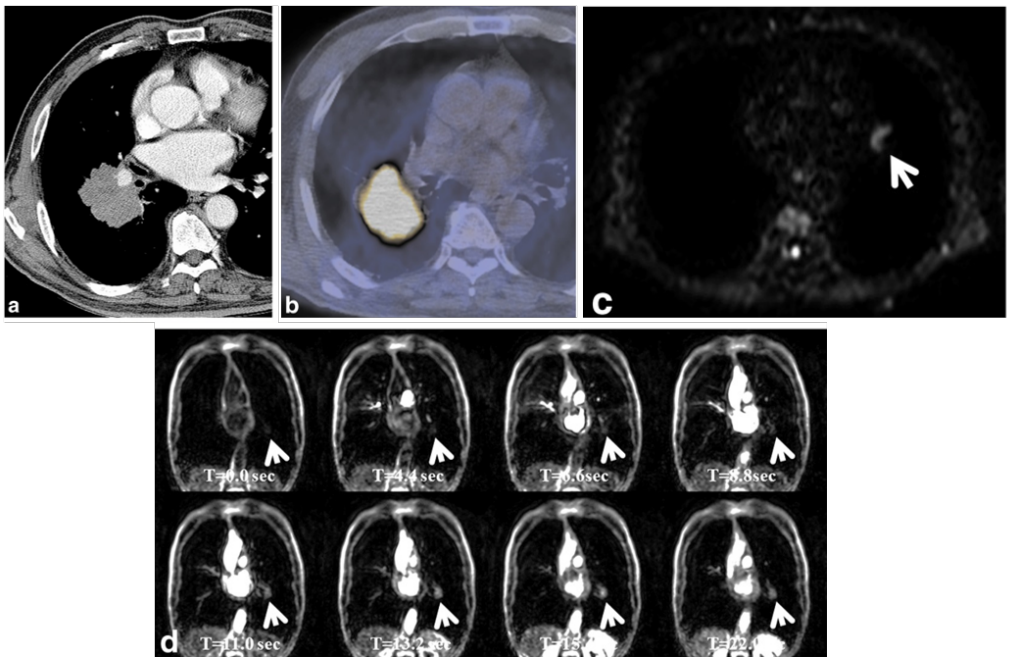


Figure 2.3: Examples of medical images a) and b) from a 68-year-old man with N2 squamous cell carcinoma of the lung and c) and d) from a 75-year-old man with lung adenocarcinoma. a), b) Medistinal window of CT (a) and PET (b) that show a 5.8-cm size mass in superior segment of right lower lobe with high glucose uptake c) Diffusion-weighted MR image manifests the nodule (arrow) as high-signal intensity lesion, suggesting a malignant one. d) Dynamic MRI with ultrafast-gradient-echo technique shows a well-enhancing nodule (arrow) in the left lower lung zone, and heterogeneous enhancement in both lungs due to pulmonary emphysema. This nodule was assessed as a malignant lesion on dynamic contrast-enhanced MRI. From Kim et al. [12].

The need to find an alternative to biopsies associated with the advantages that medical imaging presents led to the construction of prediction models, linking the genomic profiles of the tumour to the features of the images since it is possible to extract a large number of quantitative features from these images [3].

2.5 Radiomics and Radiogenomics

Finding a correlation between CT imaging and relevant gene expression signatures, such as EGFR, ALK, and KRAS, could help redefine existing staging and diagnostic paradigms and provide a better solution for some of the problems already mentioned above [10].

Radiomics is a field of study that extracts quantitative image features from medical images. Radiogenomics is the ability of radiomics to estimate the presence or absence of clinically relevant mutations and although many studies have found correlations between mutations and radiomic features, the results have not always been consistent [4].

The typical workflow of radiomics or radiogenomics is represented in Figure 2.4 and includes the acquisition and reconstruction of the images, the segmentation of the region of interest (ROI), extraction and quantification of the features and finally, the construction of predictive and prognostic models [4].

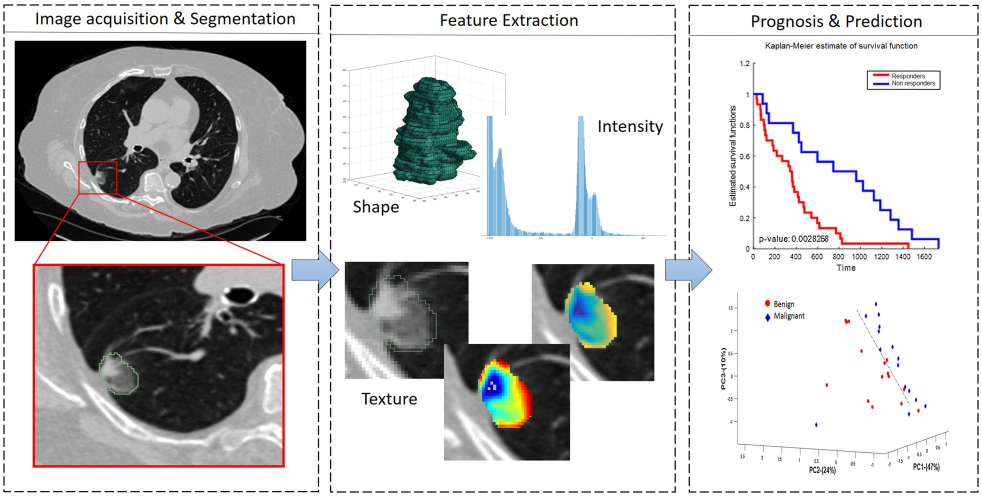


Figure 2.4: Overview of a typical radiomic workflow. From Thawani et al. [4].

After the first step that includes the acquisition of high quality and standardised imaging, then is necessary to do the segmentation of the ROI that could include only the tumour but also some surrounding tissue [13]. Previous results from the project "Lung Cancer Screening - A non-invasive methodology for early diagnosis" [1] in which this dissertation is integrated and literature suggest that the most relevant information to predict the mutation status in lung cancer might be the combination of features from the nodule and other lung structures [11, 14, 15, 16]. Also, the segmentation can be done automatically using segmentation algorithms or manually by an experienced radiologist. A combination of both methods can also be used in a semi-automatic process [13].

The next step includes not only the extraction of features (presented in more detail in section 2.5.1) but also a process of selection. Often due to the high number of features that can be obtained, it is necessary to use appropriate selection strategies to reduce the dimensionality of the problem and improve the prediction accuracy. This may help enhance the model's generalisation ability and speed up the learning process. The association and redundancy of features are two characteristics to take into account for the selection of the features to reduce the dimensionality [4, 13].

[1] https://www.inesctec.pt/pt/projetos/lucas#intro

The ultimate goal is to provide the result of the image feature selection processes to models for predicting and classifying the state of the mutated gene and evaluating their performance [13].

2.5.1 Feature Categories

The radiomic features are extracted automatically and with high throughput and could be divided into the following classes [4]:

- Histogram/Intensity-based features
- Shape/Structural features
- Texture/Gradient-based features

Intensity-based features result from graphic representations of the intensity distribution of the image - histograms. From the histograms is possible to obtain some statistical measures like the mean, median, standard deviation, kurtosis, skewness, energy, entropy, uniformity and variance of the lung nodules that can be used in a machine learning framework for mutant gene prediction [4].

Structural features include shape, volume and surface area. This type of features can give information for example about the malignant potential of the nodule that is more associated with more spiculated tumours than round ones. Apart from this, volume estimation is important in evaluating treatment response [4].

Texture features and gradient features are known to measure tumour heterogeneity and refer to the relationships and interactions between pixel intensities in a given local neighbourhood. These features extracted in multiple resolution and orientations provides rich information for lung nodule detection and diagnosis but poor information for prognosis [4].

Some features are not subvisual, but are often used together with radiomic ones to build models. Two main groups are:

- Semantic features: that are observed and described directly by radiologists and cover information like the location of the nodule, presence of emphysema, pleural effusions and ground-glass opacity [4].
- Clinical features include some patient info like sex, age, smoking status, clinical stage, and histological subtype.

2.6 Summary

The type of cancer and the mutated genes present in the tumour have a great influence on the choice of the most appropriate treatment for lung cancer especially when it comes to alternatives such as targeted therapy. NSCLC is the most frequent type of lung cancer and the most common mutated genes are EGFR, KRAS and ALK.

Nowadays, genetic testing solutions like biopsies are quite intrusive, expensive and often need to be repeated. Medical imaging is a very promising alternative to these challenges.The extraction

of a large number of so-called radiomic features from medical images is what makes the prediction of mutation status possible.

Radiomic features can be divided into three groups: intensity-based, structural and gradient-based. There exist also two groups of non-subvisual features that are often added in the construction of predictive models in order to improve performance are they clinical and semantic features.

Chapter 3

Literature Review

This chapter is divided into two main subjects. The first part compiles some relevant studies in predicting the status of mutant genes (Section 3.1) and the second part introduces ensemble learning as the proposed solution to improve the results already obtained (Section 3.2).

3.1 Predictive Models for Gene Mutation Status

The studies selected have been divided into two different areas: predictive models for gene mutation based only on nodule features (Subsection 3.1.1) and based in model and other lung structures and diseases (Subsection 3.1.2).

At the end of each section it is available a summary table (Tables 3.1 and 3.2) where it is possible to find a brief description of the objective of the study in question, the method used in the model, the number of patients making up the dataset(s) used, the results obtained presented in the form of AUC (area under the curve) and some features considered relevant. In case of more than one method has been explored in a study, the methods and results corresponding to the one that performed better are presented. In some cases, important features are not included, either because they are not discriminated in the article or because the highlighting of any of them was not justified.

3.1.1 Predictive Models for Gene Mutation Status based on Nodule Features

For this part, it was found five studies that would take into account features related only to the nodule. It should be noted that all of these are using CT images although in some cases, these are not the only ones used.

Liu et al. [17] evaluated the capability of predict EGFR mutation status in surgically-resected peripheral lung adenocarcinomas Asian patients. In this study, the regression model was shown the best results (AUC = 0.709) with the model generated with both clinical and radiomic features. The results showed a significant association between mutant EGFR and female gender, never smoker status, lepidic predominant adenocarcinomas and low or intermediate pathologic grade. Also, Liu

et al. [17] concluded that mutant EGFR status could be predicted by a set of five radiomic features that fall into three broad groups: CT attenuation energy, tumour main direction and texture defined by wavelets and laws.

Zhao et al. [18] developed a deep learning system, represented in Figure 3.1, based on 3D convolutional neural networks (CNNs) to predict EGFR mutant pulmonary adenocarcinoma automatically without requiring precise segmentation of the nodule. The method performed AUCs of 0.758 and 0.750 for the endurance test set and public test set, respectively. The results obtained from using an independent data set suggest that the model is robust and has generalizability. Zhao et al. [18] found that the use of deep learning models allowed for extra image information than conventional radiomics, called deep learning radiomics (DLR). The use of DLR showed better analytical performance but did not present an advantage when combined with conventional radiomics due to the strong correlation between these two. Although DLRs offer more representative features, deep learning models are known to be the black box of artificial intelligence lacking desirable interpretability, especially in medical contexts. Furthermore, was introduced mixup training, which is a strong data augmentation technique that helps to improve generalization. Results show that all models with mixup training significantly outperformed those without, in all of the experiments.

