FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AI-based cancer characterisation using semi-supervised learning algorithms

Cláudia Patrícia Ferreira Araújo Pinheiro

Mestrado em Engenharia e Ciência de Dados

Supervisor: Hélder Filipe Pinto de Oliveira, PhD

Co-Supervisor: Francisco Carvalho Moreira da Silva, MSc

Co-Supervisor: Cláudia Freitas, MD

September, 2022

AI-based cancer characterisation using semi-supervised learning algorithms

Cláudia Patrícia Ferreira Araújo Pinheiro

Mestrado em Engenharia e Ciência de Dados

Approved in oral examination by the committee:

Chair: Prof. Carlos Soares
External Examiner: Prof. Catarina Silva
Supervisor: Prof. Hélder Oliveira

September, 2022

Abstract

Cancer is a broad term used to refer to a large group of related diseases characterised by abnormal cell growth and division. Among all forms of the disease, lung cancer is by far the leading cause of cancer-related deaths worldwide, being the late diagnosis one of the major factors responsible for this. Among all available treatments for this disease, targeted therapies have demonstrated to improve clinical outcomes and the overall survival rate. Nevertheless, for a patient to be a candidate to receive such treatments, cancer characterisation is needed to find the genomic alterations that drive each cancer. Traditionally, this characterisation is performed by biopsy; however, despite its unarguable usefulness, it is an invasive procedure and clinical complications may arise from it. Furthermore, with this procedure, only a sample of the tumour region is obtained, which may not capture its heterogeneity.

Computer-aided diagnoses, based on computed tomography (CT) scans analysis, have the potential to provide a reliable and less invasive tumour characterisation. This is possible by establishing a link between imaging phenotype and tumour genotype — this is the aim of radiogenomics, an emerging research field. The use of deep learning methods in medical imaging has been able to deliver promising results; however, the success of such models highly relies on large, properly annotated datasets. Acquiring this volume of labelled data in the medical domain is challenging since annotating medical images is a laborious, expensive, and time-consuming process. This difficulty is increased when the label to be collected concerns on the presence of mutations as clinicians cannot obtain this information directly from images, requiring additional exams to be performed. On the other hand, raw images, without the necessary annotations, are easier to obtain as their acquisition is already part of the clinical routine. This exposes the importance of exploring methods that could mitigate the labelled data scarcity problem by using both labelled and unlabelled data to improve the accuracy of predictive models.

This work investigated the ability to use a Semi-Supervised Learning approach to predict Epidermal Growth Factor Receptor mutation status in lung cancer, in a less invasive manner, using 3D CT images. To exploit the power of adversarial training, a combination of Variational Autoencoder and Generative Adversarial Network was used to incorporate labelled and unlabelled images. With the developed method, a mean AUC of 0.7011 was achieved with the best-performing model, where only 14% of data contained label. This SSL approach improved the discrimination ability by nearly 7 percentage points over a fully supervised model.

Keywords: Deep semi-supervised learning, radiogenomics, computed tomography, lung cancer

Resumo

Cancro é um termo amplo utilizado para se referir a um grande grupo de doenças relacionadas, caracterizadas por crescimento e divisão celular anormal. Entre todas as formas da doença, o cancro do pulmão é de longe a principal causa de mortes relacionadas com o cancro a nível mundial, sendo o diagnóstico tardio um dos principais factores responsáveis por isso. Entre todos os tratamentos disponíveis para esta doença, as terapias dirigidas têm demonstrado melhorar os resultados clínicos e a taxa de sobrevivência global. No entanto, para que um paciente seja candidato a receber tais tratamentos, é necessária uma caracterização do cancro para encontrar as alterações genómicas que impulsionam cada cancro. Tradicionalmente, esta caracterização é realizada por biopsia; contudo, apesar da sua indiscutível utilidade, é um procedimento invasivo podendo surgir complicações clínicas. Além disso, com este procedimento, apenas se obtém uma amostra da região tumoral, que pode não captar a sua heterogeneidade.

Os diagnósticos assistidos por computador, baseados na análise de Tomografias Computorizadas (TAC), têm o potencial de fornecer uma caracterização fiável e menos invasiva dos tumores. Isto é possível ao estabelecer uma ligação entre o fenótipo da imagem e o genótipo tumoral – este é o objetivo da radiogenómica, um emergente campo de investigação. A utilização de métodos de *deep learning* na imagiologia médica tem levado a resultados promissores; contudo, o sucesso de tais modelos depende em grande medida de grandes conjuntos de dados devidamente anotados. Adquirir este volume de dados etiquetados no domínio médico é um desafio, uma vez que anotar imagens médicas é um processo trabalhoso, dispendioso e demorado. Esta dificuldade aumenta quando a anotação a recolher corresponde a mutações, uma vez que os clínicos não conseguem obter esta informação diretamente das imagens, exigindo a realização de exames adicionais. Por outro lado, as imagens em bruto, sem as anotações necessárias, são mais fáceis de obter uma vez que a sua aquisição já faz parte da rotina clínica. Isto expõe a importância de explorar métodos que possam mitigar o problema da escassez de dados com anotações, utilizando dados com e sem as respetivas anotações para melhorar a precisão dos modelos preditivos.

Este trabalho investigou a capacidade de utilizar uma abordagem de Aprendizagem Semi-Supervisionada para prever o estado de mutação do Recetor do Factor de Crescimento Epitelial no cancro do pulmão de uma forma menos invasiva, utilizando imagens 3D de TACs. Para explorar o poder do treino adversarial, foi utilizada uma combinação de *Autoencoder* Variacional e Rede Adversarial Generativa para incorporar imagens com e sem anotações. Com o método desenvolvido, foi alcançada, com o modelo com melhor desempenho, uma AUC média de 0.7011 onde apenas 14% dos dados continha etiqueta. Esta abordagem semi-supervisionada melhorou a capacidade descriminativa em wuase 7 pontos percentuais em relação a um modelo totalmente supervisionado.

Palavras-Chave: Aprendizagem semi-supervisionada profunda, radiogenómica, tomografia computorizada, cancro do pulmão

Acknowledgements

This dissertation is a significant milestone in a new and important cycle in my life. All the work developed during this project and throughout the last two years would not have been possible without the collaboration of several people.

First, I am extremely grateful to my supervisor team, Hélder, Francisco and Tânia, for the continuous support, advice and patience even in the most critical moments. Week after week, you were always there to answer my many questions and encourage me, giving me the right words when I most needed them.

To my boyfriend, Jorge, for encouraging me to take this step forward and supporting me during this challenging journey. Thank you for understanding and accepting all the sacrifices we had to make to turn this possible.

Special thanks to my close family and friends for their support when I decided to enrol for this degree to make a career change and, above all, for understanding my lack of time during this period.

Last but not least, I would like to thank the amazing group I found in the MDSE. I could have never expected to meet such incredible people. This was a difficult adventure, but all the mutual help, friendship and continuous support enabled us to overcome the obstacles. I cannot imagine having undertaken this journey without you! A special thank you to two friends: Miguel, for always encouraging me and for the countless hours of study and debate during these two years, and Ana, with whom I shared so many projects.

Cláudia Pinheiro

"Hard work is only a prison sentence when you lack motivation."

Malcolm Gladwell, *Outliers: The Story of Success*

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial intelligence
ALK	Anaplastic Lymphoma Kinase
ANN	Artificial Neural Network
AUC	Area Under the Curve
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
DNA	Deoxyribonucleic Acid
EGFR	Epidermal Growth Factor Receptor
GAN	Generative Adversarial Network
HU	Hounsfield Unit
KL	Kullback–Leibler
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NLST	National Lung Screening Trial
NSCLC	Non-small Cell Lung Cancer
PET	Positron Emission Tomography
ReLU	Rectified Linear Unit
RF	Random Forest
ROC	Receiver Operating Characteristics
ROI	Region of Interest
SCLC	Small Cell Lung Cancer
SSL	Semi-Supervised Learning
SVM	Support-Vector Machines
TKI	Tyrosine Kinase Inhibitors
UHCSJ	University Hospital Center of São João
VAE	Variational Autoencoder
XGBoost	Extreme Gradient Boosting
2D	Two-dimensional
3D	Three-dimensional

Chapter 1

Introduction

According to the 2019 report of the World Health Organization [5], trachea, bronchus and lung cancer deaths are ranked as the 6^{th} leading cause of death worldwide. Although it was not the most common cancer in terms of new cases in 2020, lung cancer was, by far, the most lethal cancer in the same year [2]. This high mortality is due to the difficulty in diagnosing this kind of cancer at an early stage, with more than half of the diagnoses made when the disease is already in a metastatic phase. In this stage, the 5-year survival rate (percentage of people who are still alive five years after they are diagnosed with the disease) is little more than 5% [42].

It is true that many smokers never develop this disease and that an increasing number of non-smokers are diagnosed with lung cancer yearly. Nonetheless, smoking remains the primary risk factor for developing this ailment, with 80% of the individuals diagnosed being either current or past smokers.

Lung cancer can be divided into two main histological types according to clinical differences in cells aspect under the microscope, metastatic spread, and response to therapy: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), respectively accounting for about $80-85\%$ and $15-20\%$ of the cases [64, 75]. The NSCLC can be subdivided into histological subtypes, being the most frequents adenocarcinoma, squamous cell carcinoma and large cell carcinoma. The 5-year survival rates for this kind of cancer depend on the type and stage of the disease, with NSCLC having the highest rates for the same stage when compared to SCLC.

1.1 Motivation

When diagnosed at an early stage, lung cancer can present a favourable prospect, depending on its type, with 5-year survival rates around 60% for localised cancer — limited to the lungs; however, early-stage detection remains challenging, with more than half of all cases being diagnosed when cancer has already spread to other organs, with a corresponding 5-year survival rate of only 6% [55]. For this reason, it is crucial to work on individualised treatments according to lung cancer type and stage, leaving behind the traditional approaches relying almost exclusively on chemotherapy and radiotherapy treatments for patients with advanced disease.

Targeted therapies have emerged as a strategy to improve the outcome of lung cancer, especially NSCLC, improving patient survival. They consist of treatments based on a new drug type that blocks the growth and survival of cancer cells by acting on specific proteins and genes. Some gene mutations linked to lung cancer have already been identified, being Epidermal Growth Factor Receptor (*EGFR*) and Kirsten Rat Sarcoma Viral Oncogene Homolog (*KRAS*) the most common ones. *EGFR* mutated is present in around one-third of NSCLC patients, with a higher prevalence among non-smoker patients and women [73]. On the other hand, *KRAS* — which can be found in about one out of four NSCLCs — occurs more frequently in smokers than never-smokers [47]. For their prevalence, they are important biomarkers to be identified, although only *EGFR* has approved targeted therapies. Tyrosine kinase inhibitors (TKIs) are one of the most well-known targeted therapies that treat different *EGFR* mutations. They recognise and attack cancer cells without harming the surrounding healthy cells too much, slowing down the progression of the disease and improving quality of life. Hence, these treatments are bringing new hope to lung cancer patients [68]. Consequently, assessing *EGFR* mutation has become a determinant step when deciding on the possible treatments for each individual, enabling more effective patient management in precision medicine.

EGFR mutations are usually detected using DNA extracted from tumour tissue samples obtained during biopsy or resection; however, this method has some limitations considering that it is an invasive procedure with clinical implications. Additionally, it may not always be possible to perform due to several factors, such as the existence of co-morbidities and the inaccessibility of the tumour [21]. Alternatively, cytologic samples and liquid biopsy — by extracting body fluids, usually blood — can be gathered, considering that the methods used to obtain them are usually less invasive. Even so, both present some challenges: cytologic samples can have a low tumour content, and liquid biopsy lacks standardisation and requires sensitive techniques [21, 40].

In this context, the need to find alternative non-invasive methods to determine gene mutation status arises. Computer-aided diagnosis (CAD) systems can play an essential role in this assessment. These systems allow clinicians to have more information — often not accessible to the human eye — to support decision-making. Therefore, the analysis of medical images such as computed tomography (CT) scans may be the key to overcoming the aforementioned problem. Medical images have already proven to be able to provide valuable information on the understanding of biological characteristics of cancer, and on the tumour genomic profiling [8, 24, 25]. Moreover, some publications have highlighted and revealed the connection between *EGFR* mutation status and CT scans imaging phenotypes [13, 19, 44], using supervised approaches. By establishing this link, some light is shed on a less invasive way of identifying mutations driving cancer.

However, studies so far were limited to the small size of the available datasets with the *EGFR* mutation information. More robust and reliable models could be developed if more data were used. Despite the availability of larger datasets with CT scans, they lack the intended labels. This is a recurrent problem in medical imaging analysis due to the evident difficulty in collecting such annotations — the process is expensive, time-consuming, and many times and many times requires

additional invasive exams, as is the case when collecting the label regarding mutations that drive cancer. For this reason, the usage of Semi-Supervised Learning (SSL) techniques, which make use of a combination of labelled and unlabelled data, going further than traditional supervised approaches, comes as a solution to overcome the problem of scarcity of annotated data and might enhance the predictive abilities of the models.

1.2 Objectives

This dissertation aims to contribute to supporting medical decision-making in the use of targeted therapies by developing a method for lung cancer characterisation, using thoracic CT images to predict a gene mutation status. By using an SSL approach, it is expected to create more robust predictive models by taking advantage of a broader set of data and using information that unlabelled data is able to provide.