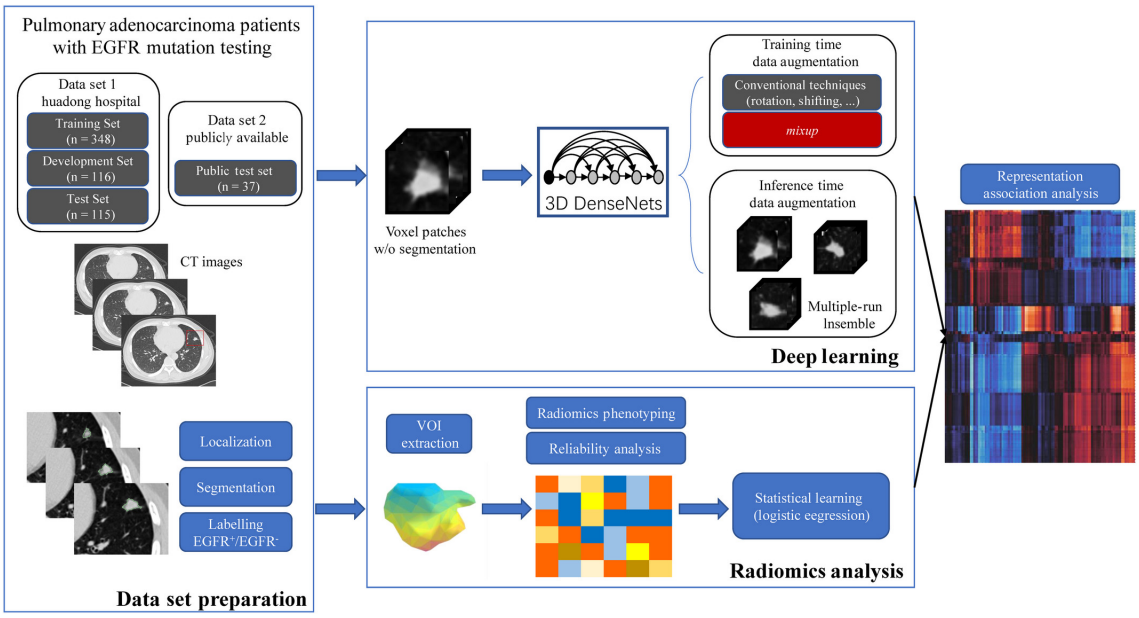


Figure 3.1: Overview of a deep learning model based on 3D convolutional neural networks to predict EGFR mutation used by Zhao et al.. From Zhao et al. [18].

Zhang et al. [19] uses a least absolute shrinkage and selection operator (LASSO) based on multivariable logistic regression to predict EGFR mutation status in patients diagnosed with non-small cell lung cancer (NSCLC). The AUC obtained in the training cohort was 0.8618, and the AUC for the validation cohort was 0.8725 for the model built by both radiomic features and clinical variables.

In this study, 485 radiomic features were extracted from the region of interest (ROI) and combined with traditional clinical features. Then it was used LASSO algorithm and 10-fold cross-validation to shorten all the features. From this result that the combination of 7 radiomic features and 3 clinical features had the potential to build a good prediction model. Moreover, it was created a radiomic signature-based nomogram for individualized mutation prediction that included age, clinical stage, gender, smoking status and Rad-signature. The Rad-signature (which was obtained by the LASSO regression model developed by radiomic features) successfully classified patients to differentiate EGFR mutation subgroup.

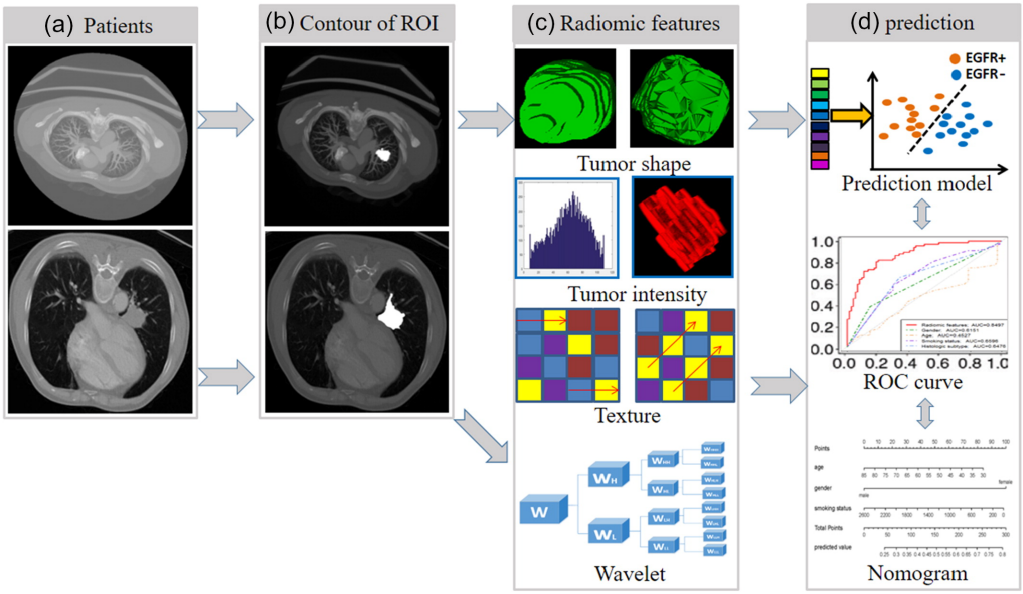


Figure 3.2: Predictive model used by Zhang et al.. From Zhang et al. [19].

Wang et al. [20] proposed an end-to-end deep learning model (illustrated in Figure 3.3) to predict the EGFR mutation status achieving in the primary cohort an AUC = 0.85 and in the independent validation cohort an AUC = 0.81. This model requires a manual selection of the ROI that should be a cubic region that contains the entire tumour. The ROI should be resized by third-order spline interpolation in each CT slice and then given to the model to predict the probability of the tumour being EGFR-mutant. The deep learning model does not demand any further pre or post-processing or image segmentation.

In this same study, three models were built for comparison to the deep learning model: a clinical model, a semantic model and a radiomics model. The clinical model involved sex, stage and age as features, and used a support vector machine (SVM) with a radius-basis kernel for EGFR mutation prediction. The semantic model used sixteen semantic features and multivariate logistic regression. The radiomics model used a random forest containing 100 trees starting from eight selected features from a total of 1108 extracted with PyRadiomics toolkit using recursive feature elimination (RFE). The deep learning model shows to outperform all the three other models.

Shiri et al. [21] compare six feature selection and twelve classifiers to predict EGFR and

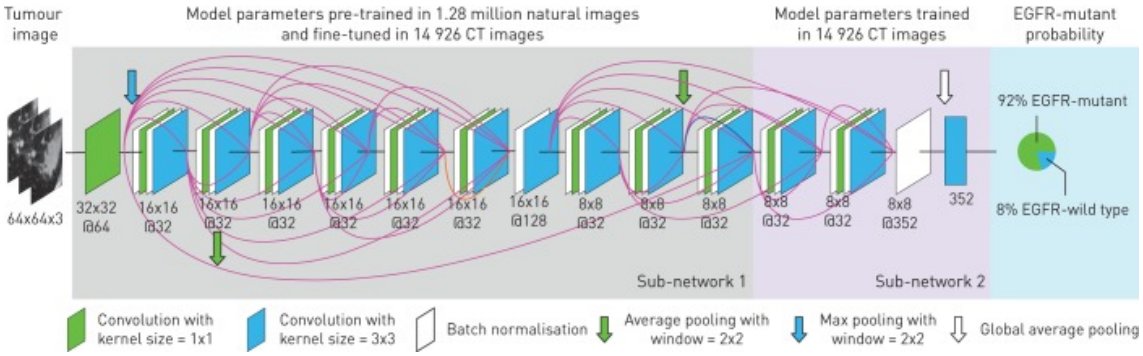


Figure 3.3: Representation of the deep learning model used by Wang et al. From Wang et al. [20].

KRAS mutation status in NSCLC patients based on radiomic features from low-dose computed tomography (CT), contrast-enhanced diagnostic quality CT, and positron emission tomography (PET) imaging modalities from 150 NSCLC patients. The radiomic features used, come not only from the original images, but also from preprocessed images and were extracted using the open-source python library PyRadiomics. In addition to the radiomic features, other conventional clinical PET biomarkers were considered, including metabolic tumour volume (MTV) and standard uptake values (SUVmax, SUVpeak). The preprocessed images were obtained with 64 bin discretization, Laplacian-of-Gaussian (LoG) and wavelet decomposition. Ten-fold cross-validation was used to do model tuning to improve robustness and the developed models were applied on an independent validation set with 68 patients. The best results were obtained when the Stochastic Gradient Descent (SGD) classifier was used for both cases, getting an AUC equal to 0.82 for EGFR and 0.83 for KRAS. In these predictions, LoG preprocessed images of PET were used in the case of EGFR and LoG preprocessed images of CT in the KRAS case. The feature selector was variance threshold and select model respectively. This process is summarized in Figure 3.4.

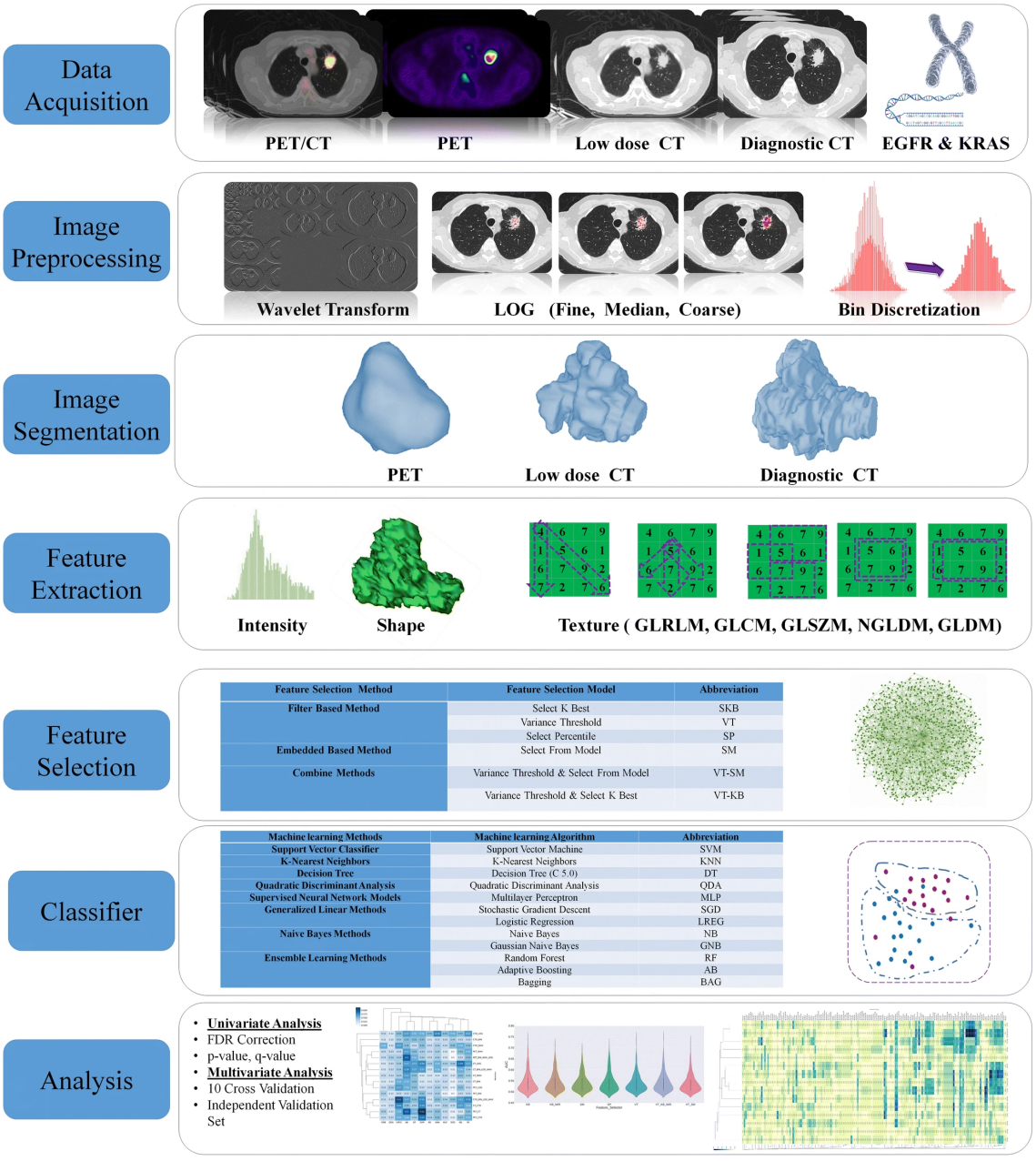


Figure 3.4: Radiogenomics framework used by Shiri et al. study. From Shiri et al. [21].

Table 3.1: Overview of published studies regarding predictive models for gene mutation status based on nodule features.

Reference	Objectives	Methods	#Patients	Results	Important Features
Liu et al. [17]	Evaluate the capability of radiomic and clinical features to predict EGFR mutation status in a cohort of patients with Adenocarcinoma.	Logistic Regression (LR)	298	**EGFR** AUC= 0.709	**Nodule features:** CT attenuation energy, tumor main direction and texture defined by wavelets and laws **Clinical features:** gender, smoking status, histologic subtype, pathologic grade
Zhao et al. [18]	Develop a deep learning system to predict the EGFR mutation status of lung adenocarcinoma based on CT images by integrating recent advances in deep supervised learning, such as dense connections and mixup training.	3D Convolutional Neural Networks (CNNs)	Dataset1: 115 Dataset2: 37	**EGFR** Dataset 1: AUC=0.758 Dataset 2: AUC=0.75	-
Zhang et al. [19]	Predict EGFR mutation status using quantitative radiomic biomarkers and representative clinical variables.	Multivariable Logistic Regression	180	**EGFR** AUC= 0.873	**Clinical features:** age, gender, smoking status **Nodule features:** Tumor shape, intensity, texture, wavelet
Wang et al. [20]	Develop an end-to-end pipeline that requires only the manually selected tumour region in a CT image.	Convolutional Neural Networks (CNNs)	844	**EGFR** AUC=0.81	-
Shiri et al. [21]	Six feature selection methods and 12 classifiers were then used for multivariate prediction of gene mutation status in PET and CT.	(Best results with) Stochastic Gradient Descent (SGD)	68	**EGFR** AUC=0.82 **KRAS** AUC= 0.83	**Nodule features:** tumor volume, shape , texture and intensity.

3.1.2 Predictive Models for Gene Mutation Status based on Nodule Features and Diseases

For this section, it was found four studies that not only considered features from the nodule but also from other lung structures. It should be noted that, like in the Section before, all of these are using CT images.

Pinheiro et al. [11] pretend to analyze results for EGFR and KRAS biological markers according to different combinations of input features. The experiments revealed that the best input combination is to collect both nodule-related and other lung structures features. The maximum mean AUC was 0.7458 for EGFR mutation using the hybrid semantic features, followed by the use of non-nodule semantic features that represented the second-best performance. The worst results correspond to the use of features only from the nodule (radiomic and semantic type). This experiments also suggests that although CT scans imaging phenotypes are related to EGFR mutation status, the same may not be true for KRAS since it was not possible to establish an acceptable model in this case.

With this study, Rizzo et al. [14] intend to validate some previously identified associations between radiological features and clinical features EGFR and KRAS alterations in an independent group of 122 NSCLC patients. The results confirmed an association between EGFR+ with an internal air bronchogram, pleural retraction, emphysema and lack of smoking and KRAS+ with round shape, emphysema and smoking. The model employed yielded AUCs of 0.82 to EGFR+ and 0.60 to KRAS+. Rizzo et al. [14] concluded that even though this study confirms the relevant association of clinical and radiological features with EGFR and KRAS, this model can not overcome the prediction of mutations using smoking history alone.

Gevaert et al. [15] try to predict EGFR and KRAS mutation status in NSCLC patients building a decision tree based in semantic image features annotated by a thoracic radiologist for each patient. Although it was found a statistically significant model for predicting EGFR the same did not happen for KRAS mutations. A possible explanation for the worse results could be the low representativeness of KRAS on the dataset used in this study. Another hypothesis may simply be that this gene does not have as strong radiographic manifestations as EGFR. For the construction of the decision tree that will lead to EGFR prediction were used four variables: emphysema, airway abnormality, the percentage of ground glass component and the type of tumour margin. The decision tree can be observed on Figure 3.5 and its implementation presented as a result an AUC = 0.89.

Song et al. [16] combined clinical, conventional CT and radiomic features to predict ALK mutations with a database that included 335 patients with lung adenocarcinoma. In total were extracted one thousand two hundred and eighteen quantitative radiomic features from the semi-automatically delineated volume of interest (VOI) of the entire tumour obtained with PyRadiomics tool, twelve conventional CT features and seven clinical features. These features come from original and the pre-processed CT images with high-pass and low-pass wavelet filters or Laplacian of Gaussian (LoG) filters. Then all of these features were selected using a sequential of the F-test-

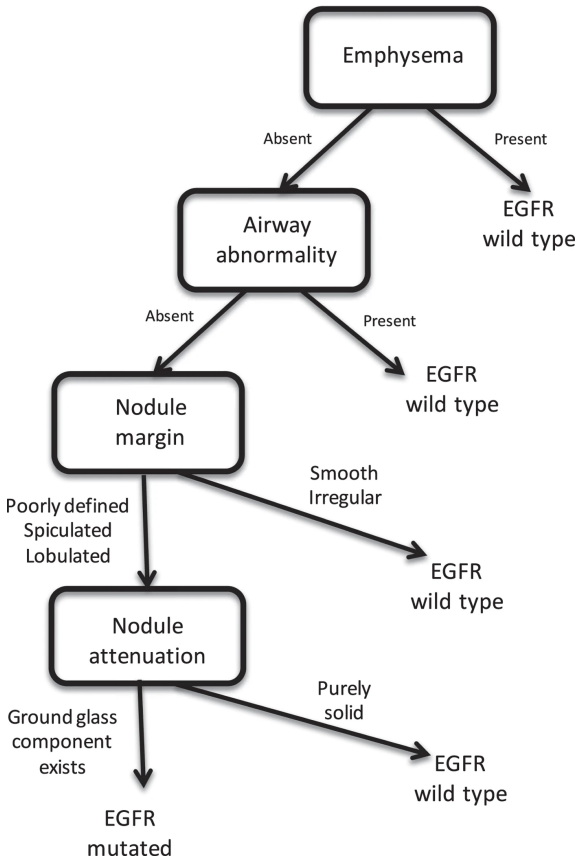


Figure 3.5: Decision tree using semantic image features used by Gevaert et al.. From Gevaert et al. [15].

based me, the density-based spatial clustering of applications with noise and the recursive feature elimination methods. Finally, three predictive models were built by a soft voting classifier, each one corresponding to a different set of characteristics - the radiomic model with the radiomic features, the radiological model with radiomic plus conventional CT features and the integrated model that included radiomic, conventional CT and clinical features. It was employed grid search that maximized the AUC of the repetitive 10-fold cross-validation to found better hyper-parameters for feature selection and model training. As for the conclusions concerning the characteristics, it was noted that for conventional CT features, pericardial effusion, local lymphadenopathy, lobulated margin, and the absence of pleural retraction sign were correlated with ALK-rearranged status. The majority of radiomic features associated with ALK mutations reflected information around and within the high-intensity voxels of lesions. In addition, the intra-tumoural cavity and left lower lobe location were also associated with the ALK mutation status whereas clinical stage I, male sex and current smoking were inversely associated with it. Smoking history was the most powerful factor to differentiate ALK mutated and non-mutated lung adenocarcinomas. It has also been verified that add clinical information and conventional CT features improved the performance of the radiomic model in the primary cohort (AUC = 0.83–0.88), but not in the test cohort (AUC

= 0.80–0.88). The methodology adopted by

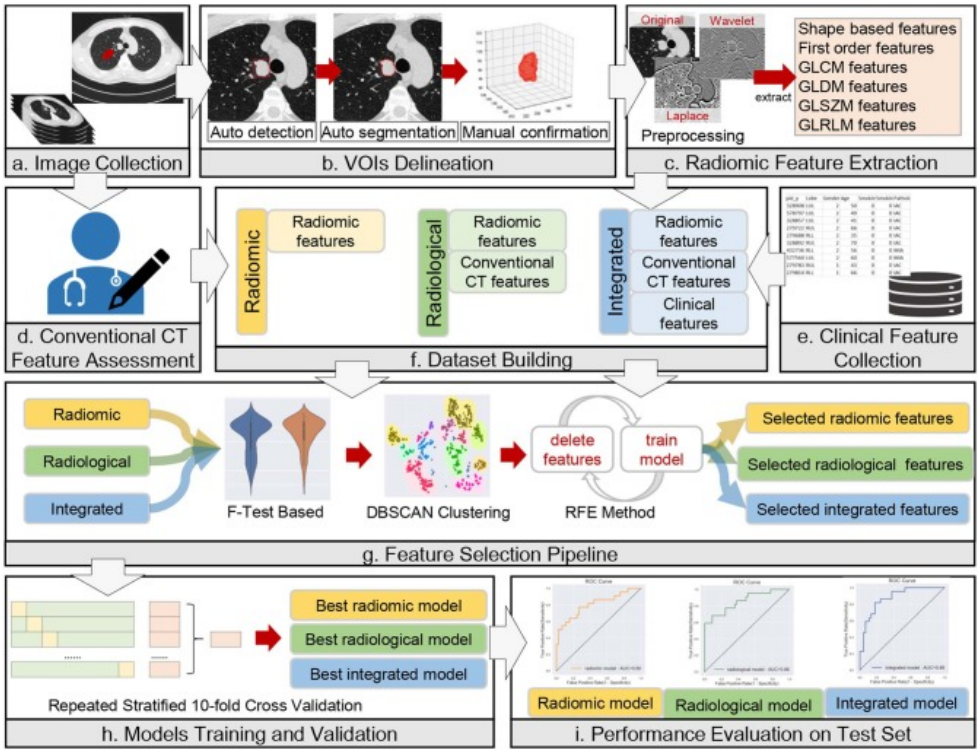


Figure 3.6: Workflow of data analysis used by Song et al..From Song et al. [16].

Table 3.2: Overview of published studies regarding predictive models for gene mutation status based on nodule features and other lung structures and diseases.

Reference	Objectives	Methods	#Patients	Results	Important Features
Pinheiro et al. [11]	Analyse the results for EGFR and KRAS biological markers according to different combinations of input features.	Gradient Tree Boosting	211	**EFGR** AUC=0.746	**Clinical Features:** gender and smoking status
Rizzo et al. [14]	Validated the significant association of clinical and radiological features with EGFR/KRAS alterations.	Univariate Analysis	122	**EGFR+** AUC=0,82 **KRAS+** AUC=0,60	-
Gevaert et al. [15]	Investigated whether EGFR and KRAS mutation status can be predicted using imaging data.	Decision Tree	186	**EGFR** AUC=0.89	**Clinical features:**smoking status **Nodule features:**percentage of ground glass component and the type of tumor margin **Structures and Diseases:** emphysema, airway abnormality
Song et al. [16]	Predict the anaplastic lymphoma kinase (ALK) mutations in lung adenocarcinoma patients noninvasively with machine learning models that combine clinical, conventional CT and radiomic features.	Recurrent Convolutional Neural Networks	335	**ALK** AUC=0.80–0.88	**Clinical features:** age, sex, smoking history, smoking index, clinical stage **Nodule features:** distal metastasis, and pathological invasiveness of the tumor,maximum diameter, mean CT attenuation, lesion location, involved lobe, density, margin, cavity, calcification **Structures and Diseases:**pleural retraction sign, pleural effusion, pericardial effusion, and local lymphadenopathy.