Exploiting the power of adversarial training, the approach presented in this work consists of combining a Variational Autoencoder and a Generative Adversarial Network, intending that the features extracted from unlabelled data to discriminate images can help in the classification task. This method is expected to significantly reduce the necessary labelled data required to train such a classification model.

1.3 Contributions

This dissertation includes the following contributions:

- Assessment of gene mutation status in lung CT scans, using a Semi-Supervised Learning approach;
- Comparison of the obtained model with a fully supervised learning method — baseline model;
- Method for an end-to-end identification of important biomarkers status in CT images, incorporating both labelled and unlabelled data.

1.4 Document structure

This document comprises seven chapters, as follows. The present chapter is an introduction to the work, containing the context and the motivation for the problem under study, as well as identification of the objectives and a list of the contributions of the project. Chapter 2 details some relevant clinical concepts required for a better understanding of the described problem, including a more comprehensive description of lung cancer, computed tomography, targeted therapies, and mutation testing. Chapter 3 details some relevant works regarding Artificial Intelligence-based

solutions in radiogenomics and the use of Semi-Supervised Learning approaches in medical imaging. Additionally, in this chapter, some basic concepts about Semi-Supervised Learning are briefly described. Chapter 4 provides a description of the datasets used in this work as well as an explanation of the steps required in data preparation. In Chapter 5 the methodology employed is explained, and a detailed description of the experiments conducted is provided. Chapter 6 presents the results obtained with the developed model, including some important comparisons and a discussion of the outcomes. Lastly, Chapter 7 summarises the main findings of the work done in this dissertation alongside any shortcomings and some considerations for future research works.

Chapter 2

Background

Often referred to as "the disease of the century", cancer is not a modern ailment. Earliest evidence of cancer in humans were found in ancient Egyptian manuscripts over 3500 years ago, and the first description of this disease traces back to 3000 BC. It was only around 400 BC that it was first called "cancer" by the Greek Hippocrates, known as "the father of Medicine". When describing tumours, Hippocrates used the words *karkinos* and *karkinoma*, which in Greek mean *crab*. It is thought that these names were chosen as the spread of a malignant tumour brings to mind a crab with its many legs [22, 1].

Nowadays, in this era of major technological advances, the ever-increasing understanding of cancer is making this field of research a promising area and one of the most rapidly evolving fields of modern medicine. In fact, cancer is not a single and specific disease but rather a broad term given to a group of more than one hundred related diseases, with some forms being more frequent than others. Breast and lung cancer were the most common types in 2020, accounting for 12.5% and 12.2% of all new cases, respectively. Despite figures concerning mortality also varying across the different forms of the disease, lung cancer is the leading cause of cancer-related deaths, with 18.2% of all cancer deaths in the same year [2].

This chapter intends to provide a clinical overview of the proposed work, covering some important key concepts. It comprises a brief introduction to the anatomy of the lungs (Section 2.1) and a more detailed description of lung cancer (Section 2.2), computed tomography (Section 2.3), targeted therapy (Section 2.4), and mutation testing (Section 2.5).

2.1 Anatomy of the lungs

The lungs are part of the respiratory system, and their principal function is to supply oxygen from the atmosphere to the bloodstream, which takes it to every cell in the human body. In there, it is exchanged for carbon dioxide that is transported to the lungs to be removed from the circulation and the body [63].

Lungs are paired and located in the thorax — chest cavity — and the left one, due to the presence of the heart, is smaller than the right one. At the same time, the right lung is shorter so that the liver can be accommodated underneath it. These organs are divided into lobes, having the right lung three lobes — upper, middle, and lower — whereas the left lung has only two — upper and lower, as represented in Figure 2.1.

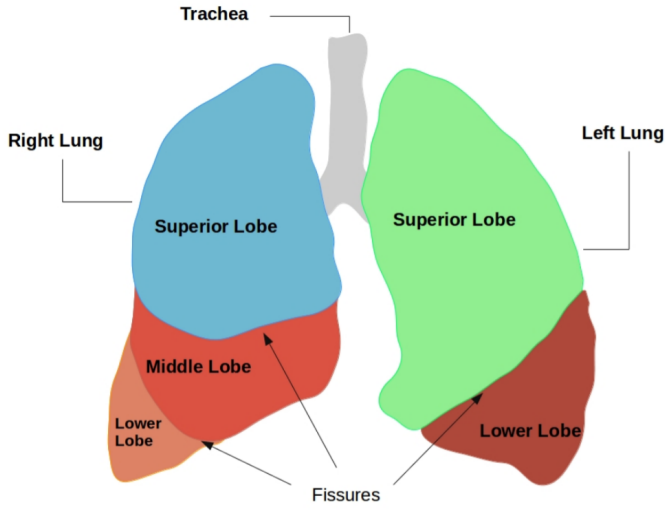


Figure 2.1: Representation of the lungs with the respective lobes. From [23].

2.2 Lung cancer

The human body is composed of trillions of cells that divide to originate new ones as the body needs them. In a typical process, when cells are damaged or become too old, they die, and new cells replace them. However, some modifications in the DNA can occur and disturb this control mechanism, leading cells to frantic growth and multiplication, usually forming a mass of tissue called a tumour, which can be classified as benign or malignant — cancerous tumour. However, some cancer types do not create a tumour, as is the case of leukaemia and most types of lymphoma.

It is called primary tumour to the initial cancer tumour, and metastases occur when the cancer cells spread to adjacent tissues and other organs. The metastases are considered the primary cause of cancer mortality since metastatic tumours are not only difficult to treat due to their anatomically dispersed distribution in different organs but, typically, they are also resistant to cytotoxic agents [28, 45]. In lung cancer, the primary tumour is located in the lungs, and two main broad types, with characteristic behaviour, can be observed:

- **Non-small cell lung cancer:** accounts for circa $80-85\%$ of all cases and can be subdivided into three major subtypes:
  - **Adenocarcinoma:** is the dominant histological subtype of lung cancer, accounting for almost half of all lung cancers and 60% of NSCLC cases. It usually arises in the lung

periphery and is associated with better clinical outcomes. There has been an increased incidence of adenocarcinoma over the last decades [75, 53];

  - **Squamous cell carcinoma:** this subtype is the most strongly associated with a smoking history. Although it is most commonly detected in the central part of the lungs, an increasing number of peripheral tumours have been observed. The incidence of this subtype of NSCLC has been decreasing during recent decades [75, 53];
  - **Large cell carcinoma:** represent a minority of all NSCLC cases and is commonly peripherally located [75].

- **Small-cell lung cancer**: accounts for $15-20\%$ of all lung cancers and is typically caused by tobacco smoking. Most cases are diagnosed at an advanced stage [75].

Besides their cells outlook, these disease subtypes have a different evolution: SCLC has a faster spread, being more aggressive; NSCLC is usually less aggressive, growing at a slower pace. Nevertheless, the latter has a drawback: it is typically identified at a later stage. This late diagnosis holds obvious implications: the 5-year survival rate for patients with localised NSCLC is 63%, whereas for the metastatic phase is only 7% [42]. However, a new generation of treatments has paved the way to increase these figures by acting directly on specific molecular targets — targeted therapies. This is possible by identifying the biomarkers associated with cancer growth, progression, and metastasis [38].

2.3 Computed Tomography

Screening, initial diagnosis and tracking of lung cancer are performed using imaging tools such as X-rays, CT, magnetic resonance imaging (MRI) and positron emission tomography (PET) scans. For providing detailed images with a simple procedure, CT is usually the most used technique, being considered the gold standard for lung nodule detection [56, 43].

CT scans combine data from multiple X-ray images captured from different angles to generate detailed images of tissues inside the body. It produces cross-sectional images, called slices, that can be used to construct a 3D image, as illustrated in the example of Figure 2.2. As different body parts absorb X-rays in different amounts according to their density, the fundamental principle behind CT is that the density of the tissue can be measured from the calculation of the attenuation coefficient of the X-rays which pass through [12].

The density values are measured in a quantitative scale known as the Hounsfield scale (named after Godfrey Hounsfield, one of the developers of the CT technique). On such a scale, one has reference values: at standard pressure and temperature, the air is -1000 Hounsfield Units (HU), and water is 0 HU [12]. The HU comprise the grayscale in CT images. Thus, tissues like air, which have little attenuation, appear black on the X-ray, whereas bones, which have high attenuation, appear white; soft tissue, like the heart or liver, are displayed in shades of grey in a CT image.

Figure 2.2: Illustrative example of a CT slice. From the NSCLC-Radiogenomics dataset [11].

2.4 Targeted therapies

Once a lung cancer diagnosis is confirmed, and depending on its type and stage, common treatment options include surgery, radiotherapy, chemotherapy and targeted therapy. Unlike traditional radiotherapy and chemotherapy, which are indiscriminative about their target (they not only kill cells responsible for cancer but also destroy healthy cells), targeted therapy is based on drugs that act with greater precision on the intended cells. This is an emerging trend regarding the drug delivery methodologies to treat lung cancer for its ability to avoid harming normal cells. As a result of their specificity, targeted therapies allow to improve the outcomes and diminish the side effects associated with the treatment [52].

This targeted action is possible by identifying the genomic alterations driving each cancer. It can be predicted whether a patient would benefit from treatments with specific pharmacological agents by assessing the presence of particular mutations. This is the base of precision medicine that aims to provide every single patient with the proper cancer treatment at the correct dose [51]. Consequently, testing for genetic mutations is an essential part of diagnosing patients with many types of cancer, such as lung cancer. Several oncogenes can be evaluated, being *EGFR* and Anaplastic Lymphoma Kinase (ALK) the most common targets for mutations due to their prevalence and having approved targeted therapies [72]. Although many mutations are present in different cancer types, *EGFR* mutations are specific for lung cancer [10] and have the most effective targeted therapies.

2.5 Mutation testing

Being considered a critical step in diagnosis standards, molecular testing for important biomarkers can be done with different clinical approaches, depending on clinical criteria and tissue availability. When the tumour tissue is easily available, the biopsy is the gold standard for detecting these mutations status. During this procedure, tumour tissue samples are collected so that DNA that allows the direct genetic sequencing can be extracted [21]. However, several biopsies may be

needed to achieve an accurate diagnosis due to the dynamic nature of cancer that leads to tumour heterogeneity — biomarkers are not evenly present in all cancer cells [49]. As a result, and since it is an invasive procedure associated with severe pain for patients and an inherent risk of complications, it might not be the best option. Furthermore, as many patients are only diagnosed when the disease is already in an advanced stage (as is the case for lung cancer), the biopsy may not be feasible due to clinical limitations, such as co-morbidities [21].

Less invasive methods have been developed and gained ground in the last years to overcome the above-mentioned limitations, namely liquid biopsy and cytologic samples. Liquid biopsy uses non-solid biological tissue (typically blood), instead of tumour tissue, to assess DNA that has been shed from tumour cells. Consequently, this is a non-invasive procedure, easily repeatable, and quicker. This promising technology has the potential to capture tumour heterogeneity. However, it also has shortcomings: it lacks standardised protocols, requires very complex and costly technologies, and has lower sensitivity and precision in comparison with tumour tissue [49]. Cytologic sample consists in analysing the cells from bodily tissues and fluids collected in a minimally invasive and rapid procedure. Although a smaller sample is required with this method, it has some drawbacks, such as the low tumour content [68].

2.6 Summary

Targeted therapies have emerged as a standard treatment for patients with NSCLC, improving their clinical outcomes and the overall survival rate. To be eligible for target therapy, patients must have a targetable driver mutation and, therefore, biomarkers testing plays an essential role in precision medicine. Biopsy has long been the standard for both the confirmation of a cancer diagnosis and assessment of gene mutation status. However, as detailed in this chapter, this is an invasive procedure, and several complications might arise from it.

In light of the above, computer-aided diagnosis, based on CT scans analysis, arises as an option to provide a trustworthy and less invasive tumour characterisation. These systems, designed to aid clinicians in the decision-making process, are being increasingly adopted in the medical field, representing a paradigm shift in healthcare. Since CT is usually performed as a preliminary step toward lung cancer diagnosis, these scans can be used to provide further information. From these medical images, a large number of valuable features can be extracted — radiomic features —- that have been proven to allow the identification of alterations within tumour DNA [59]. In this context, radiogenomics refers to the identification of molecular biology behind these imaging phenotypes by integrating the study of data from medical images and the genomic scales [74, 59].

Chapter 3

Literature Review

Artificial intelligence (AI), regarded by many as a threat, does not intend to replace clinicians but rather to provide them with additional and accurate tools. AI has the potential to assist health professionals, helping them in time-consuming tasks like image analysis and decision support systems. Leveraged by an increasing amount of available data and easier access to ever greater computing capacity, AI represents a paradigm shift in healthcare.

In this chapter, a selection of relevant studies regarding AI-based solutions in radiogenomics (Section 3.1) and Semi-Supervised Learning approaches in medical imaging (Section 3.2) is presented and described. Section 3.3 comprises an overview with some conclusions that can be drawn from the analysis carried out.