3.1.3 Discussion

It was possible to collect studies with predictive models for the three most common mutated genes. In a total of nine studies, eight aimed to predict the EGFR mutated status, two to KRAS mutation and only one to predict ALK-rearrangement. This suggests that EGFR has a higher predictive ability and therefore a higher correlation with radiomic features. KRAS has been referenced by some authors due to the impossibility of creating acceptable models for predicting the status of this gene. As for ALK, it is not possible to conclude about the predictive ability since only one predictive model has been found which may only suggest that the possibility to predict the status of this gene by radiomics was simply not studied enough yet.

According to the literature, it was possible to establish some correlations between the different genes and some specific features used in the prediction models used. It stands out among all, the smocking status that was shown a strong relation with all three genes taken into consideration.

The authors, who made comparative studies, concluded that the combined use of radiomics with clinical features generally tended to improve performances.

Looking now more closely at Subsection 3.1.2 , which compiles the studies using features of lung structures in addition to those associated with the nodule, the conclusion is that the use of the two combined has the potential to improve the model's results.

From all the literature, the predictions made using deep learning techniques that allowed for extra image information, called deep learning radiomics (DLR), outperformed the results obtained with conventional radiomic features. However, these methods require a large amount of data that is often not available.

3.2 Proposed Solution: Ensemble Learning

Ensemble learning is based on the principle of combining several predictions to obtain a more accurate final result analogous to the way humans gather diverse opinions combine them to make complex decisions [6]. The use of ensemble methods present many advantages over single machine learning algorithms that are usually used in prediction of mutated genes. For this reason, the use of assembly methods emerges as a proposed solution with the objective of overcoming the barriers of the methods previously used.

3.2.1 Ensemble Methods

Ensemble learning is described as a technique that combines multiple base learners to make a decision/solve a problem employing an aggregation rule. It is typically used in supervised machine learning (ML) tasks [22]. This approach is based on the idea that by combining multiple models, the errors of a single learner will be compensated by the others. At the same time averaging different learners reduces the chance of choosing incorrectly a single one [6].

In Figure 3.7 is represented a scheme of the typical design of an ensemble approach. The single classifiers 1 to n take a set of labelled examples inputs to produce a model, these models

are built to solve the problem by themselves so all the classifiers return a preliminary output. The different outputs of every learners will then be combined by an aggregation rule and produce a final output that must be more accurate than the first ones that came from the single techniques [5, 22]. An intelligent combination rule often proves to be a more efficient approach [6].

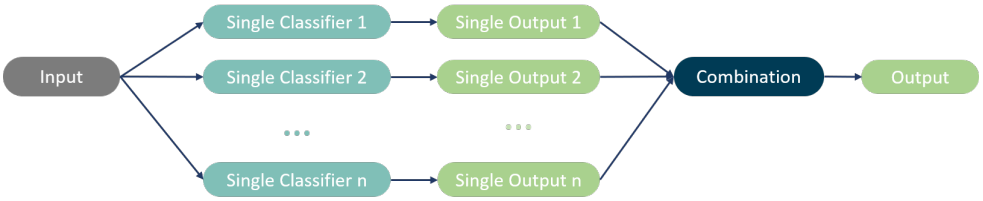


Figure 3.7: Ensemble methods approach.

There are homogeneous and heterogeneous ensembles according to the use of learners. Homogeneous ensembles combine at least two different variants of the same technique while heterogeneous ones combine at least two different ML techniques [5, 22].

Some of the most popular learning algorithms are decision trees, neural networks, naive bayes classifier, k-nearest neighbor, support vector machines and kernel methods. As for combiners, some of the most used are the majority vote, probabilistic, regression and weighted average [5].

3.2.2 Advantages of use ensemble methods

There are some reasons that justify why ensemble methods have better performances and overcome single ML algorithms [6]:

- **Overfitting avoidance:** Overfitting is a thing that occurs when a statistical model fits its training data exactly. When this happens, the algorithm cannot work accurately in unseen data i.e. the model has no generalisation capacity. As ensemble learning allow take into account different hypothesis this will reduce the risk of choosing an incorrect hypothesis and improves the overall predictive performance.

- **Computational advantage:** By the combination of several learners, the risk of getting stuck in local optima decreases.

- **Representation:** By combining different models, the search space can be extended and, therefore, a better fit to the data space is achieved. As can be seen in the Figure 3.8 the optimal solution is very hard to get with a single learner but possible by the combination of multiple single learners.

There are also some challenges that sometimes ensemble methods can mitigate like class imbalance in the dataset and concept drift (when distribution of features and the labels tend to change over time) [23].

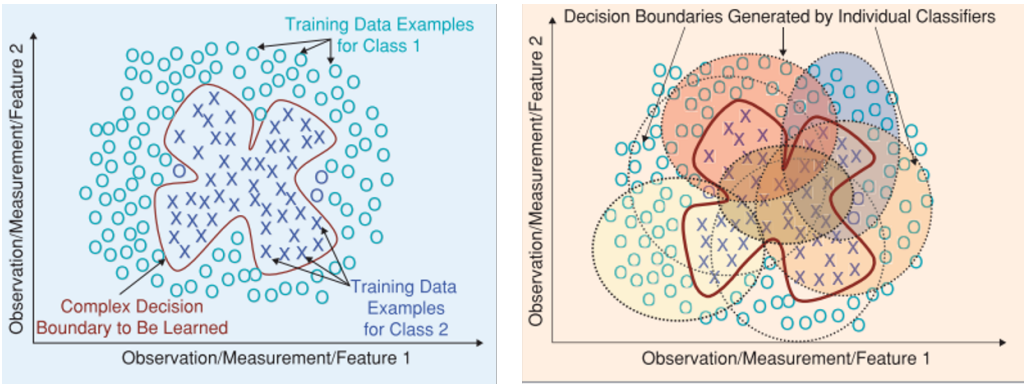


Figure 3.8: Contribution of different single predictors and importance of diversity of outputs in a good final ensemble prediction. From Polikar [23].

3.2.3 Construction of an ensemble model

Construct an ensemble models implies to select a method for training the participant models and choosing a proper process to combine the baselearners output. But that is not all,there are other important factors to be taken into consideration if it is intended to build a good ensemble method [5, 6]:

- **Diversity:** The participating baselearners must be sufficiently diverse to achieve a good final predictive performance which implies that two baselearners at the same data instance produce different outputs.
- **Predictive performance:** It is important that each of the baselearners also has a performance as high as possible so that it does not become prejudicial to other methods when combining the outputs of the various baselearnerss.

These two principles may seem contradictory but, in other words, the main objective is to combine predictors with high performances but that have uncorrelated errors [6]. The predictive performance of the whole ensemble is higher, the less correlated are the errors made by individual baselearners [24].

3.2.3.1 Baselearners

Although the ensemble is all based on the same principle there are different ways of including different baselearners [6]:

- **Input manipulation:** consists of the use of different training subset to ensure a variety of inputs are used for the different base models. This method is especially effective for cases where small changes in the training set may result in a completely different model. The distribution of the data among the different baselearners may be random or determined according to the class distribution in the entire dataset.

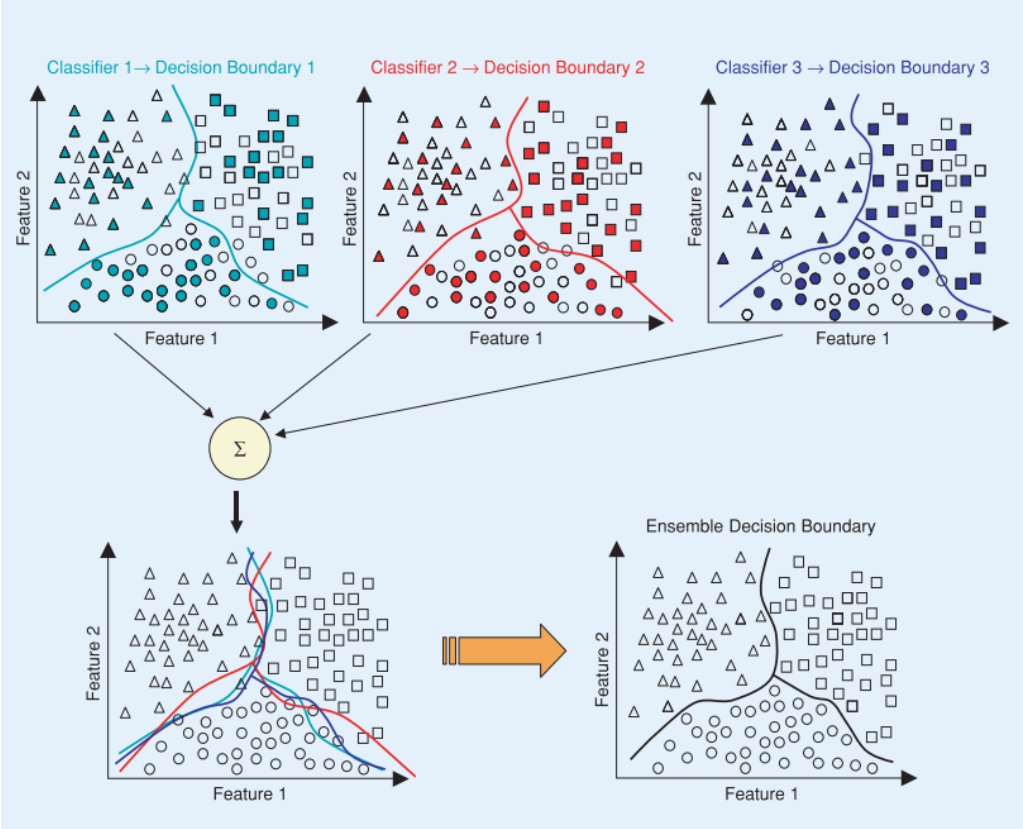


Figure 3.9: Representation of the contribution of different classifiers in an ensemble approach. From Polikar [23].

- **Manipulated Learning Algorithm:**where the use of each base model is altered. This alteration could be done several ways like make the model choose different convergence paths, inducing randomness by selecting one out of k best splitting attributes at each split, distributing neighbours expanding the feature space generating different combinations of the original features, train the base models with varied hyperparameter values, etecetera.

- **Partioning:** dividing the original dataset into smaller subsets and use different ones to train the different baselearners. This partitioning can be done horizontally in which each baselearner uses all the features but different features or vertically where each baselearner uses the same instances but with different features.

- **Output manipulation:** each class is encoded as an L-bit code-word where L is the number of classifiers participating in the ensemble. The purpose of each classifier is to predict a bit L of the code-word. Classifiers are then applied for new instances to generate L-bit strings that represent the predictions. The chosen class to be predicted for a given instance is the class whose code-word is the closest to the instance string. Closeness can be measured using different methods such as euclidean decoding and Hamming distance.

- **Ensemble hybridization:** as the name suggests this approach combines at least two strategies to build an ensemble. Random forests are the best-known manifestation of this method as it manipulates the instances when building each tree and manipulates the learning algorithm by choosing randomly a subset of features at each node.

3.2.3.2 Output Fusion

Just as there are different ways of combining baselearners, there are also different methods for merging outputs. It's possible to distinguish two different procedures the weighting methods and meta-learning methods [6].

In the weighting method, the base model outputs are combined by assigning weights to each one. This technique is most suitable when the performance of the base models is comparable. Majority voting is the simplest weighting method and selected the class pursuant the one who gets more votes [6].

As for meta-learning method, there is one more learning stage, the outputs from the base learners are inputs to a new learner that generates the final output. This approach is good when base learners have different performances on different subspaces, i.e. when base models consistently correctly classify or consistently misclassify certain instances. Stacking is one of the most popular meta-learning techniques [6].