3.1 AI-based solutions in radiogenomics

Radiogenomics, a combination of the words "radiology" and "genomics", aims to establish associations between medical imaging and gene expression patterns [65, 54]. This new ongoing field of research intends to understand how molecular characteristics of particular diseases are reflected in imaging phenotype. Figure 3.1 presents an illustration of typical research methods of radiogenomics. The establishment of such associations has the potential to improve diagnostic methods for precision cancer care. However, the alterations induced by genetic mutations can be very challenging, if not impossible, to be perceived with the human eye [65].

In this context, AI has proven to be a very powerful tool, equipping clinicians with methods to uncover patterns in imaging data such as CT scans and MRIs. Many studies, in a few cancers, tried to establish a link between these imaging phenotypes and the underlying mutations. In one of the most well-documented areas, Choi *et al.* [16] presented a model to assess the isocitrate dehydrogenase (IDH) mutation status of gliomas — the most common type of brain tumour — using preoperative MRIs. To this end, an approach based on Convolutional Neural Networks (CNNs)

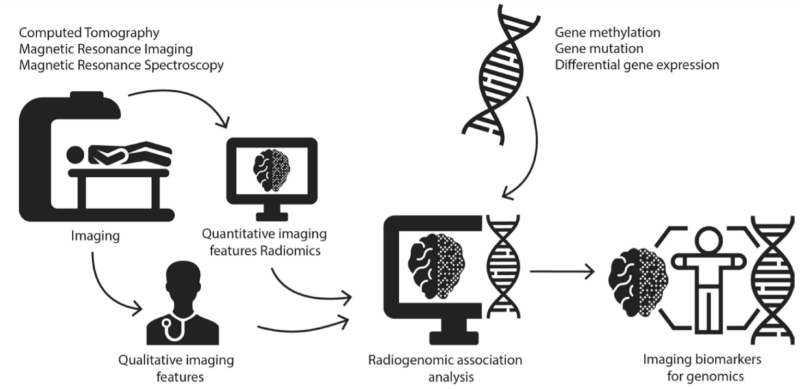


Figure 3.1: Typical research methods of radiogenomics. From [34].

was used, both for automated tumour segmentation and to classify IDH status. The classifier integrated information from 2D tumour signal intensity, 3D tumour shape and location, and age. Three patient cohorts were used, leading to classification performances ranging from 0.86 to 0.96, using the area under the receiver operating characteristic (ROC) curve (AUC) as the evaluation metric.

A similar study was performed before by Chang *et al.* [14] using a residual neural network, a modified version of the well-known ResNet, originally proposed by He *et al.* [30]. The implemented model achieved an accuracy of 0.857 and AUC of 0.94 within the testing set. When age at diagnosis was incorporated into the model, the predictive performance increased to 0.891 and 0.95 with regard to the accuracy and AUC, respectively.

Based on the same architecture, He *et al.* [29] also used ResNet, a pre-trained model, to estimate the mutation status of *KRAS* oncogene in colorectal cancer patients, using contrast-enhanced CT images in three positions: axial, coronal, and sagittal directions. Also, an evaluation in terms of predictive performance between models in which only the region of interest (ROI) was used and holistic ones, where surrounding areas were also added, was made. Due to the limited size of the available dataset, data augmentation in the training set was performed by introducing random rotations and translations. The higher AUC of 0.90 was achieved by the ResNet model in the axial direction, increasing to 0.93 when the surrounding tumour area was also provided as input. In their study, the authors also compared the performance of this DL method and a radiomics model with Random Forest (RF) as classifier. To develop this last model, features from the primary tumours were extracted based on the manually delineated ROI. Results showed that the DL model had better predictive capacity since the radiomics model could only achieve an AUC of 0.818.

Motivated by the possibility of improving outcomes of patients with kidney cancer through personalised treatments, Kocak *et al.* [36] developed models to predict the mutation status of the gene encoding the protein polybromo-1 (PBRM1) in patients with clear cell renal cell carcinoma (RCC). The study aimed to evaluate the predictive value of ML-based high-dimensional quantitative CT texture analysis. To this end, two different classifiers were used — Artificial Neural

Network (ANN) and RF — and data augmentation was applied to build stable ML models and balanced classes. The highest accuracy and AUC were observed using the RF model (0.950 and 0.987, respectively).

Regarding lung cancer, due to the importance of molecular profiling in such a deadliest disease, as already mentioned in Chapter 2, several studies have been conducted, with a predominance of those aiming to predict *EGFR* mutation. Pinheiro *et al.* [44] investigated the possibility of predicting *KRAS* and *EGFR* mutation status using different combinations of input features, encompassing clinical, radiomic (which described only the nodule), and semantic features (qualitative imaging features defined by radiologists during their evaluation). The latter were further divided into features describing only the tumour, features describing lung structures other than the nodule, and hybrid semantic futures (which comprises the previous two sets). To develop the work, the radiomic features were extracted from CT scans from the NSCLC-Radiogenomics dataset [11], and the Extreme Gradient Boosting (XGBoost) algorithm was applied as the classifier, which turned possible to determine the most relevant features in the construction of the boosted decision trees. The attained results evidence the likely association between *EGFR* mutation status and CT scan imaging phenotypes, being the best and worst-performing models, respectively, the one that used hybrid semantic features and the model that received as input only radiomic features extracted from the nodule, recording averaged AUCs of 0.7458 ± 0.0877 and 0.5797 ± 0.1238. Additionally, when evaluating the importance scores of the used features, the most relevant ones for the classification problem were tumour external. Regarding the *KRAS* prediction, no satisfactory model could be reached.

Having shown the importance of a holistic lung analysis, Morgado *et al.* [41] presented an approach extending the latter, using the same dataset. *EGFR* mutation status was assessed using radiomic features extracted from the entire volumetric region of the lung containing the tumour instead of focusing on the nodule region only. Additionally, a combination of six ML algorithms and feature selection methods was employed. The model comprising the Linear Support-Vector Machines (SVM) classifier and feature selection performed with Principal Component Analysis (PCA) with 70% explained variance achieved the highest averaged AUC — 0.737 ± 0.018. Again, with these results, it was shown that better results could be achieved when assessing *EGFR* mutation status using radiomic characteristics from an extended ROI, comprising the entire lung containing the nodule. Despite these outcomes, a common limitation is mentioned by the authors of both studies: the small sample size. The utilisation of a larger dataset so that more robust models can be obtained is pointed out as a future direction.

Wang *et al.* [67] presented an end-to-end pipeline, based on DL, towards identifying *EGFR* mutation status in CT scans. The proposed method, which is not based on feature engineering, instead of using an accurate tumour segmentation or even features manually extracted, requires only as input the tumour region in a CT image, previously manually identified. The proposed approach makes use of the ability of DL models to automatically learn feature representations relevant for each task, in this case, *EGFR* mutation-related features to predict whether a tumour is *EGFR*-mutant or wild-type. To perform this study, 844 patients who met well-defined inclusion

criteria were selected from two independent hospitals in China. The patients from each hospital were divided into two cohorts: 603 patients from one hospital were included in the primary cohort (a total of 14926 CT images from this cohort were used for training), and 241 patients from the other hospital made part of the validation cohort. The applied pipeline of the *EGFR* mutation status prediction can be seen in Figure 3.2 and comprises two subnetworks. The first one shares the same structure with the first 20 layers in DenseNet, firstly introduced by Huang *et al.* [31], with weights acquired from the ImageNet dataset [48] in a transfer learning manner. The second subnetwork is composed of four convolutional layers and was trained with a dataset consisting of nearly 15000 CT images to identify the *EGFR* mutation status. With this DL model, AUCs of 0.85 and 0.81 and accuracy of 0.7702 and 0.7386 were obtained in the primary cohort by 5-fold cross-validation and in the validation cohort, respectively. These similar outcomes seem to support the generalisation capability of the model in predicting the *EGFR* mutation status of new patients.

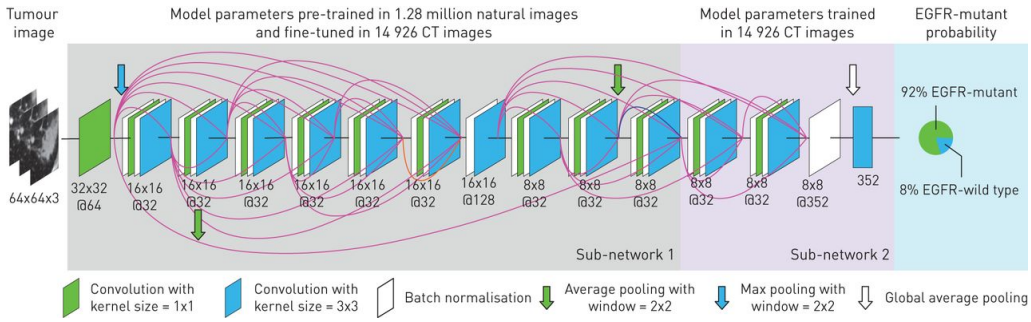


Figure 3.2: Pipeline of the *EGFR* mutation status prediction model proposed by Wang *et al.*. From [67].

3.2 Semi-supervised approaches in medical imaging

DL is revealing to be a promising and trendy field in medical imaging; however, the success of its models usually relies on large, properly labelled datasets to fine-tune the parameters. Such labelled data is generally difficult to acquire, particularly in the medical field, considering that the imaging annotation process is highly time-consuming, complex, and even subjective, requiring some expertise from the clinicians. In contrast, unlabelled data are typically easier to obtain. Therefore, applying semi-supervised learning approaches may be one valuable solution to overcome this problem by using both labelled and unlabelled data. Hence, SSL can be seen as being halfway between unsupervised and supervised learning. Figure 3.3 presents a general scheme of a SSL approach.

"If sufficient unlabelled data is available and under certain assumptions about the distribution of the data, the unlabelled data can help in the construction of a better classifier" [66]. The mentioned assumptions draw attention to essential requirements about the data distribution — illustrated in Figure 3.4 — so that the unlabelled data can be helpful for the models:

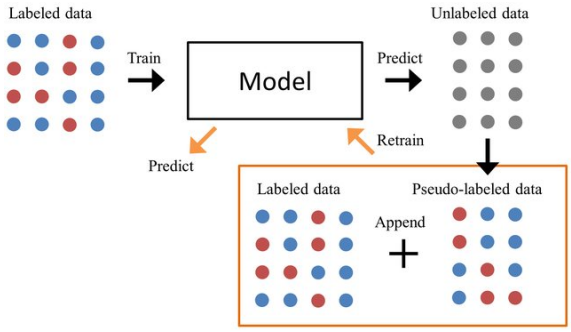
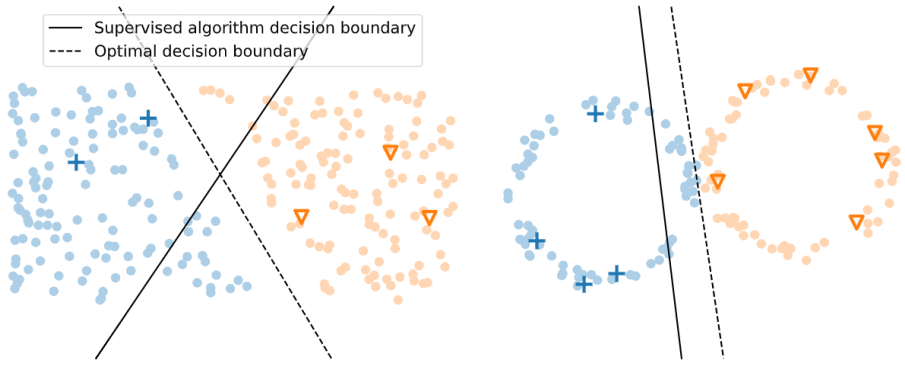


Figure 3.3: Semi-supervised learning general overview. From [34].

- **Smoothness assumption —** if two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2 [15]. In other words, data points located close to each other in data space are likely to share the same label;
- **Cluster assumption —** if points are in the same cluster, they are likely to belong to the same class. This assumption is many times referred to as low-density assumption and can be formulated in an equivalent form: "the decision boundary should lie in a low-density region" [15];
- **Manifold assumption —** data points on the same low-dimensional manifold should have the same label [66].



(a) Smoothness and low-density assumptions. **(b)** Manifold assumption.

Figure 3.4: Illustrations of the semi-supervised learning assumptions. From [66].

In the last years, several SSL methods have been proposed, and the most common ones are usually grouped in the following categories:

- **Self-training —** an incremental algorithm and one of the most straightforward techniques. Initially, one single supervised classifier is trained with annotated data only. The classifier is then iteratively used to predict labels for the unlabelled data, being its own highly confident

predictions added to the labelled training dataset and then retrained. This process is repeated until some termination criterion is met (such as no more unlabelled data remain) [66];

- **Co-training —** an extension of self-training to multiple classifiers. In this case, in each iteration, the most confident predictions of each classifier are added to the labelled dataset of the other classifiers [66];
- **Generative models —** act by modelling the process that generates the data, estimating the probability of any instance belonging to each class. These models learn the inherent features of data to better model data distributions [71];
- **Graph-based methods —** represent data as a graph, connecting similar examples, in a way that the label of unlabelled observations can be inferred from it. In this graph, all the observations are represented by nodes, and the weighted edges that connect them measure the similarity between each pair [60]. With this information, labels from annotated data can be propagated through the graph. These algorithms are called transductive algorithms, given that they only predict the labels of the unlabelled examples, not inferring any general decision rule [15];
- **Consistency regularisation —** these models are also known as perturbation-based and are usually used with neural networks. They encompass methods in which data points are perturbed with a small amount of noise, encouraging the predictions for both the original and the noisy data to be similar [71]. One of the most frequent and successful methods of this type that can be found in the literature is the Teacher-Student model.

Some studies have already applied these approaches in medical imaging in order to deal with the small proportion of labelled data in the training datasets, with a greater focus on segmentation tasks.

Martins & Silva [39] recently used a teacher-student-based pipeline for semi-supervised learning in the classification of chest X-ray images, based on a previous work of Yaniz *et al.* [70]. The used method consists of two steps: firstly, a teacher model is trained only with the labelled data; secondly, the student model is trained to reproduce the best outputs of the teacher. A DenseNet network was utilised for both student and teacher models. This work compared the developed method and a fully supervised approach for different proportions of labelled data. Analysing these results, a significant increase in performance was achieved when using small labelled datasets: when only 2% of the data had label, an AUC of 0.8222 was obtained, which increased to 0.879 when the percentage of labelled data was higher (10%) and going as high as 0.893 with 20% of the data with label; the fully supervised method achieved an AUC of 0.75 using only the same 2% of the labelled data and 0.807 with 10% of the labelled data. Using the entire labelled dataset with the fully supervised method resulted in an AUC of 0.906. Another important assessment presented concerns the effect of integrating a data augmentation step in the pipeline. Augmentations such as random horizontal flip and random rotation were tested when training the teacher and student models. It could be noticed that strong augmentation functions lead the student model to achieve

better results in the labelled dataset. The same cannot be said regarding the teacher model or the initial training of the student model with the unlabelled dataset in which little difference was observed.

Performing a similar comparative study on the effect of using different proportions of labelled data, Sun *et al.* [62] developed a graph-based semi-supervised approach, using CNN, for breast cancer diagnosis in digital mammography. The SSL method was carried out by hiding the labels of most of the used dataset to replicate unlabelled data. The proposed pipeline begins with ROI extraction, followed by a feature engineering process which involves computing identical features for both labelled and unlabelled ROIs. The subsequent scheme comprises three modules: data weighing (to minimise the effect of noisy data or outliers), feature selection, and a dividing co-training data labelling using a graph-based method, as shown in Figure 3.5. The latter was designed to obtain confident labelling and is similar to majority voting but, rather than dividing the dataset into many subgroups (being each one too small), the authors have shuffled and divided it five times (obtaining ten subgroups, each one with half the dataset, to train the classifiers). With this technique, the achieved AUC and accuracy were, respectively, 0.8818 and 0.8243 for a dataset with less than 5% of the data with label. A comparison in terms of performance to a CNN model that would use only the labelled data could not be made, considering that there was not sufficient data to train the network properly. However, using other models to make this assessment, the highest AUC achieved was 0.8236, observed when using SVM (an AUC of 0.8535 was obtained when utilising the same scheme with the entire dataset). When using the total uncovered dataset with the CNN model, the accuracy increased 3.75% showing that unlabelled data could never replace labelled data. However, when there is a small number of annotated examples and a large amount of unlabelled data, SSL can help build more robust algorithms with superior performance.

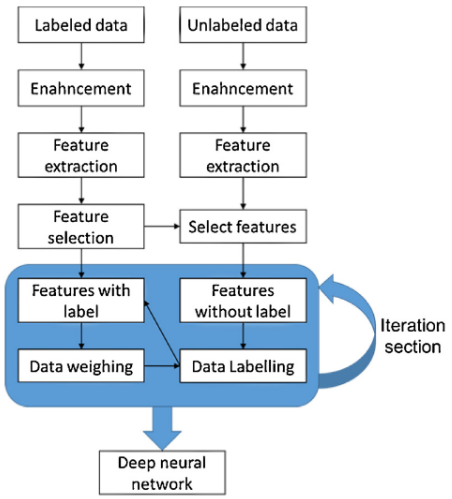


Figure 3.5: Sun *et al.* proposed scheme overview. From [62]

In the same medical field, Al-Azzam & Shatnawi [9] conducted a broader comparative research on the difference between supervised and SSL approaches. To this end, nine different ML

classification algorithms were used — Logistic regression, Gaussian Naive Bayes, Linear SVM, RBF SVM, Decision Tree, RF, XGBoost, Gradient Boosting, and K-Nearest Neighbors — for breast cancer diagnosis. For the SSL method, performed with a self-training technique, the authors divided the dataset into two equal-sized subsets that acted as labelled and unlabelled data, removing the target variable from the second group. They concluded that the accuracy of SL and SSL algorithms were very similar, being the latter a capable solution to address the problem of scarcity of annotated data. The pipeline for the general SSL approach used is presented in Figure 3.6.

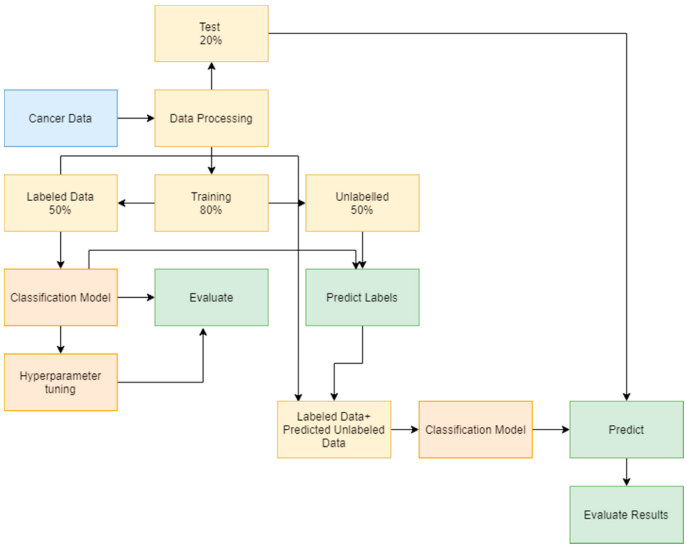


Figure 3.6: SSL approach used by Al-Azzam & Shatnawi. From [9].

Driven by the capacity of semi-supervised generative adversarial networks (GANs) to better capture the data distribution by utilising both unlabelled and labelled examples during training, Das *et al.* [18] proposed a semi-supervised GAN model for breast cancer grading (a scale expressed numerically as 1, 2 or 3) through histopathological images. The presented architecture consists of a generator model and two distinct supervised and unsupervised discriminator models, arranged in a stacked fashion, with parameters sharing. As in traditional GANs, the unsupervised discriminator is applied to be a binary classification model that predicts whether the examples are real or false. While performing this, the model extracts features from the data, learning its distribution. The supervised discriminator is trained to use these extracted features to classify real data, predicting their label. Figure 3.7 illustrates the proposed architecture. The model was tested in two different datasets and, as in other previously mentioned works, the unlabelled sets were obtained by simply removing the labels of a large set of data from both data collections. The recorded accuracies were 0.9822 and 0.9844 with corresponding AUCs of 0.9158 and 0.9327.

Xie *et al.* [69] proposed a semi-supervised adversarial classification model to classify lung nodules as benign and malignant on chest CT scans. The presented scheme comprises two parts: an unsupervised adversarial autoencoder reconstruction network and a supervised classification

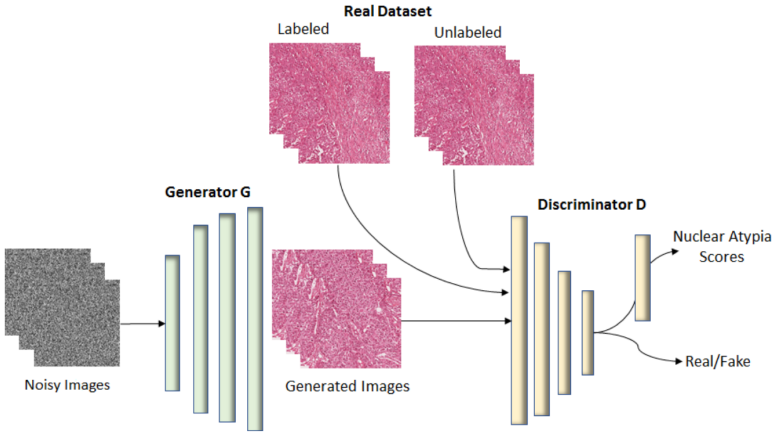


Figure 3.7: The proposed semi-supervised GAN. From [18].

network, to integrate both labelled and unlabelled data. These two components do not share parameters but are rather connected by learnable transition layers, which convey to the classification network the representation capacity learned by the reconstruction network, as illustrated in Figure 3.8. The constructor network consisted of a generator (containing an encoder and a decoder), which reconstructed each input image. The classification network comprised two components separated by a global average pooling. An extension of the model (including 27 submodels with the architecture mentioned above) was used to characterise each overall features of the nodule and tested on the LIDC-IDRI dataset. With this extension, the recorded accuracy and AUC were, respectively, 0.9253 and 0.9581. To demonstrate the advantages of adversarial training in this type of classification task, 10 patches from single view ROIs were randomly selected, and the corresponding reconstructed patches, generated with this model with and without adversarial training, were visualised. The obtained images showed that using adversarial training could generally improve the quality of reconstructed ROIs and, consequently, allowed the reconstruction network to have a stronger ability to characterise lung nodules.

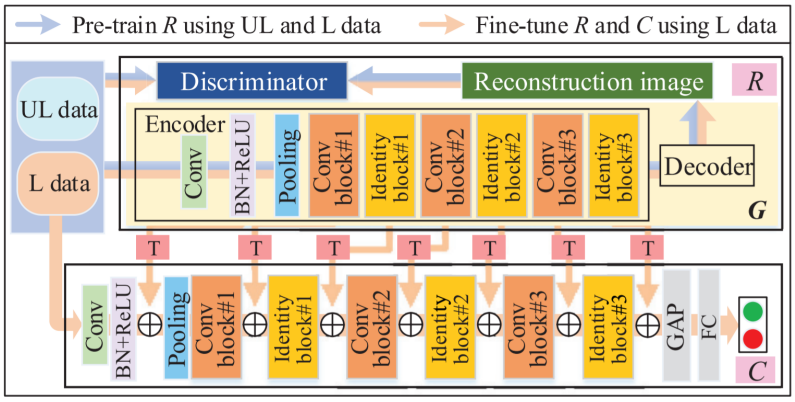


Figure 3.8: Semi-supervised adversarial classification model. "UL": unlabelled data; "L": labelled data; "R": adversarial auto encoder-based reconstruction network; "C": classification network; and "G": generator which contains an encoder and decoder. "T": learnable T layer. From [69].

3.3 Summary

This chapter presented relevant studies regarding both AI solutions in radiogenomics and Semi-Supervised Learning approaches in medical images. Taking into consideration the review performed, some conclusions can be drawn:

- Deep learning is the state-of-the-art approach regarding medical imaging, achieving remarkable success in several tasks in this domain. One of the most appealing characteristics of these algorithms is the ability to tackle with more challenging problems in a very efficient way, without having to extract features beforehand. This aspect is particularly important when assessing mutation status through CT images since it is not yet fully known which structures/tissues can exhibit alterations induced by genetic mutations, so the entire image might contain many mineable data;

- Several works presented models capable of predicting the mutation status of some important biomarkers. When focusing on lung cancer, a link between CT image features and the underlying *EGFR* mutation status could be established. However, a limitation that has often been flagged up concerns the limited size of the publicly available datasets;

- Many studies revealed the effort made to address the small data challenge often faced in the medical domain, with data augmentation-based methods being included in the models pipeline. Although these approaches achieved enhanced performance, they still depend on the number of training data;

- Semi-supervised learning has demonstrated to be a promising and robust approach to overcome the shortage of labelled data. The experiments presented in this chapter evidence the advantage of using a semi-supervised approach over fully supervised training, under certain

conditions. This advantage is even more notorious when the proportion of labelled data included is smaller.

A thorough search of the relevant literature yielded no article in which an end-to-end solution to assess *EGFR* mutation status using a semi-supervised approach is presented.

Chapter 4

Data Description and Preparation

In this chapter, the datasets available for the presented research are detailed (Section 4.1), as well as the steps that were taken to prepare all the data for the task (Section 4.2). A brief summary mentioning the number of considered samples from each data collection is presented in Section 4.3.

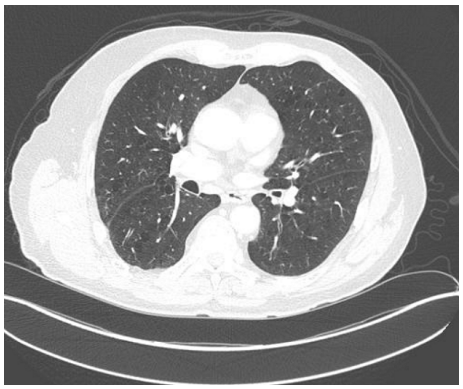
4.1 Data characterisation

Three datasets with CT images were used to develop the proposed work: two including clinical data with *EGFR* mutation status label — NSCLC-Radiogenomics and UHCSJ — and the other, much larger, without this label — NLST. In all datasets, imaging data are originally provided in DICOM (Digital Imaging and Communications in Medicine) format [33], an image format introduced to ensure consistency among different types of medical image modalities. A detailed description of each dataset is provided hereafter.