3.3 Summary

This literature review has shown the potential of radiogenomics. Analysing the collected studies there is a great incentive to use features that not only take into account the nodule but also other structures of the lung. Additionally, it is concluded that the combination of radiomic and clinical features tends to improve the performance of the models. As regards the different mutated genes analysed, EGFR shows a higher predictive capacity than KRAS since it was often not possible to establish acceptable models for the latter. As for ALK, only one study combining characteristics of the nodule with those of other lung structures was found.

By the need to overcome some difficulties still existing in the prediction of mutated genes, the possibility of combining radiogenomics with ensemble learning has arisen. The ensemble method consists of combining more than one single classification technique under a specific combination rule. This process confers some characteristics that provide advantages over the use of simple ML approaches when used alone. Although this technique presents promising properties there are factors as the diversity and the quality of the performance of the methods used as base-learners that can compromise its good performance.

Chapter 4

Mutant Prediction

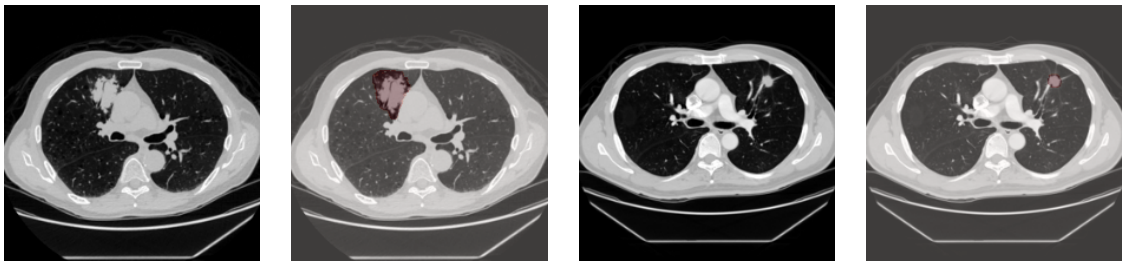
This chapter covers the whole generation and development process of the EGFR status prediction following the ensemble approach. The model construction itself, discussed in Section **??**, is only one part of the process. It also includes the selection of the dataset, which is presented in Section 4.1, the feature extraction discussed in Section 4.2.2, and the preprocessing of those features, Sections 4.2.3 and 4.2.4. After the initial model building, there were some modifications and experiments that are briefly described in the present chapter, Section 4.3. These experiments will be better explained with the complementary information present in the next chapter, Results and Discussion, since many of the changes are motivated by obtained results.

4.1 Public Data

To develop the present study it was necessary to find a dataset that would satisfy the following needs:

- Availability of CT image
- Availability of tumour segmentation
- Availability of EGFR mutation status label

The NSCLC-Radiogenomics Dataset [25] was chosen because it fulfils all the required conditions. This dataset is publicly available and includes clinical and imaging data on 211 NSCLC patients collected between 2008 and 2012 by Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System. Out of the 211 subjects, only 117 were considered since only these owned tumour binary masks and a EGFR mutation test results [25]. The CT scans included in this dataset were obtained using different scanner models and scanning protocols, presenting variations in slice thickness from 0.625 to 3 mm (median: 1.5 mm) and X-ray tube current from 124 to 699 mA (mean: 220 mA) at 80–140 kVp (mean: 120 kVp). Semantic tumour annotations and clinical history of the patients are also available. Figure 4.1 represents two CT slices examples of EGFR wild type patient and an EGFR mutated patient from the NSCLC-Radiogenomics Dataset [25].



(a) (b) (c) (d)

Figure 4.1: Examples of CT scan in axial projection for a) *EGFR* wild type and; b) with color overlay in the semitransparent mode to enhance the nodule region; c) CT scan in axial projection for *EGFR* mutated patient; d) CT slice with color overlay in the semitransparent mode to enhance the nodule region. Both cases are from the NSCLC-Radiogenomics Dataset [25].

4.2 Data Preparation

With the dataset chosen to train and test the model to be developed, it is necessary to prepare the data to be used. Motivated by the literature, the data preparation was based on the premise that in addition to features extracted from the nodule, information present in other structures in the lung is relevant for gene prediction [11, 14, 15, 16]. Clinical data was also considered since that some literature also suggests that the combination of these ones with radiomics improve the performance of the models [14, 17, 16].The distribution of the clinical data from the 117 patients considered for this study is presented in Table 4.1.

Table 4.1: Clinical data distribution from 117 patient dataset.

Gender		**Smoker Status**		
Male	**Female**	**Non-Smoker**	**Former**	**Current**
76,07%	23.93%	13.68%	65.81%	20.51%

4.2.1 CT Images Pre-Processing

The CT scans use Hounsfield Units (HU) to represent the information.This scale is based on radiodensity that considers that at standard pressure and temperature the radiodensity of water is zero HU and radiodensity of air is -1000 HU. The formula for calculating HU value based on this scale is shown in Equation 4.1, where where μ is the original linear attenuation coefficient of substance, μ_{water} is the linear attenuation coefficient of water and μ_{air} is the linear attenuation of air.

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{4.1}$$

Sometimes, the content of CT does not come in HU units and it is necessary to convert it. This conversion, represented in Equation 4.2 is made using Rescale Slope and the Resclape Intercept fields present in the metadata associated with the scan and are defined by the hardware

manufacturer [26].

$$HUValue = PixelValue \times RescaleSlope + RescaleIntercept \tag{4.2}$$

All the images were then normalized using the min-max normalization method [27]. All values below -1000 HU (radiodensity of air) are set to 0 and all values above 400 HU (representing hard tissues, not relevant for this study purposes) are set to 1. For the values between is performed a linear transformation to put these values in the [0,1] range.

Since the CT images were obtained using different scanner models and scanning protocols, it is important to perform resampling so that all images are represented in a standard form. This is accomplished by setting the space between consecutive slices to 1mm and the space between pixels in the same slices to [1mm; 1mm]. Each slice dimension was calculated to match this new spacing, including the tumour masks,and the resampled image is obtained by interpolation.

For the segmentation of the lung containing the nodule a 2D segmentation model was used with the lung binary masks that is based on the U-Net architecture [28].

4.2.2 Feature extration

The feature extraction of features was performed using the open-source package *PyRadiomics* [29]. A total of 1316 features were extracted divided into seven categories: shape-based (14 features), first-order (intensity-based) (18 features), Gray Gray Level Co-occurrence Matrix (GLCM) (24 features), Gray Level Dependence Matrix (GLDM) (14 features), Gray Level Run Length Matrix (GLRLM) (16 features), Gray Level Size Zone Matrix (GLSZM) (16 features) and Neighboring Gray Tone Difference Matrix (NGTDM) (5 features). This features are extracted from both filtered and unfiltered images except for shape-based ones that are independent from intensity values and therefore only were extracted from unfiltered images [29].

The filtered images are originated by wavelet and Laplacian-of-Gaussian transformations. With wavelet filtering, the original image is decomposed into low and high frequencies. The 3D image can be constructed as separable products of 1-D wavelets. The volume F(x, y, z) is filtered along the x, y and z dimension, with low-pass (L) and high-pass filters(H), resulting in eight sub-volumes: LLL, LLH, LHL, LHH, HLL, HLH, HHL and HHH [30].

The LoG filter produces a derived image for each sigma value applied, to emphasize areas of grey level change. The sigma defines how coarsely the emphasized texture should be in this study 5 images were generated from this filter corresponding to sigma values equals to 1.0 mm, 2.0 mm, 3.0 mm, 4.0 mm and 5.0 mm [31].

4.2.3 Data Augmentation

After obtaining the radiomic features, it was necessary to add the clinical features, which underwent a binarization process using the one-hot-encoder method. This method creates a binary column for each category and returns a sparse matrix. Thus increased the number of features from a total of 1316 to 1321 [32].

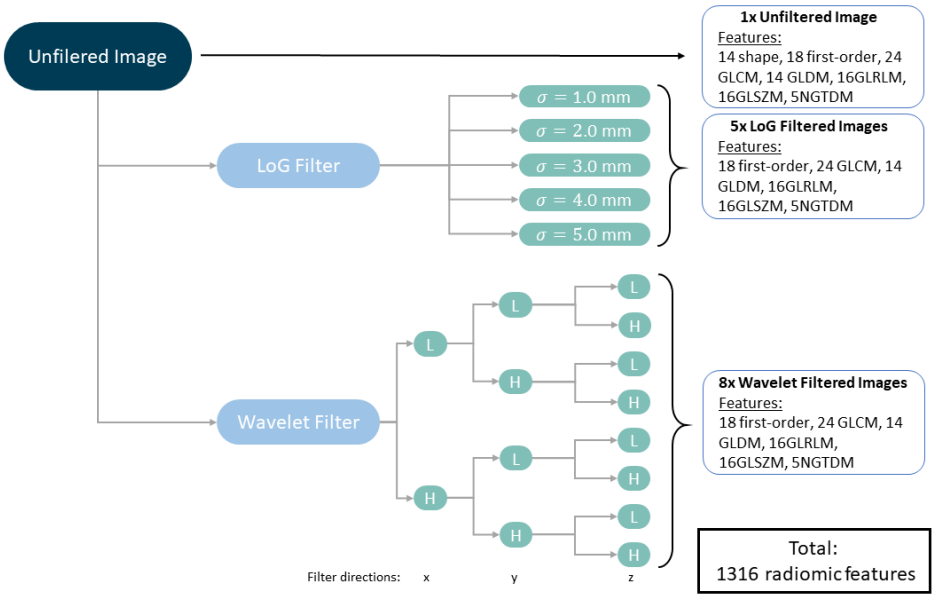


Figure 4.2: Feature extraction overview.

The dataset used in this study has a distribution of 80% of wildtype EFGR to 20% mutated EFGR, this will result in a model biased towards the negative class. To overcome this class imbalance a Synthetic Minority Over-Sampling Technique (SMOTE) was performed. SMOTE is an over-sampling methodology in which the minority class is over-sampled by creating "synthetic" examples. The synthetic examples of the minority class are generated using k-minority class nearest neighbours. For instance, if the amount of over-sampling needed is 200%, two neighbours from the five nearest neighbours of a point are chosen randomly. Then a random point in the segments between the original point and its neighbours is selected as a new synthetic example [33].

4.2.4 Dimensionality Reduction

A common practice that allows reducing not only the overfitting but also the computational cost, which also allows faster processes, is to reduce the dimensionality of the dataset. For this purpose, it was used PCA in the dataset after SMOTE. Principal component analysis (PCA) is a statistical technique employed to reduce the dimensionality of datasets while maintaining the maximum amount of variance. This process works by creating a set of new variables, the principal components (PC), through linear combinations of the original variables [34]. From previous work [35], PCA feature selection method was implemented with 70% of variance in the feature set for best performance. This results in a total of 1316 features removed and 5 retained features.

4.3 Model Implementation

The main objective of this study is to develop a model using ensemble learning capable of predicting mutated genes. An overview of the pipeline used is presented in Figure 4.3.

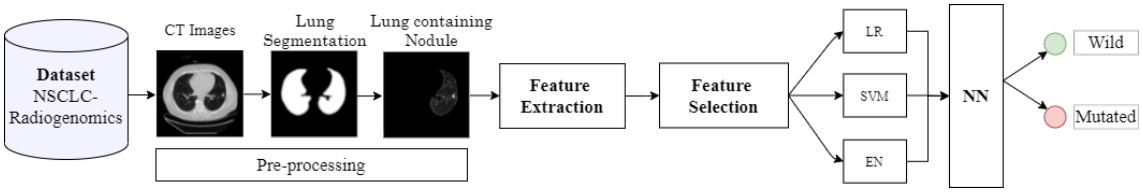


Figure 4.3: Overview of the model with the best performance studied.

4.3.1 Baseline

As a starting point we used as baseline a model developed by Xiao et al. [36]. This model proved to be accurate and effective in predicting cancer in the three RNA-seq datasets from Lung Adenocarcinoma, Stomach Adenocarcinoma and Breast Invasive Carcinoma. An overview of the model is shown in Figure 4.4.