4.1.1 NSCLC-Radiogenomics

The NSCLC-Radiogenomics dataset [11] is a publicly available collection developed from a cohort of 211 NSCLC patients, comprising clinical and imaging data. The records were acquired between 2008 and 2012 and are related to patients from the Stanford University School of Medicine and the Palo Alto Veterans Affairs Healthcare System. This is a unique dataset containing imaging data paired with genomic data — mutation status information for *EGFR* (172 patients — 43 mutant and 129 wild-type), *KRAS* (171 patients — 38 mutant and 133 wild-type) and *ALK* (157 patients — 2 translocated and 155 wild-type). Besides CT and PET/CT scans, this dataset provides semantic annotation of the tumours in a controlled vocabulary and binary tumour masks. The latter result from a manual delineation made by a radiation oncologist. However, not all data are available for the entire cohort, with only 117 patients having all the previously-mentioned data types. The CT scans contained in this database were acquired using different CT scanners and imaging protocols, resulting in a slice thickness variation from 0.625 to 3 *mm* (with a median value of 1.5 *mm*) and an X-ray tube current from 124 to 699 *mA* (with a mean value of 220 *mA*) at $80-140$ kVp (mean value: 120 kVp) [11].

(a) Mutant



(b) Wild-type

Figure 4.1: Examples of CT scan slice for a) *EGFR* mutant and; b) *EGFR* wild-type. From the NSCLC-Radiogenomics dataset [11].

From this dataset, just 117 patients were considered in the present work and regarding the distribution of the *EGFR* mutation status for this subset, the wild-type is predominant — mutant: $\approx 20\%$ and wild-type: $\approx 80\%$ of the cases. Figure 4.1 displays two examples of CT slices from the NSCLC-Radiogenomics database, one from an *EGFR* mutant patient and the other from an *EGFR* wild-type patient.

4.1.2 NLST

The National Lung Screening Trial (NLST) [6] was a randomised trial of lung cancer screening tests with 53 454 registered participants between 2002 and 2004. All the subjects were individuals considered at high risk: smokers or former smokers, with ages between 55 and 74 and at least 30 pack-year smoking history (a pack-year is a measure used to assess the risk of lung cancer and heart disease caused by smoking: it is determined by multiplying the number of packs of cigarettes — 20 cigarettes — smoked per day by the number of years the person has smoked, and not, as many times thought, by dividing the total packs of cigarettes smoked by the same number of years). The study aimed to evaluate the clinical effectiveness of lung screening with chest CT. Screenings took place from 2002 to 2007 at 33 medical institutions in the United States. From the cohort, 26 722 participants were randomly assigned to screening with low-dose CT and 26 732 to screening with chest radiography. Participants were offered three exams (T0, T1, and T2) performed annually, being the first (T0) done soon after enrolment. All abnormalities found in the exams were recorded and, for a CT scan to be considered positive (suspicious for lung cancer), the radiologist had to observe a non-calcified nodule or mass with at least 4 *mm* of diameter or other suspicious findings for lung cancer. The confirmation of lung cancer was made by the NLST through medical records abstraction, and participants diagnosed with this disease did not undergo any posterior screening test in this trial. In these cases, information was documented in an additional dataset containing data about each confirmed lung cancer case, including tumour size and location. The latter encompasses the following: carina, left hilum, lingula, left lower lobe,

left main stem bronchus, left upper lobe, mediastinum, right hilum, right lower lobe, right middle lobe, right main stem bronchus, right upper lobe, other and unknown.

All screening examinations were performed in line with a standard protocol, which specified acceptable machine characteristics and acquisition variables, resulting in a variation of the slice thickness from 1.0 to 2.5 *mm*, tube current-time product from 40 to 80 mAs with 120 to 140 kVp of voltage [6, 7]. With the data collected in this trial, one of the most extensive chest CT datasets publicly available was built. The NLST database also includes clinical data, which, along with the images, are only available for researchers through the Cancer Data Access System[1]. Out of the 26 722 patients assigned to screening with CT, only 1 089 had a confirmed cancer diagnosis and, from those, just 622 had paired image data. This last subset was the initial collection of data considered in this work and, for not carrying information regarding the *EGFR* mutation status, was the unlabelled set.

4.1.3 University Hospital Center of São João Dataset

This dataset is the result of a collaboration previously established with the University Hospital Center of São João (UHCSJ). It is a database of 141 lung cancer patients, containing clinical and imaging data, collected between 2019 and 2020 (recently, the inclusion of new patients in this database is being resumed). This dataset also includes information regarding the mutation status for a wide set of genes, including *EGFR*, and nodule masks segmented by radiologists. For this project, 108 patients from this dataset were initially considered as only these had both CT scan and binary tumour mask as well as a test result for *EGFR* mutation status. Considering the latter, the distribution is the same as in the NSCLC-Radiogenomics (mutant: $\approx 80\%$; wild-type: $\approx 20\%$).

4.1.4 Comparison of labelled datasets

A crucial point concerns the differences observed in images of the previously described labelled datasets (NSCLC-Radiogenomics and UHCSJ). When integrating different datasets, acquired with different protocols, it is of utmost importance to understand their similarity.

From CT scans analysis of both datasets, it was noticed that the nodules contained in the UHCSJ database were larger than the ones in the NSCLC-Radiogenomics. In fact, certain nodules occupy the entire lung region in some slices. Figure 4.2 displays the difference in terms of the percentage of the lung volume occupied by the nodule for each dataset. The extended right tail for the UHCSJ images is noticeable, with a larger range of values — the measurement of the nodule volume can go as high as around 46% of the total lung volume. On the other hand, the NSCLC-Radiogenomics scans contain smaller nodules, with more than 70% of the cases with tumours affecting less than 1% of the lung (the maximum volume percentage observed is around 15%). Some additional statistics also demonstrate this difference: for the NSCLC-Radiogenomics dataset, the tumours cover, on average, 1.27% of the lung with a median value of 0.34%; the

[1] https://cdas.cancer.gov/plco/

nodules found in the UHCSJ dataset occupy, on average, 4.27% of the lung, with a median value of 1.34%.

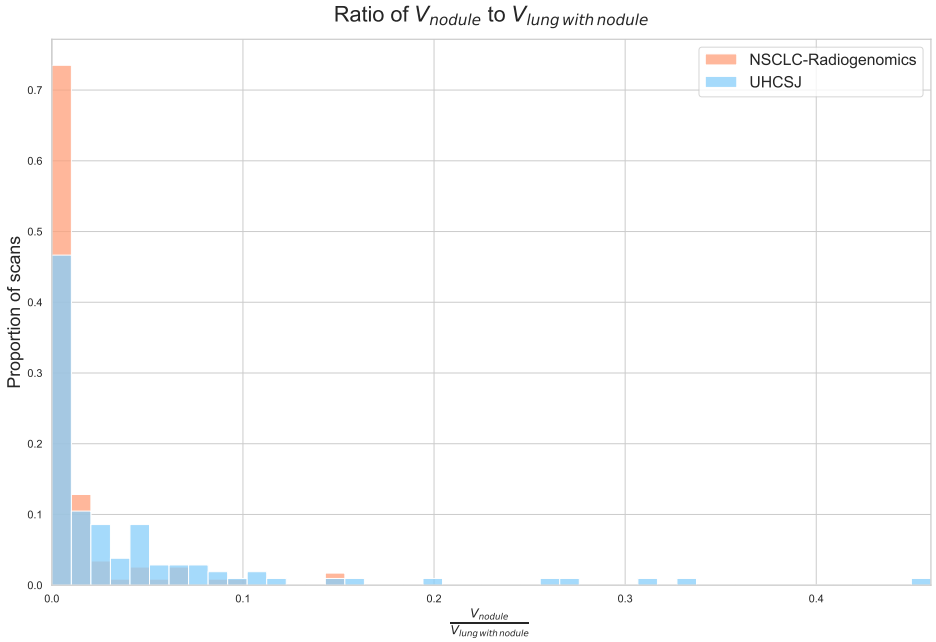


Figure 4.2: Comparison of the ratio V_{nodule} to $V_{lung\,with\,nodule}$ for each labelled dataset. V_{nodule} represents the nodule volume and $V_{lung\,with\,nodule}$ represents the volume of the lung containing the nodule.

4.2 Data preparation

4.2.1 Medical Image Data to Image Arrays

As mentioned before, the CT images were originally available in DICOM format. Therefore, after selecting the ones to be used in the experiments, according to the inclusion criteria specified above, they had to be converted to image arrays using the *Pydicom* package [4]. As these datasets have already been used in previous research works from the supervisor team, the CT scans were made available with some pre-processing, as follows:

- conversion to an image array format;
- pixel values converted to Hounsfield units (HU) by applying a linear transformation using the *Rescale Slope* and *Rescale Intercept* attributes with tags (0028, 1053) and (0028, 1052), respectively [3];

- rescaling of the HU values, using the *min-max* normalisation and assigning values under -1000 HU – corresponding to the air density, as previously mentioned in Section 2.3 – to 0 and values above 400 HU – relating to the density of hard tissues – to 1;
- resize to (256×256).

4.2.2 Data Pre-processing

The performed study used a holistic approach based on the entire lung containing the nodule in confirmed lung cancer cases. To restrict the CT scans to the desired ROI, some pre-processing steps had to be executed. In the case of the NSCLC-Radiogenomics and UHCSJ datasets, binary masks for the nodule were available. In the NLST dataset, no segmentation is given; the only information provided that could be used was the size of the tumours and corresponding locations as well as the CT scan slice number containing the largest nodule diameter. As it was intended to use an extended ROI containing the entire lung, binary masks for the lung containing the nodule, in the case of NSCLC-Radiogenomics, and for both lungs, in the case of NLST and UHCSJ, were obtained using a lung segmentation algorithm developed in previous research works [57]. It was noticed that the segmentation model was not able to return only the background when no lung area was present, as can be seen in Figure 4.3. Additionally, in some cases, the tumour area was not included in the lung masks. For these reasons, additional steps were required and performed differently accordingly to the dataset:

- **NSCLC-Radiogenomics and UHCSJ:** as in these datasets tumour binary masks were available, they were merged with the masks obtained with the segmentation model to achieve lung binary masks which included the nodule. After exploring the UHCSJ dataset, 3 CT scans were discarded as the available masks and images could not provide an acceptable lung region for analysis. This situation is illustrated in Figure 4.4, where one of these CT scans is provided. In this case, the left lung (the one on the right side of the image) contains the nodule, but it is almost not visible. For this reason, the corresponding image provided after the segmentation algorithm was applied does not provide an acceptable lung area, as demonstrated in Figure 4.5. It is noticeable the nodule region but lacks almost the entire lung. Being intended to detect a mutation that led to cancer, based on a holistic approach considering the entire lung containing the nodule, it is of major importance to have both the nodule and lung well represented.

  In the case of the UHCSJ images, as only masks for both lungs were made available, the images were cropped to a single side (left or right), accordingly to the nodule position. To correct the segmentation and remove the noise that could be found in the first and last slices of each scan, as already illustrated in the example of Figure 4.3, these new masks were used to determine the contours present in each slice. The number of contours was used to limit the initial and final slice of the scan so that, ideally, only the lung region would be exhibited. In the initial slices and given the enormous amounts of contours found that mainly belong

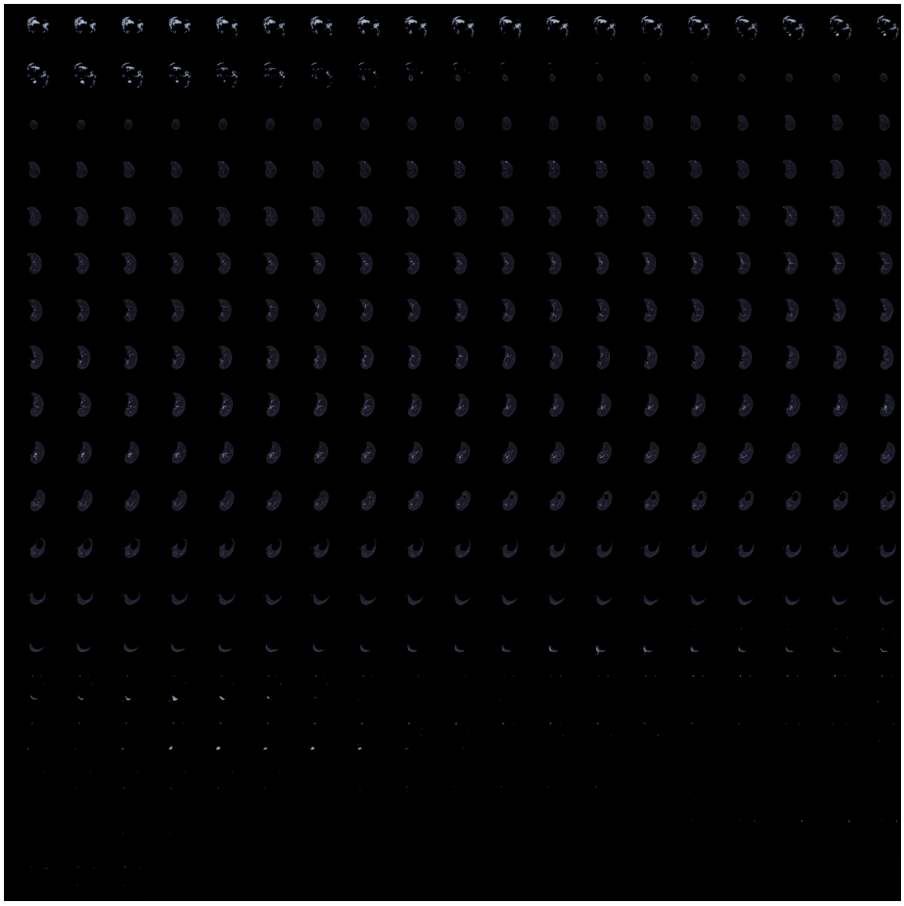


Figure 4.3: Illustrative example of a CT scan before pre-processing. From the NSCLC-Radiogenomics dataset [11].

to structures other than lungs, the maximum area exhibited was compared with the area of an enclosing circle since the initial part of the lung should be approximately circular. This method returned the first and last slice to be considered in each case, leading to cleaner images. After this process, each scan was cropped to the desired ROI — the entire lung — to eliminate unnecessary pixels. The result is illustrated in Figure 4.6, which corresponds to the same CT scan presented in Figure 4.3 after the described method has been applied.

- **NLST:** after the lungs segmentation, a similar method as the one above described was applied. However, in this case, as no nodule binary mask was provided, and since in some of the resultant images the nodule region was not included in the masks, the contours of the lungs were used to fill these gaps. With these resulting masks, and using the information in *The Lung Cancer* dataset (included in the NLST database), each scan was cropped to the lung containing the nodule. In 4.1.2, the possible locations for the tumour were stated. Since the carina is located at the base of the trachea (the area where the trachea splits into left and right bronchus), and the mediastinum is also located in the region that separates the lungs, individuals that only presented tumours in these locations were excluded. Likewise, when

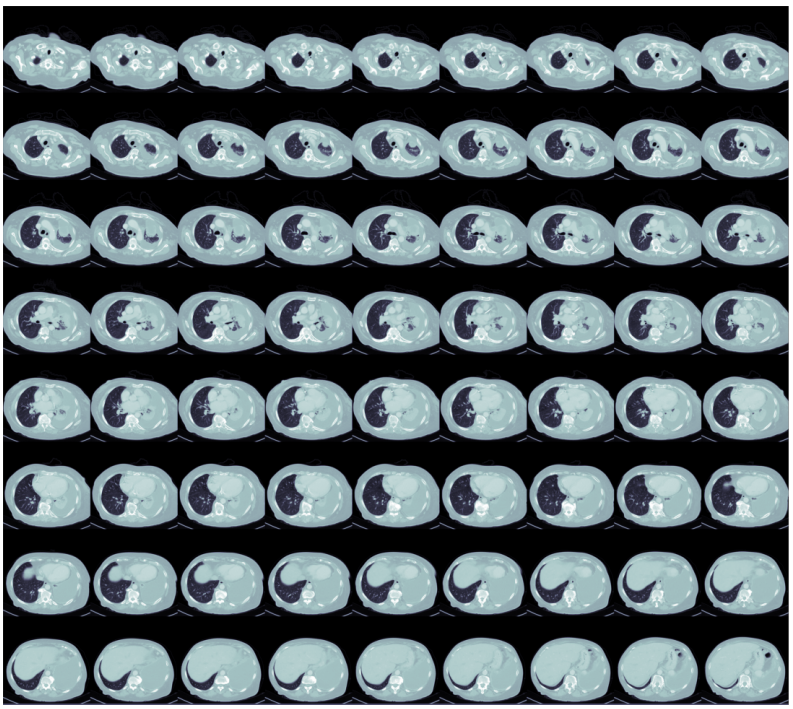


Figure 4.4: Example of a CT scan discarded from the UHCSJ dataset.

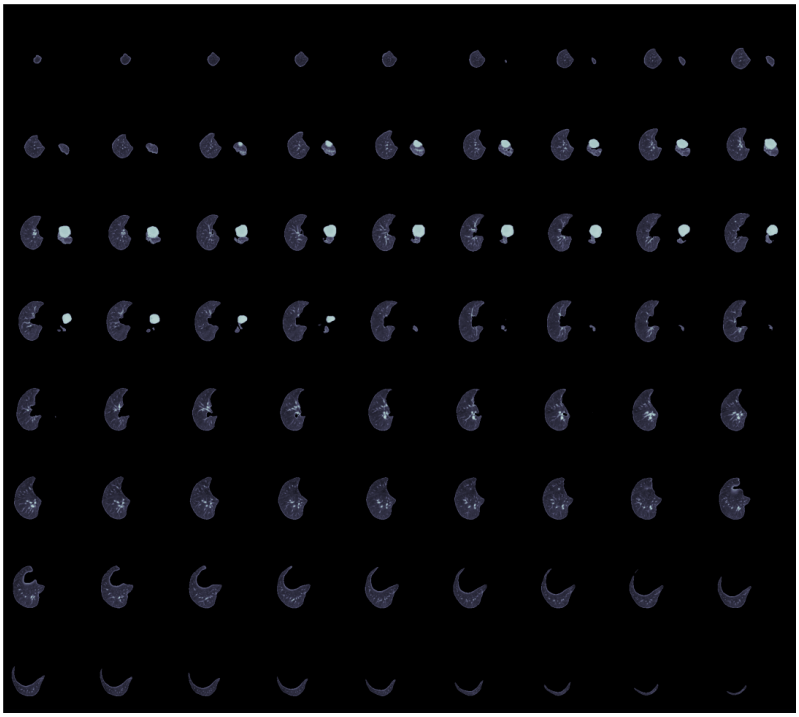


Figure 4.5: Lungs segmentation of a CT scan discarded from the UHCSJ dataset.

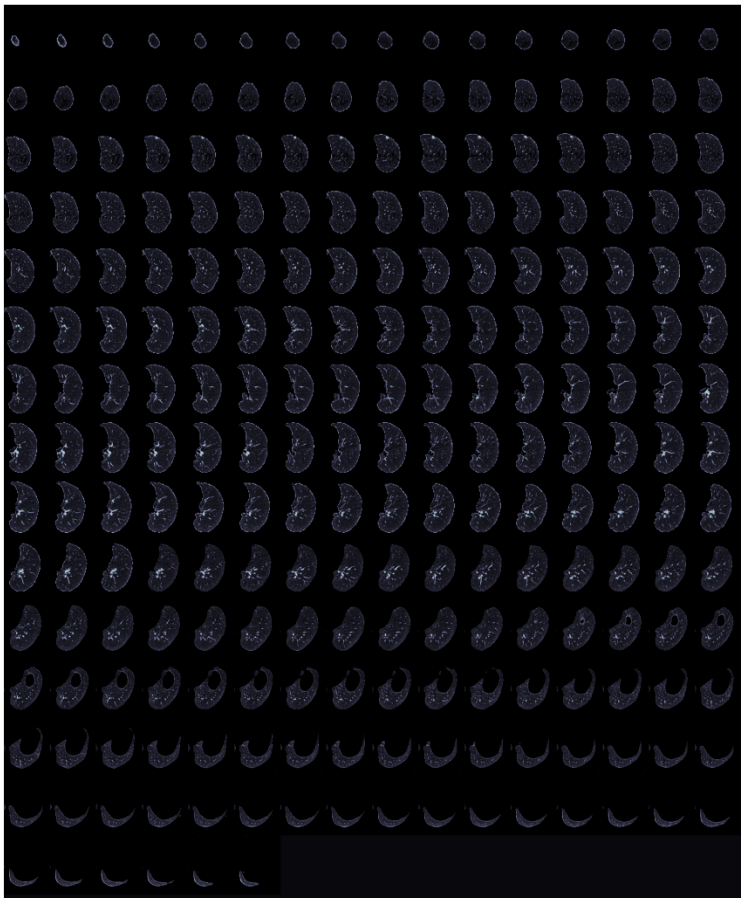


Figure 4.6: Illustrative example of a CT scan after pre-processing.

the only location available was *unknown* or *other*, the CT scans were not considered. Additionally, inconsistencies between the information presented in *The Lung Cancer* dataset and the *Spiral CT Abnormalities* dataset (also integrated in the NLST database) were identified. When they could not be resolved (for instance, in the cases where the malignant tumour was identified on the left side of the lung and there were only recorded nodules on the right side), the scans were not considered. Moreover, one case was found in which both lungs exhibited primary tumours. In this situation, the two lungs were considered distinct samples (as if they belonged to different patients). This resulted in a total of 574 volumes.

In this work, it was intended to take as input 3D volumes, providing information about the lung as a whole. For this reason, data uniformisation was an essential step as CT scans from the considered datasets had a varying number of slices: from 245 to 635 in the NSCLC-Radiogenomics, 61 to 261 in the UHCSJ, and between 46 and 545 in the NLST dataset. Therefore, to obtain volumes with the same number of slices, a standard depth of 64 was selected. Considering this value, it was preferable not to have CT scans with an inferior number of slices. For this reason, the only CT scan with less than 64 slices was excluded from the NLST data collection. The same criterion was applied to the UHCSJ data collection, resulting in 3 CT scans being rejected. Additionally, the

axial image size was also a challenge due to resource limitations. Therefore, each cropped slice was resized to (128×64) by interpolation. To achieve the desired standard depth, two different strategies were tested:

- rescale the volume, using interpolation, by a scaling factor given by

$$factor = \frac{desired\ depth}{scan\ depth} \tag{4.1}$$

  where *desired depth* was set to 64 and *scan depth* represents the number of slices of the CT scan.

- select 64 linearly spaced slices, simulating a wider space between slices. In this selection, it was imposed that the slice containing the maximum nodule mask area was included.

4.3 Summary

Having performed a description of the datasets available to develop the proposed investigation, the final number of considered images from each one, according to the aforementioned inclusion criteria, is summarised in Table 4.1.

Table 4.1: Number of CT scans considered from each dataset.

Dataset	EGFR mutation status	# CT scans considered
NSCLC-Radiogenomics	Available	117
UHCSJ	Available	102
NLST	Unavailable	574

Chapter 5

Lung Cancer Characterisation

Multiple approaches to tackling SSL tasks were presented in Chapter 3. The current chapter provides a detailed description of the proposed semi-supervised methodology to perform lung cancer characterisation. The presented approach uses a holistic analysis with an extended ROI, including the entire volumetric region of the lung containing the nodule instead of using the nodule region only. This methodology is detailed in Section 5.1. The experiments conducted to achieve this goal are addressed in Section 5.2.

5.1 Method overview

As previously mentioned, the proposed method aimed to develop a more robust classification model for *EGFR* mutation status assessment using CT scan images as input in a combination of labelled and unlabelled data. To achieve this, a generative model was used, particularly a GAN, to explore the power of adversarial training. Traditional GANs, being unsupervised ML techniques, can deal with a vast amount of unlabelled data to produce new, fake images on the same domain.

5.1.1 Generative Adversarial Networks

Generative Adversarial Networks were first proposed by Goodfellow *et al.* [26] in 2014 and are now widely used in image, video and voice generation. In a GAN, two neural networks —- a generator and a discriminator — are trained, competing against each other: the generator tries to deceive the discriminator with fake data, whereas the latter attempts to distinguish between samples drawn from the training data and examples drawn from the generator [27]. Typically, the generator network receives as input some random noise and learns to transform it into a realistic image (or other kinds of data). Rather than be trained to minimise a distance to a specific input or distribution, it is trained to fool the discriminator, which in turn learns to discriminate better. By training the discriminator to differentiate these produced images from real data, it is expected that this network learns feature representations related to the training data.

The underlying concept in semi-supervised GAN is using this ability to learn feature extraction and expose the network to a larger amount of data — the unlabelled data — so that it can extract information from many more samples, which can be helpful in the classification task. Hence, the discriminator is trained not only in the labelled data but also in the unlabelled set, extracting features from this collection to discriminate between real and synthetic data. Therefore, in an SSL GAN, the discriminator is trained not only to differentiate real images from synthetic ones but also to predict the label for the input data. By using this semi-supervised approach, it was expected, on the one hand, that the GAN could be able to provide additional information to the classifier and, on the other hand, that the classification results could act as a regulariser to the generative model. Given the difficulty of training GANs, both in terms of convergence and stability, instead of using a typical adversarial network, where the starting point is some random vectors, a combination of Variational Autoencoder (VAE) and GAN was used. With this, an image-to-image approach was obtained, bringing more stability to the training.