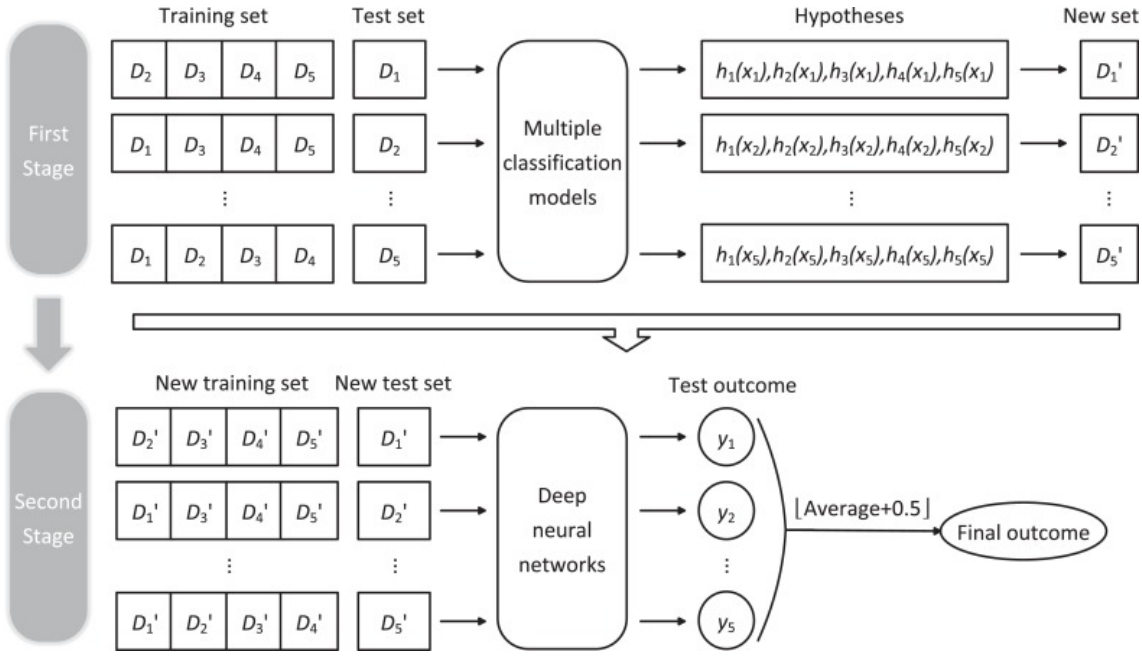


Figure 4.4: Deep learning-based ensemble method. From Xiao et al. [36].

The model is divided into two different stages. In the first stage, S-fold cross-validation is employed to divide the initial data into S groups of training and testing datasets. After that, multiple classifiers (first-stage models) are trained with the data of S - 1 groups and then tested in the remaining group set. Subsequently, each model predictions are assembled in a set of binary

hypotheses, represented in Figure 4.4 by H_i=[$h_1(x_i)$, $h_2(x_i)$, ... , $h_j(x_i)$]. The numbers associated with the h represented the j different classifiers used, and the x_i represented each input data corresponding x_1 to the input when D_1 is the data test, x_2 when is D_2 and so on. Each H_i set will generate a new set D_i'. Then, in the second stage of the model, a deep neural network classifier (second-stage ensemble model) is used and has as inputs the new sets D_i'. It is expected that the outcome of the second stage can be more accurate and the generalisation error can be reduced [36].

4.3.2 First-stage Implementation

To replicate the model presented above, a 5-fold cross-validation was applied in the NSCLC-Radiogenomics Dataset, after the features extraction. This step will not only have effect for single classifier separately, but also generates new datasets for the ensemble stage. This should help to avoid overfitting and reduce the generalisation error.

For the first-stage classification step the five models chosed to be implemented in the first approach were da same used by Xiao et al.: k-nearest-neighbour (KNN), support vector machines (SVMs), decision trees (DTs), random forests (RFs), and gradient boosting decision trees (GBDTs).

In order to improve the performances of each of the models the hyperparameters were tuned using Grid Search CV (cross-validation) on the training data. Grid search CV is an approach that methodically build and evaluate a model for each combination of algorithm parameters specified in a grid [37]. The used hyperparameters that make up the grid in this study are presented in the Table 4.2.

Table 4.2: Hyperparameters of the ML model values used in the Grid Search CV.

Algorithm	**Hyperparameter**	**Values**
KNN	Number of Neighbours	1, 3 , 5, 7, 9, 11, 13, 15, 17, 19, 21
	Weights	uniform, distance
	Metric	euclidean, manhattan, minkowski
SVM	Kernel	poly, rbf, sigmoid
	C	0.01, 0.1, 1, 10, 50
	Gamma	scale
DT	Criterion	gini, entropy
	Splitter	best, random
	Maximum Depth	1, 10, 100.1000
RF	Number of Estimators	10, 100, 500,1000
	Maximum Features	sqrt, log2
GBDT	Number of Estimators	10,100,1000
	Learning Rate	0.001, 0.01, 0.1
	Subsample	0.5, 0.7, 1
	Maximum Depth	3, 7, 9

4.3.3 Second-stage Implementation

The second-stage model is trained to combine the predictions from first stage models to make a final prediction. In this stage, a 3-layer NN model was used with the input layer that contains five neurons corresponding to the five models used in the first stage of the model. Grid Search CV was used to choose the number of neurons of the hidden layer. The output layer has one neuron whose output was 0 or 1, which correspond to wild type mutant EGFR type, respectively.

4.3.4 Experiments

As seen above, the good performance of the ensemble model is intrinsically linked to the performance of the base-classifiers. As such, and based on the study of Morgado et al. [35] we replaced the models with the worst performance by three others that were more satisfactory, resulting in a first stage composed of the following models: logistic regression (LR), support vector machine (SVM) (now including linear SVM in in the choice made by grid search CV), elastic network (EN) and random forest (RF) and gradient boosting decision trees (GBDT). For including methods that need to make linearizations there was a need to include a new normalization through the min-max method, only radiomic features are normalized, clinical are not considered [27].

Analysing the influence of baselearners on the ensemble method result, we studied the impact of using only three methods with better performance as inputs to the network that works as a combination rule instead of using all five methods. The two worst performers are, in this case, the RF and the GBDT.

To provide as much information as possible to the ensemble, occurred the idea of alternatively feed the second stage a set of probabilities as input instead of a set of binary hypotheses. For that, instead of gather a binary prediction from each of the methods of the first stage, the probability of the gene being mutated was collected.

Alternatives to the model were also explored by replacing the neural network with other methods such as: majority voting (MV) and random forest (RF).

4.4 Summary

In the attempt to implement a model capable of predicting the mutated EGFR gene, several steps had to be followed.

At first, it was necessary to choose a dataset that met the necessary conditions. In this case, to have available the CT images, tumour segmentations and EGFR mutation status labels. It was possible to consider 117 of 211 patients from the chosen dataset, NSCLC-Radiogenomics Dataset.

The collected CT images then underwent pre-processing before radiomic features extraction for a total of 1316 from filtered (LoG and wavelet) and unfiltered images. The gender and smoking status of each patient were also added to these features.

Finally, the data obtained were exposed to data augmentation processes through SMOTE and dimensionality reduction through PCA.

Now that the data are ready to be used by the model, let's move on to the model building process.

As a baseline we used an existing model developed with the objective of cancer prediction but based on another type of inputs other than medical images.

This model can be divided into two phases. In the first phase, different classifiers are used to make binary predictions that will be used in the second phase as input for a new classifier that aims to make a final prediction more accurate than the one generated by the previous models.

In an intent to improve this model several experiments were made. This includes, swapping some classifiers in both the first and second stages, changing the amount of classifiers used as input in the second stage, and alter the binary predictions by probabilistic ones to serve as inputs as well.

The results obtained are presented in the next chapter, Chapter 5

Chapter 5

Results and Discussion

The results obtained during the various experiments were decisive to define the guidelines along the way to obtain the best model. The results obtained in the different experiments are presented and discussed in this chapter. Bear in mind that all AUCs presented are the result of the average of 50 different dataset combinations. The seeds that generate the different combinations were saved, so that they could be replicated in the different experiments to make the comparative study as precise as possible.

Starting with the use of SMOTE, this technique to reverse class imbalance showed to improve model performance. The implementation of SMOTE proved to be most effective when done on both the first and second phase input data of the model. In fact, when not used in the second phase the neural network was shown to be unable to make predictions consistently returning AUCs equal to 0.5 for predicting only one of the classes.

Table 5.1: Hyperparameters of the ML model values used in the Grid Search CV.

Algorithm	Hyperparameter	Values
LR	Solver	newton-cg, lbfgs, liblinear, sag, saga
	Penalty	l1, l2, elasticnet
	C	0.01, 0.1, 1 ,1.5, 2.0, 2.5, 10, 100
SVM linear	C	0.1, 1, 10, 100
	Gamma	0.001, 0.01, 0.1, 1
EN	Maximum Iterations	1, 5, 10
	Alpha	0.0001, 0.001, 0.01, 0.1, 1, 10, 100
	l1 ratio	0.1, 0.2, 0.4, 0.5, 0.6, 0.7
RF	Number of Estimators	10, 100, 500,1000
	Maximum Features	sqrt, log2
GBDT	Criterion	friedman mse, mse, mae
	Maximum Depth	3, 4, 5, 6, 7, 8, 9

In the choice of multiple classifiers for the first phase, the five used in the model proposed by Xiao et al. [36] were initially chosen but they did not immediately produce results as satisfactory as those presented in the study. The main reason pointed out for this is the use of a different type of

data, with special emphasis on the dataset dimension that has special relevance in model training. Due to the importance associated with the good performance of these first models for the effective operation in ensemble learning, it was decided to replace some of them. Using previous work [35], it was chosen to substitute the three methods with the worst performance (KNN and SVM and DT), by three others with greater potential (LR, linear SVM and EN).

Also for these new models, the hyperparameters, presented in Table 5.1, were tuned using the GridSearchCV on the training data.

The AUCs obtained for each one of this new methods are presented in Table 5.2 in the format AUC$\pm\sigma$ with σ being the standard deviation.

Table 5.2: AUC of the ML single methods of the first stage.

Method	**AUC$\pm\sigma$**
LR	0.712 $\pm$0.119
SVM	0.711 $\pm$0.119
EN	0.712 $\pm$0.120
RF	0.656 $\pm$0.136
GBDT	0.642 $\pm$0.130

Analysing the results, three methods stand out for their best performance: LR, linear SVM and EN. For the hypothesis that the two methods that present lower AUCs could be detrimental to the learning of the neural network in the second phase, it was decided to make a comparative study of the performance of this model using as input the predictions of three or five initial classifiers. The results obtained are shown in table 5.3.

Table 5.3: AUC of the initial NN second stage.

5 methods (LR, SVM, EN, RF, GB)	**3 methods (LR, SVM, EN)**
0.624 $\pm$0.143	0.644 $\pm$0.125

With this new information, the conclusions are that the replacement of the worst performing methods by others that allowed better results favoured the operation of the neural network and that the subsequent elimination of two of the single methods with worse performance also suggested advantageous.

At this point, the results obtained through NN still do not satisfy the purpose of using ensemble, which would be to outperform the predictions made by the classifiers when used separately. With this in mind, the next approach was aimed to increase the information given to the second stage of our model. To do this, we replaced the binary predictions of every single method that constitute the D' datasets with ensembles of probabilistic predictions.

As illustrated in Table 5.4, use probabilistic outputs from the baseclassifiers instead of binary outputs as inputs for the second stage shows to bring advantages for the neural networks both

Table 5.4: Comparison of final prediction of the models with binary and probabilistic inputs at the second stage with the initial NN.

Input of the second stage	NN (5 input methods)	NN (3 input methods)
Binary	0.624 ±0.143	0.644 ±0.125
Probabilistic	0.687 ±0.120	0.706 ±0.122

when using the inputs of five classifiers and when using three. This study also reinforced the idea that using only 3 better methods would be beneficial.

Finally, the replacement of the second phase method by two other methods was tried. For the simplicity of the method the majority voting (MV) was chosen, and for being a method commonly used in this type of learning and for showing positive results in this field the random forest (RF) [38].

Table 5.5: AUC of the RF second stage.

Input of the second stage RF	AUC $\pm\sigma$
Binary	0.646 ±0.128
Probabilistic	0.625 ±0.140

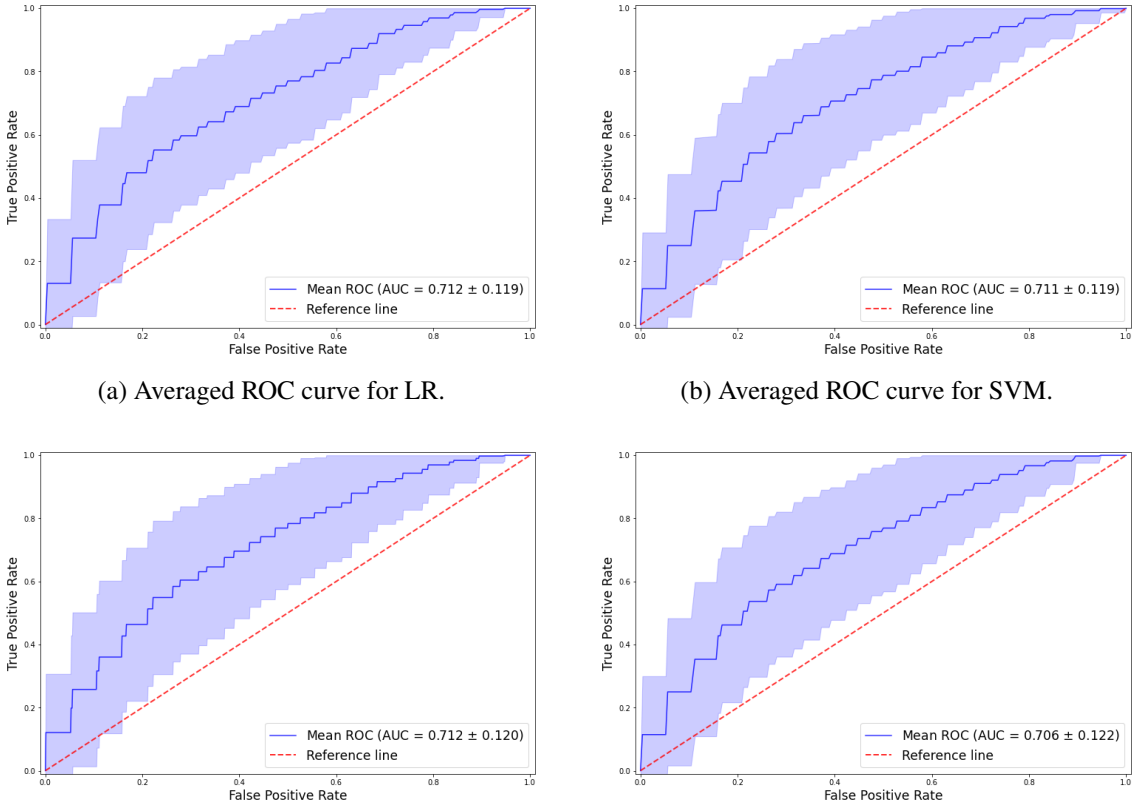
The use of random forest in the second stage showed a curious behaviour. Looking at table 5.5 and analyzing the behavior of the same RF, it performs better when using binary inputs and is also better than the NN using binary inputs. However, the result of the RF with binary input still loses to both NN with probabilistic inputs. Note that for evaluating RF only the three methods were uses as inputs.

When using MV, the inputs used had to be binary and predictions were made according to the predictions of each of the first phase classifiers. The final prediction will be the prediction given by the largest number of classifiers. Therefore, when calculating the AUC, it was done assuming a threshold that can be considered equivalent to 0.5.

To make a more accurate comparative study of the MV behaviour the AUCs of the different experiments were recalculated using a threshold of 0.5 and are presented at Table 5.6.

Table 5.6: Comparison of final prediction of the models with AUC calculated with fixed threshold =0.5.

	Input of the second stage	
Second Stage Method	**Binary**	**Probabilistic**
NN (5 input methods)	0.615 ±0.120	0.623 ±0.114
NN (3 input methods)	0.651 ±0.116	0.648±0.111
RF (3 input methods)	0.651 ±0.116	0.585 ±0.128
MV (5 input methods)	0.648 ±0.121	
MV (3 input methods)	0.651 ±0.117	



(a) Averaged ROC curve for LR.

(b) Averaged ROC curve for SVM.

(c) Averaged ROC curve for EN.

(d) Averaged ROC curve for NN.

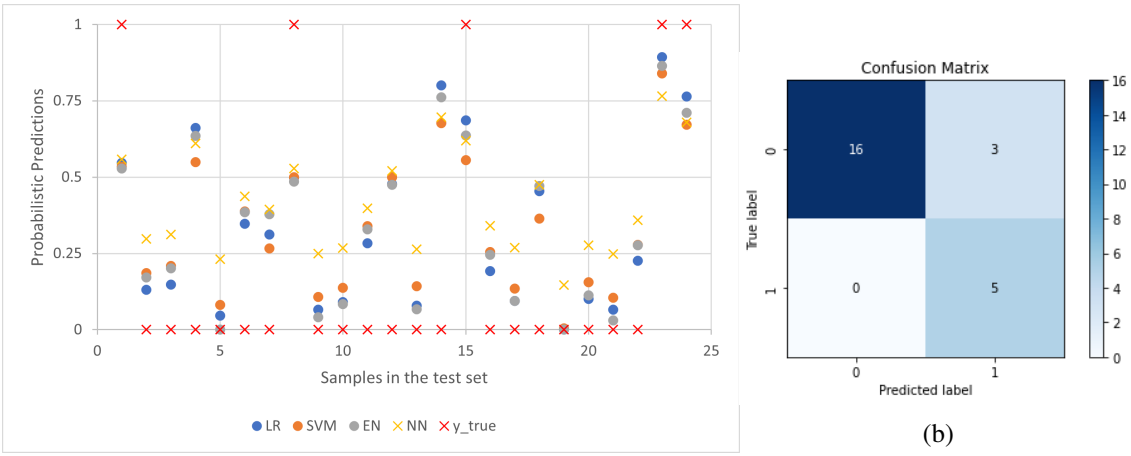
Figure 5.1: Averaged ROC curve obtained by the ensemble method from 50 runs. The blue line depicts the arithmetic average ROC curve and the shading the standard deviation.

There have been attempts to make some variations of the neural network by changing the activation loss and optimiser functions, the number of hidden layers and the number of neurons.We present only the results obtained by two of them, which showed promising results, but under the same conditions as the previous experiments, they were not superior.

The best result obtained corresponded to the model that used only three base learners (LR, SVM and EN) and that had as a combination method the neural network. The mean Receiver Operating Characteristic (ROC) curve was computed for all random data splits, being represented in Figure 5.1. The resultant curves show the similar performance obtained by the base learners and, as a consequence, the final classification. However, the result obtained AUC = 0.706 ($\pm$0.122) did not outperform the single models of the first state with AUCs of 0.712 ($\pm$0.119) to LR, 0.711 ($\pm$0.119) to SVM and 0.712 ($\pm$0.120) to EN. A possible reason that could explain this behaviour and that is suggested by the very close results of the single classifiers is that the methods misclassified the same examples causing the NN used for ensemble to also be induced to make this same mistake.

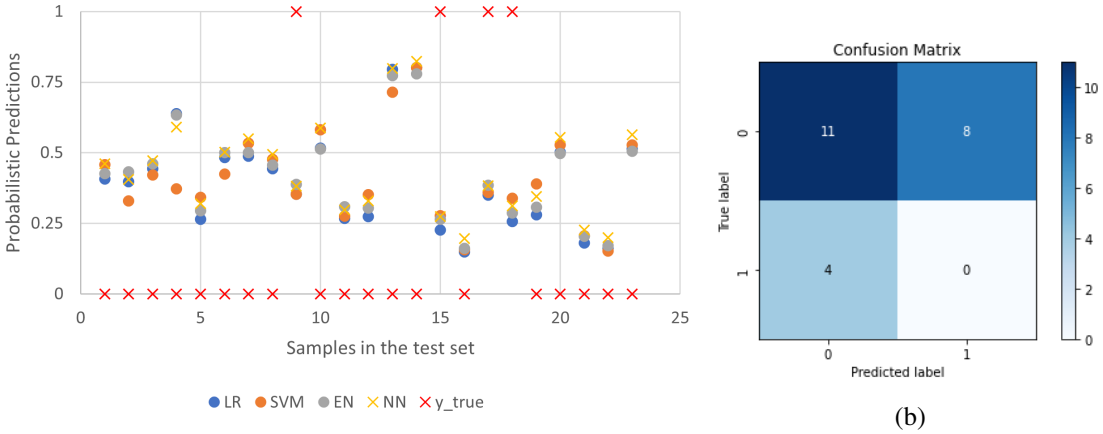
To analyse this hypothesis the best and the worst datatest predictions made by the best performing model were plotted in the graphics below Figure 5.2 and Figure 5.3, respectively, as well as their confusion matrices. In Figure 5.2 it can be observed that there is only a small variation in

the three baselearner methods represented by circles and that when they fail they fail in the same samples. The prediction made by NN follows that same pattern for using that methods prediction as inputs. In Figure 5.3, which represents the worst prediction, it is noticeable the poor performance of the EN method that predicts all values close to 0.5, affecting the NN prediction. The analysis of this figure confirms the great level of influence that a bad baselearner performance can have on ensemble methods.



(a) (b)

Figure 5.2: Detailed test prediction graphical representation (a) for the best performed datatest with NN model with 3 input probabilistic methods as combination rule and respective confusion matrix (b).



(a) (b)

Figure 5.3: Detailed test prediction graphical representation (a) for the worst performed datatest with NN model with 3 input probabilistic methods as combination rule and respective confusion matrix (b).

5.1 Summary

This chapter compiles the results of the different experiments and an analysis of them from which some conclusions can be drawn:

- The use of SMOTE in the model implemented at the beginning of both stages leads to higher performance which suggests the importance of the class distribution in the training data used.
- It was reinforced the idea of the great influence of the baselearners performances in the moments of replacement and elimination of the initial classifiers that presented the worst performance as well as in the particular case of the datatest analysis with worse predictions of the best model studied.
- The use of probabilistic forecasts as combination rule inputs in the ensemble method as a source of more information proved to be advantageous when using NN in the second stage but not for RF.

Chapter 6

Conclusions and Future Work

Lung cancer has a huge mortality rate. Therefore, progress must be made both in early diagnosis and in the choice of the most appropriate treatment to try to invert it. A factor to be taken into account when choosing the best treatment for a patient is the gene mutation status. One way to get this information is trough biopsy but this method presents some disadvantages. Biopsies are a process extremely invasive and often painful. For all of these reasons is very important to continue investing in the study of non-invasive methods to predict and be able to choose the best treatment.

The literature has shown that medical imaging is a potential alternative to biopsies. Some image features have shown to identify genomic alterations within tumour DNA, a field that is now called radiogenomics. Although most studies used the features that are related to the nodule, there is a percentage that shows that using external features of the nodule might improve the model's performance. Also, ensemble methods shows a big potential to overcome some barriers of the single models used previously in gene mutation prediction.