5.1.2 Variational Autoencoder

Autoencoders, an efficient feature extraction method, are neural networks used to learn lower-dimensional codifications — latent space -– to, afterwards, generate input reconstructions. Their architecture comprises two networks: an encoder and a decoder. The former transforms the input data into an encoded representation, and the latter reconstructs, as close as possible, the original input from the low dimensional latent space. Thus, the decoder acts similarly to a GAN generator, projecting a low-dimensional vector to an image. A shortcoming of this kind of representation learning algorithm is that it does not allow the generation of new samples as it uses a deterministic approach. VAEs [35], generative models with a similar structure to autoencoders and a solid probabilistic foundation, replace the encoded representation by a stochastic sampling operation, learning, instead, parameters of a probability distribution using a Bayesian approach. As the posterior distribution $p(z|x)$ (where z is the latent variable and x the input) is an intractable probability distribution, using this variational inference, the encoder learns $q(z|x)$, a simpler and tractable, distribution, as an approximation [20]. Typically, $q(z|x)$ is a Gaussian.

5.1.3 Proposed method

In this study, the proposed method encompasses two main structures connected by a shared network: a VAE and a semi-supervised GAN, where the decoder of the VAE acts as the GAN generator, as illustrated in Figure 5.1. The discriminator has two different outputs, both for binary classification tasks: one for the likelihood of an image being a real CT (belonging to the training data) or a reconstructed one, and the other to classify labelled data as *EGFR* mutant or wild-type. In fact, this can be seen as having a discriminator and a classifier with a common backbone (common feature extraction layers) and two different output layers. The encoder of the VAE receives as input CT scan images from the training set (both labelled and unlabelled) and maps them to a

distribution, providing as outputs two vectors: one representing the mean (μ) and the other representing the log-variance (σ) of $q(z|x)$. Latent vectors z, sampled from these distributions, are passed to the decoder/generator that maps the code to an image. In the generation of each of these samples z, a reparameterisation must be performed to enable backpropagation (backpropagation cannot be performed across a stochastic node due to the impossibility of differentiating the sampling operation) [35]. Thus, the variable z can be obtained by $z = \mu + \sigma \odot \varepsilon$, where $\odot$ represents the element-wise product and ε is an auxiliary noise variable (the stochastic component), $\varepsilon \sim \mathcal{N}(0, 1)$. The reconstructed and original images are then provided as input to the discriminator/classifier, whose role is to undertake the two classification tasks mentioned above. To stabilise the training of the generator, avoiding a faster convergence of the discriminator early in training, the VAE part was initially fixed to be trained alone, giving the decoder/generator a better starting position.

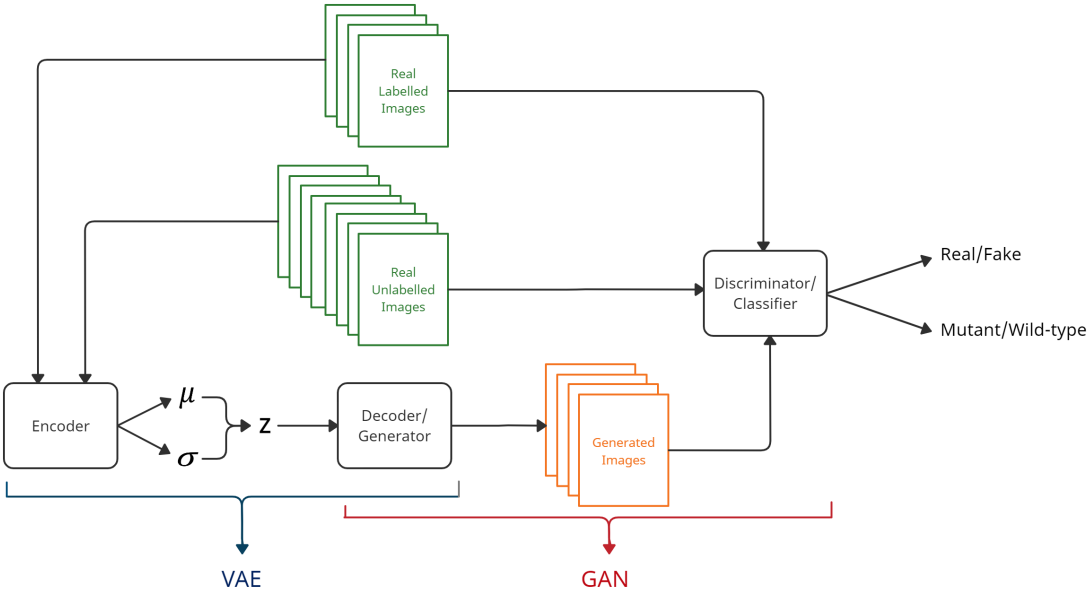


Figure 5.1: Proposed method, which encompasses two main structures connected by a shared network: a VAE and a semi-supervised GAN, where the decoder of the VAE acts as the GAN generator.

5.1.4 Evaluation

For all the experiments performed (which are fully detailed in Section 5.2), the labelled dataset considered was randomly split into two different sets, one for training (80%) and the other for testing (20%). The unlabelled set utilised was added to the training set, and the model was trained until the classifier (the discriminator) converged. To achieve a model as robust as possible, and to explore data variance, 10 different random train-test splits were performed. During this training

and testing process, different evaluation metrics were computed and averaged over the 10 random splits: AUC, precision, sensitivity and specificity.

5.2 Experiments

A long path of advances and setbacks was paved until a final model was achieved. Different experiments were conducted in order to test some possible solutions for the model described above. In addition to hyperparameter tuning, different discriminator network architectures were tested as well as some variations to the loss functions. Furthermore, a 2D approach was also tested.

5.2.1 Networks architectures

The proposed VAE base architecture was kept unchanged during the experiments and is represented in Figure 5.2. As can be seen, the encoder is composed of five convolution blocks (C1 to C5) and two bottleneck dense layers. Each convolution block comprises convolutional layers, with an increasing number of filters, each one followed by a Rectified Linear Unit (ReLU) activation function and batch normalisation [32]. To reduce the size of each feature map by half, $4 \times 4 \times 4$ kernels with stride 2 and padding 1 were used. To reconstruct the original input, a latent vector with the size of the latent dimension (l_{dim}) is passed to a dense layer to, afterwards, be reshaped into 256 activation maps. This is used as input to four transposed convolution blocks (TC1 to TC4), each one composed of a $4 \times 4 \times 4$ transposed convolutional layers, with a decreasing number of filters, with stride 2 and padding 1, followed by a Leaky ReLU activation (with a negative slope of 0.2) and batch normalisation. The last transposed convolutional layer (TC5) is followed by a Sigmoid activation function, to ensure that all output pixel values belong to the original range of values — $[0, 1]$.

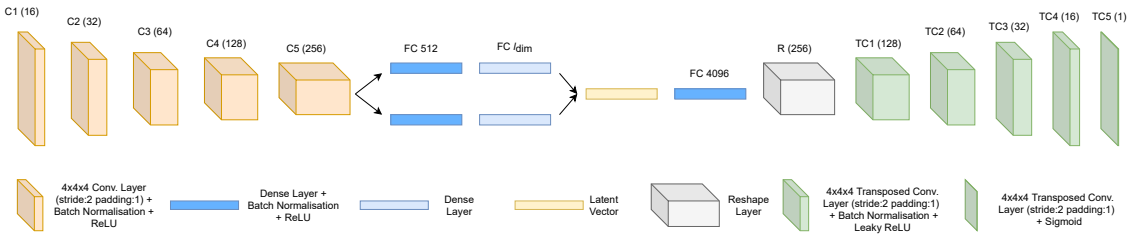


Figure 5.2: Proposed VAE architecture.

Considering the GAN architecture — largely based on the DCGAN [46] — as previously mentioned, the VAE decoder acts as the generator network. Therefore, only a scheme of the discriminator base architecture is represented in Figure 5.3. As illustrated, this network embodies four blocks of convolutional layers with $4 \times 4 \times 4$, stride 2 and padding 1 for down-sample, and one dense layer, in the backbone, followed by two classification heads. Each convolutional layer is followed by batch normalisation and uses Leaky ReLU (with a negative slope of 0.2) as the activation function. A dropout layer [61] was added after each of these convolutional blocks to

decrease the number of trainable parameters, reducing overfitting. This regularisation strategy, which consists in randomly dropping out neurons during training with a certain defined probability, has the additional benefit of promoting a more robust feature extraction. Similarly to the decoder network, the number of filters increases as the network goes deeper. Lastly, a dense layer with a variable number of neurons was included prior to the classification heads. Additionally, it was tested a slight variation of the base architecture depicted in Figure 5.3 by adding in each classification head another dense layer, with a smaller number of neurons than the one used in the previous layer.

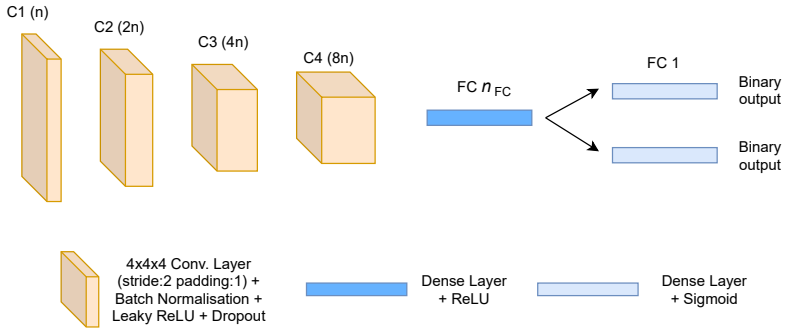


Figure 5.3: Proposed discriminator architecture.

Before this final base architecture being achieved, a different design for the discriminator was tested and is illustrated in Figure 5.4. This approach was introduced by Salimans *et al.* [50] and the main difference between the two implementations concerns the output layer: instead of having two output layers, a single one is used with two nodes (the same number of classes in the initial supervised classification problem) and, therefore, a Softmax activation function. The unsupervised classification task uses the outputs before the activation function, and a normalised sum of the exponential outputs is calculated, returning the probability of the input being fake (please refer to [50] for further detail about this variation).

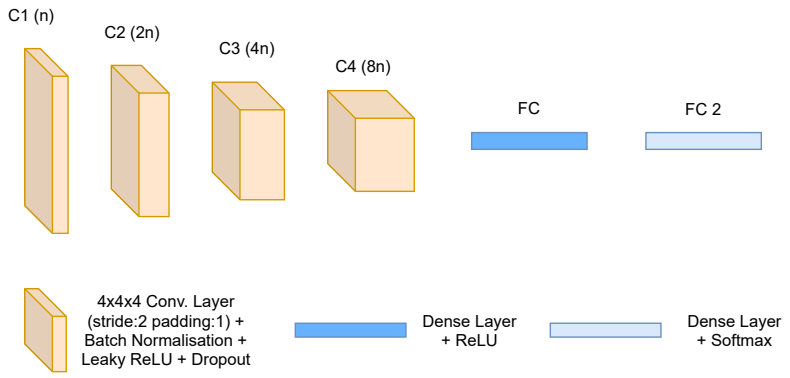


Figure 5.4: Alternative discriminator implementation.

5.2.2 Loss functions

During the training process, different loss functions were tested for optimisation of all the networks involved, trying to find the best way to combine the VAE with the GAN, with the best possible performance. Starting with the discriminator, the loss functions considered for the optimisation of this network only suffered slight modifications during the model development and comprised an adversarial loss and a supervised classification loss, used separately or combined (both options were tested). In the adversarial part, as proposed in the original GANs paper [26], the goal of the discriminator is to maximise the function presented below:

$$\mathscr{L}_{adversarial} = \mathbb{E}_{x\sim p_{data}(x)}\left[log\, D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[log(1-D(G(z)))\right] \tag{5.1}$$

being: $\mathbb{E}_{x\sim p_{data}(x)}$ and $\mathbb{E}_{z\sim p_z(z)}$ the expected values over the real data inputs and over the fake images $G(z)$, respectively; $D(x)$ the discriminator probability estimation that a real image x is real; $G(z)$ the generator output for a given input z and $D(G(z))$ the discriminator probability estimation that an image produced by the generator is real.

Additionally, it was tested another version of the Equation 5.1, by adding another term. The goal of such a term is to enforce the discriminator to distinguish not only images generated from latent vectors sampled from the distributions outputted by the encoder but also from random noise vectors sampled from a Gaussian distribution. Hence, the alternative adversarial loss to be maximized is given by:

$$\mathscr{L}_{adversarial} = \mathbb{E}_{x}\left[log\, D(x)\right] + \mathbb{E}_{z}\left[log(1-D(G(z)))\right] + \mathbb{E}_{z_p}\left[log(1-D(G(z_p)))\right] \tag{5.2}$$

with $z_p \sim \mathscr{N}(0,1)$. The supervised classification loss considered was the Binary Cross-Entropy (BCE):

$$\mathscr{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(Y_i \cdot log\,\hat{Y}_i + (1-Y_i)\cdot log(1-\hat{Y}_i)\right) \tag{5.3}$$

where Y_i is the ground truth label and $\hat{Y}_i$ is the predicted probability for the i$^{\text{th}}$ image, and N the mini-batch size.