The best result obtained correspond to the model that used only three base learners (LR, SVM and EN) and that had as a combination method the neural network. However, the result obtained AUC = 0.706 ($\pm$0.122) did not outperform the single models of the first state with AUCs of 0.712 ($\pm$0.119) to LR, 0.711 ($\pm$0.119) to SVM and 0.712 ($\pm$0.120) to EN. A possible reason that could explain this behaviour and that is suggested by the very close results of the single classifiers is that the methods misclassified the same examples causing the NN used for ensemble to also be induced to make this same mistake.

Although the results have not surpassed the ML single methods, the ensemble approach still has great potential to be used in this field. Continuing to study this approach remains important. The use of other datasets, prediction of other mutated genes, more refined base learners and combination rules and different combinations of these last two components in the construction of other models opens the way for future work and research.

References

[1] Zhencong Ye, Yongmei Huang, Jianhao Ke, Xiao Zhu, Shuilong Leng, and Hui Luo. Breakthrough in targeted therapy for non-small cell lung cancer. *Biomedicine Pharmacotherapy*, 133:111079, 2021. URL: http://www.sciencedirect.com/science/article/pii/S0753332220312725, doi:https://doi.org/10.1016/j.biopha.2020.111079.

[2] Erminia Massarelli, Marileila Varella-Garcia, Ximing Tang, Ana C. Xavier, Natalie C. Ozburn, Diane D. Liu, Benjamin N. Bekele, Roy S. Herbst, and Ignacio I. Wistuba. Kras mutation is an important predictor of resistance to therapy with epidermal growth factor receptor tyrosine kinase inhibitors in non–small-cell lung cancer. *Clinical Cancer Research*, 13(10):2890–2896, 2007. URL: https://clincancerres.aacrjournals.org/content/13/10/2890, arXiv:https://clincancerres.aacrjournals.org/content/13/10/2890.full.pdf, doi:10.1158/1078-0432.CCR-06-3043.

[3] Madeleine Scrivener, Evelyn E.C. 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. doi:10.21037/tcr.2016.06.18.

[4] Rajat Thawani, Michael McLane, Niha Beig, Soumya Ghose, Prateek Prasanna, Vamsidhar Velcheti, and Anant Madabhushi. Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer*, 115(October 2017):34–41, 2018. URL: https://doi.org/10.1016/j.lungcan.2017.10.015, doi:10.1016/j.lungcan.2017.10.015.

[5] Mohamed Hosni, Ginés García-Mateos, Juan M. Carrillo-de Gea, Ali Idri, and José Luis Fernández-Alemán. A mapping study of ensemble classification methods in lung cancer decision support systems. *Medical and Biological Engineering and Computing*, 58(10):2177–2193, 2020. doi:10.1007/s11517-020-02223-8.

[6] Omer Sagi and Lior Rokach. Ensemble learning : A survey. (August 2017):1–18, 2018. doi:10.1002/widm.1249.

[7] Geewon Lee, Ho Yun Lee, Hyunjin Park, Mark L. Schiebler, Edwin J.R. 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. URL: http://dx.doi.org/10.1016/j.ejrad.2016.09.005, doi:10.1016/j.ejrad.2016.09.005.

[8] Wenxiao Jiang, Guiqing Cai, Peter C. Hu, and Yue Wang. Personalized medicine in non-small cell lung cancer: a review from a pharmacogenomics perspective. *Acta Pharmaceutica Sinica B*, 8(4):530 – 538, 2018. SI: Targeted Cancer Therapy. URL: http://www.sciencedirect.com/science/article/pii/S2211383518300030, doi:https://doi.org/10.1016/j.apsb.2018.04.005.

[9] S. E.D.C. Jorge, S. S. Kobayashi, and D. B.Jorge Costa. Epidermal growth factor receptor (EGFR) mutations in lung cancer: Preclinical and clinical data. *Brazilian Journal of Medical and Biological Research*, 47(11):929–939, 2014. doi:10.1590/1414-431X20144099.

[10] Stefania Rizzo, Francesco Petrella, Valentina Buscarino, Federica De Maria, Sara Raimondi, Massimo Barberis, Caterina Fumagalli, Gianluca Spitaleri, Cristiano Rampinelli, Filippo De Marinis, Lorenzo Spaggiari, and Massimo Bellomi. CT Radiogenomic Characterization of EGFR, K-RAS, and ALK Mutations in Non-Small Cell Lung Cancer. *European Radiology*, 26(1):32–42, 2016. doi:10.1007/s00330-015-3814-0.

[11] Gil Pinheiro, Tania Pereira, Catarina Dias, Cláudia Freitas, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P.Pinheiro Oliveira. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *bioRxiv*, 2019. doi:10.1101/794123.

[12] Hyun Su Kim, Kyung Soo Lee, Yoshiharu Ohno, Edwin J.R. Van Beek, and JuergenKim Biederer. PET/CT versus MRI for diagnosis, staging, and follow-up of lung cancer. *Journal of Magnetic Resonance Imaging*, 42(2):247–260, 2015. doi:10.1002/jmri.24776.

[13] Philippe Lambin, Emmanuel Rios-velazquez, Ralph Leijenaar, Sara Carvalho, Patrick Granton, Catharina M L Zegers, Robert Gillies, Ronald Boellard, and Lee Moffitt. HHS Public Access. 48(4):441–446, 2015. doi:10.1016/j.ejca.2011.11.036.Radiomics.

[14] Stefania Rizzo, Sara Raimondi, Evelyn E.C. de Jong, Wouter van Elmpt, Francesca De Piano, Francesco Petrella, Vincenzo Bagnardi, Arthur Jochems, Massimo Bellomi, Anne Marie Dingemans, and PhilippeRizzo Lambin. Genomics of non-small cell lung cancer (NSCLC): Association between CT-based imaging features and EGFR and K-RAS mutations in 122 patients—An external validation. *European Journal of Radiology*, 110(November 2018):148–155, 2019. URL: https://doi.org/10.1016/j.ejrad.2018.11.032, doi:10.1016/j.ejrad.2018.11.032.

[15] 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, Daniel L Rubin, Sandy Napel, and Ann N Geavert Leung. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Nature Publishing Group*, pages 1–8, 2017. URL: http://dx.doi.org/10.1038/srep41674, doi:10.1038/srep41674.

[16] Lan Song, Zhenchen Zhu, Li Mao, Xiuli Li, Wei Han, Huayang Du, Huanwen Wu, Wei Song, and Zhengyu Song Jin. Clinical, Conventional CT and Radiomic Feature-Based Machine Learning Models for Predicting ALK Rearrangement Status in Lung Adenocarcinoma Patients. *Frontiers in Oncology*, 10(March):1–14, 2020. doi:10.3389/fonc.2020.00369.

[17] WilberLiu Quispe-Tintaya. HHS Public Access. *Physiology & behavior*, 176(3):139–148, 2017. doi:10.1016/j.cllc.2016.02.001.Radiomic.

[18] Wei Zhao, Jiancheng Yang, Bingbing Ni, Dexi Bi, Yingli Sun, Mengdi Xu, Xiaoxia Zhu, Cheng Li, Liang Jin, Pan Gao, Peijun Wang, Yanqing Hua, and Ming Zhao Li. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Medicine*, 8(7):3532–3543, 2019. doi:10.1002/cam4.2233.

[19] Liwen Zhang, Bojiang Chen, Xia Liu, Jiangdian Song, Mengjie Fang, Chaoen Hu, Di Dong, Weimin Li, and Jie Zhang Tian. Quantitative Biomarkers for Prediction of Epidermal Growth Factor Receptor Mutation in Non-Small Cell Lung Cancer. *Translational Oncology*, 11(1):94–101, 2018. URL: https://doi.org/10.1016/j.tranon.2017.10.012, doi:10.1016/j.tranon.2017.10.012.

[20] Shuo Wang, Jingyun Shi, Zhaoxiang Ye, Di Dong, Dongdong Yu, Mu Zhou, Ying Liu, Olivier Gevaert, Kun Wang, Yongbei Zhu, Hongyu Zhou, Zhenyu Liu, and Jie Wang Tian. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. 2019. URL: http://dx.doi.org/10.1183/13993003.00986-2018, doi:10.1183/13993003.00986-2018.

[21] Isaac Shiri, Hasan Maleki, Ghasem Hajianfar, Hamid Abdollahi, Saeed Ashrafinia, Mathieu Hatt, Habib Zaidi, Mehrdad Oveisi, and Arman Shiri Rahmim. Next-Generation Radiogenomics Sequencing for Prediction of EGFR and KRAS Mutation Status in NSCLC Patients Using Multimodal Imaging and Machine Learning Algorithms. *Molecular Imaging and Biology*, 22(4):1132–1148, 2020. arXiv:1907.02121, doi:10.1007/s11307-020-01487-8.

[22] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman Hall/CRC, 1st edition, 2012.

[23] Robi Polikar Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–44, 2006. doi:10.1109/MCAS.2006.1688199.

[24] Kamal M. Ali and Michael J. Pazzani. Error reduction through learning multiple descriptions. *Machine Learning*, 24(3):173–202, 1996. doi:10.1007/bf00058611.

[25] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann N.C. Leung, Michael Kadoch, Chuong D. Hoang, Joseph Shrager, Andrew Quon, Daniel L. Rubin, Sylvia K. Plevritis, and Sandy Napel. Data descriptor: A radiogenomic dataset of non-small cell lung cancer. *Scientific Data*, 5:1–9, 2018. doi:10.1038/sdata.2018.202.

[26] Elmar Rendon-Gonzalez and Volodymyr Ponomaryov. Automatic lung nodule segmentation and classification in ct images based on svm. pages 1–4, 06 2016. doi:10.1109/MSMW.2016.7537995.

[27] S Gopal Krishna Patro and Kishore Kumar. Normalization : A Preprocessing Stage Normalization : A Preprocessing Stage. (May), 2016. doi:10.17148/IARJSET.2015.2305.

[28] J. Frade J. Mendes C. Freitas V. Hespanhol J. L. Costa A. Cunha H.P. Oliveira F. Silva, T. Pereira. The impact of interstitial diseases patterns on lung ct segmentation. Submitted In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Guadalajara, Mexico 31 October–4 November 2021.

[29] Joost J M Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G H Beets-tan, Steve Pieper, and Hugo J W L Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. 77(21):104–108, 2017. doi:10.1158/0008-5472.CAN-17-0339.

[30] A. Procházka, L. Gráfová, and O. Vysata. Three-dimensional wavelet transform in multidimensional biomedical volume processing. 2011.

[31] Sergei V Fotin, David F Yankelevitz, Claudia I Henschke, and Anthony P Reeves. A multi-scale Laplacian of Gaussian (LoG) filtering approach to pulmonary nodule detection from whole-lung CT scans. pages 1–16. arXiv:arXiv:1907.08328v1.

[32] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. *Degree Project Technology*, page 41, 2018. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426{%}0Ahttp://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf.

[33] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE : Synthetic Minority Over-sampling Technique. 16:321–357, 2002.

[34] Hristo Todorov, David Fournier, and Susanne Gerber. Principal Components Analysis : Theory and Application to Gene Expression Data Analysis. 4(2), 2018.

[35] Joana Morgado, Tania Pereira, Francisco Silva, Cláudia Freitas, Eduardo Negrão, Beatriz Flor de Lima, Miguel Correia da Silva, António J. Madureira, Isabel Ramos, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P.Morgado Oliveira. Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. *Applied Sciences (Switzerland)*, 2021. doi:10.3390/app11073273.

[36] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Xiao Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153:1–9, 2018. doi:10.1016/j.cmpb.2017.09.005.

[37] G S K Ranjan, Amar Kumar Verma, and Sudha Radhika. K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. (March), 2019. doi:10.1109/I2CT45611.2019.9033691.

[38] Mounir Hamza and Denis Larocque. An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8):629–643, 2005. URL: https://doi.org/10.1080/00949650410001729472, arXiv:https://doi.org/10.1080/00949650410001729472, doi:10.1080/00949650410001729472.