Moreover, it was tested if the addition of a manifold regularisation term would improve the overall performance. A manifold regularisation, $\Omega_{manifold}$, as outlined in Section 3.2, should enforce the discriminator to yield similar features for nearby points in the latent space.

$$\Omega_{manifold} = \left\| D(G(z)) - D(G(z+\delta\cdot 1\times 10^{-5})\right\|, \delta \sim \mathscr{N}(0,1) \tag{5.4}$$

For the optimisation of the encoder and decoder/generator networks, different combinations of loss functions were tested and are now described (in the equations that follow, the parameters γ and λ represent loss weights and the reduction or aggregation method selected is the mean):

- Decoder/generator optimisation:

C_1) with a loss function composed of a reconstruction term $\mathscr{L}_{reconstruction}$ (Equation 5.5), in this case, the Mean-Squared Error (MSE) between the decoder reconstruction $\hat{x}$ and the original input x, and a generator term, $\mathscr{L}_{adversarial}$. The latter is determined by maximising the log-probability of the discriminator considering generated images as belonging to the training data or, analogously, minimizing the log-probability of the discriminator correctly classifying fake images. This was done by applying the loss introduced by Goodfellow *et al.* [26] given by Equation 5.6 or, alternatively, the non-saturating version (Equation 5.7). In such a case, the decoder/generator loss is provided by Equation 5.8;

$$\mathscr{L}_{reconstruction} = \frac{1}{N}\sum_{i=1}^{N}(x-\hat{x})^2 \tag{5.5}$$

$$\mathscr{L}_{adversarial} = \mathbb{E}_{z\sim p_z(z)}\left[log(1-D(G(z)))\right] \tag{5.6}$$

$$\mathscr{L}_{adversarial} = \mathbb{E}_{z\sim p_z(z)}\left[-log(D(G(z)))\right] \tag{5.7}$$

$$\mathscr{L}_{generator} = \lambda \cdot \mathscr{L}_{reconstruction} + \mathscr{L}_{adversarial} \tag{5.8}$$

C_2) maintaining the same reconstruction loss mentioned above (Equation 5.5) and substituting the adversarial loss (Equation 5.7) by the feature matching loss, introduced by Salimans *et al.* [50], where the generator is encouraged to synthesise data that minimises the statistical difference between the features of the real and fake data on an intermediate layer of the discriminator. Therefore, this loss is defined as follows:

$$\mathscr{L}_{feature\ matching} = \left\|\mathbb{E}_{x\sim p_{data}} f(x) - \mathbb{E}_{z\sim p_z(z)} f(G(z))\right\|^2 \tag{5.9}$$

where $f(x)$ represents activations on an intermediate layer of the discriminator. Consequently, in this case,

$$\mathscr{L}_{generator} = \lambda \cdot \mathscr{L}_{reconstruction} + \mathscr{L}_{feature\ matching} \tag{5.10}$$

C_3) inspired by Larsen *et al.* [37], using the feature matching loss (Equation 5.9), instead of a reconstruction loss, combined with an adversarial loss achieved by:

$$\mathscr{L}_{adversarial} = \mathbb{E}\left[log(1-D(G(z)))\right] + \mathbb{E}\left[log(1-D(G(z_p)))\right] \tag{5.11}$$

where z_p is a sample from the prior $\mathscr{N}(0,\,1)$. For this situation,

$$\mathscr{L}_{generator} = \lambda \cdot \mathscr{L}_{feature\ matching} + \mathscr{L}_{adversarial} \tag{5.12}$$

- Encoder optimisation:

  C_1) with the traditional VAE loss, which incorporates both a reconstruction loss (Equation 5.5) and a latent loss $\mathscr{L}_{prior}$ (Equation 5.13), the Kullback-Leibler (KL) divergence loss or relative entropy. The latter is a statistical measure and quantifies the distance between two probability distributions [17], in this case, the distribution of the encoder output and a Gaussian of mean 0 and variance 1. Therefore, for this situation, the encoder loss can be obtained by Equation 5.14;

$$\begin{aligned}\mathscr{L}_{prior} &= \mathbb{D}_{KL}(q(z|x)\,||\,\mathscr{N}(0,1)) \\ &= -\frac{1}{2}\sum_{i=1}^{N}(1+log\,(\sigma^2)-\sigma^2-\mu^2)\end{aligned} \tag{5.13}$$

$$\mathscr{L}_{encoder} = \gamma \cdot \mathscr{L}_{reconstruction} + \mathscr{L}_{prior} \tag{5.14}$$

  C_2) replacing the previously mentioned reconstruction term by the feature matching loss (Equation 5.9), maintaining the KL divergence loss. Thus, in this case,

$$\mathscr{L}_{encoder} = \gamma \cdot \mathscr{L}_{feature\ matching} + \mathscr{L}_{prior} \tag{5.15}$$

5.2.3 Hyperparameters

The networks described in 5.2.1 required careful hyperparameters optimisation for fine-tuning. Due to hardware limitations (a NVIDIA Tesla P100 16GB graphics processing unit was used), some hyperparameters had to be restrained, namely, the mini-batch size and the number of neurons in the hidden layers. Table 5.1 presents the list of values considered for the hyperparameters manual search applied.

As the training sets used included different proportions of labelled and unlabelled data, it was chosen to keep the same ratio in the mini-batch size. That is, if the used training set had, for instance, a proportion of 15% of annotated data and the remaining 85% of data without label, the mini-batch comprised data with a similar division.

5.2.4 Imbalanced data

As usually occurs when dealing with medical diagnosis, the labelled datasets described in Section 4.1 have an uneven class distribution, with the mutant type as the under-represented class. If a classification model were built with this imbalance, without any further attention given, the model would tend to be biased towards the negative classification, failing to capture the minority class. To tackle this, two strategies were tested:

Table 5.1: List of values used for hyperparameters optimisation in the SSL model.

Hyperparameter	Values
Mini-batch size	8, 16, 32
Dropout discriminator	0.3, 0.4, 0.5
Learning rate discriminator	1×10^{-5}, 2×10^{-5}, 5×10^{-5}, 1×10^{-4}, 3×10^{-4}, 1×10^{-3}
Learning rate VAE	1×10^{-5}, 1×10^{-4}, 2×10^{-4}, 1×10^{-3}
Optimisers	Adam, AdamW, SGD [a]
Weight decay	1×10^{-7}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}
Momentum	0.1, 0.5, 0.9
l_{dim}	128, 256, 512
n [b]	16, 32, 64
n_{FC} [c]	512, 1024
γ	1, 5
λ	5, 10, 15

[a] Stochastic Gradient Descent.

[b] Number of filters in the first hidden layer of the discriminator network.

[c] Number of neurons in the dense layer of the discriminator network.

- oversampling the minority class by applying data augmentation techniques – horizontal and vertical flip, random rotation, and adding Gaussian noise – *on-the-fly*, this is, without physically storing transformed images. Instead, in each mini-batch, the same number of samples for each class is used, allowing repetition of the minority samples, applying transformations to 75% of them;

- using a weighted loss function during training, including in the classification loss an argument with class weights given by $\left[\frac{n_0}{n_0}, \frac{n_0}{n_1}\right]$, where n_0 represents the number of examples in the negative class (the majority class) and n_1 represents the number of examples in the positive class (the minority class) in the training set.

5.2.5 Proportions of unlabelled data

A final experiment relates to the percentage of unlabelled data used. To evaluate the performance when different amounts of data without label are used to build the SSL approach, once a final model was achieved, the number of utilised training samples from the NLST dataset was reduced.

When training the model only with the NLSCL-Radiogenomics dataset as labelled collection, using the entire NLST subset summarised in Table 4.1 (Section 4.3) corresponds to a percentage of around 14% of labelled data. Once the UHCSJ dataset is combined with the NSCLC-Radiogenomics, this percentage rises to circa 23%, as summarised in Table 5.2. To investigate if a variation in the unlabelled dataset size would affect the classification performance, and up to which point, different values for the percentage of NLST data used were tested: 100% (the base model), 80%, 60%, and 40%. The corresponding proportions labelled-unlabelled data, for both possible cases (using only one labelled data collection or combining both) are detailed in Table 5.3.

Table 5.2: Percentage of labelled data when using the entire NLST dataset.

Dataset	Available # samples	# Samples for training	% labelled data
NSCLC-Radiogenomics	117	94	14.07%
NLST	574	574	
NSCLC-Radiogenomics + UHCSJ	219	175	23.36%
NLST	574	574	

Table 5.3: Proportions labelled-unlabelled data for different NLST data percentages.

Labelled set	% NLST	% labelled data	% unlabelled data
NSCLC-Radiogenomics	100	14	86
	80	17	83
	60	21	79
	40	29	71
NSCLC-Radiogenomics + UHCSJ	100	23	77
	80	28	72
	60	34	66
	40	43	57

5.2.6 2D approach

When a 3D model was achieved, it was decided to test a 2D counterpart to verify if a better model could be built, having fewer parameters and a much smaller input space for pattern search and, therefore, less probability of suffering from overfitting. To achieve this, it was used as input, from each CT scan, the slice containing the largest area of the nodule. In the case of the NLST dataset, this information was available in a complementary file. For the NSCLC-Radiogenomics and UHCSJ data collections, the corresponding slice was determined using the nodule masks. The model used for this approach was achieved by simply adapting the 3D model obtained for the 2D version.

Chapter 6

Results and Discussion

In this chapter, the results of the achieved model are presented and discussed. In Section 6.1, the best set of hyperparameters is presented as well as the corresponding attained results. Moreover, some relevant comparisons that may guide future works are provided. In Section 6.2 a detailed analysis of the findings is made.

6.1 Results

Numerous experiments were performed to achieve the final model, whose results are presented in this chapter. The experiments were conducted in phases to optimise the available time. The best outcomes were provided when using the base architecture represented in 5.3, as outlined in the previous chapter, with the following combination of loss functions:

- discriminator — using the $\mathscr{L}_{adversarial}$ depicted in 5.2 and the $\mathscr{L}_{BCE}$ propagated separately through the network;
- decoder/generator -– applying the $\mathscr{L}_{generator}$ summarised in 5.12;
- encoder — optimised with the $\mathscr{L}_{encoder}$ presented in 5.15.

The base model with the best performance was obtained with the final discriminator architecture depicted in Figure 6.1 and the set of hyperparameters presented in Table 6.1. The model with all these settings defined was trained in the remaining scenarios:

- tackle with the imbalance present in the labelled set by oversampling or by using a weighted loss;
- rescaling the image depth by interpolation or by selecting 64 linearly separated slices, as already detailed in Section 4.2.2;
- adding a manifold regularisation term (Equation 5.4) to the discriminator loss;

- using only the NSCLC-Radiogenomics as labelled data or combining it with the UHCSJ dataset.

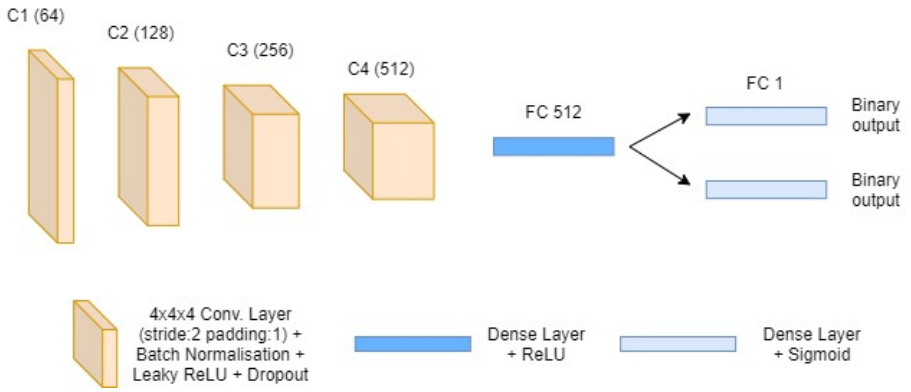


Figure 6.1: Final discriminator architecture that yielded the best results.

Table 6.1: Set of hyperparameters that led to the best-performing SSL model.

Hyperparameter	**Values**
Mini-batch size	32
Dropout discriminator	0.5
Learning rate discriminator	1×10^{-5}
Learning rate VAE	2×10^{-4}
Optimisers	Adam
Weight decay	1×10^{-4}
l_{dim}	128 $^{(a)}$
n $^{(b)}$	64
n_{FC} $^{(c)}$	512
γ	5
λ	10

$^{(a)}$As no significant differences between the three tested values were found, a size of 128 was selected to reduce the number of trainable parameters.
$^{(b)}$Number of filters in the first hidden layer of the discriminator network.
$^{(c)}$Number of neurons in the dense layer of the discriminator network.

Starting with the strategy to overcome the imbalance found in the dataset, Table 6.2 presents the achieved performances when considering the two tested approaches using only the NSCLC-Radiogenomics dataset as the labelled set. The results are also discriminated by the technique used to rescale the depth of the considered volumes. Using a weighted loss function instead of oversampling the minority class provided better results, being more noticeable when the rescaling of the number of slices per image is performed with interpolation, registering an AUC of 0.7011 ± 0.1143 averaged over 10 random train-test splits. Thus, in the results provided hereafter, this was the technique applied.

Table 6.2: Performance results for the *EGFR* mutation prediction when considering different strategies to handle imbalanced data. The evaluation metrics are presented as mean ± standard deviation, averaged over 10 random train-test splits.

Strategy imbalanced	Image depth rescaling	AUC	Precision	Sensitivity	Specificity	Running time (*min*)[(a)]
Weighted loss	Interpolation	0.7011 ± 0.1143	0.4246 ± 0.1395	0.6600 ± 0.2538	0.7421 ± 0.1341	65
	Linearly spaced	0.5974 ± 0.1060	0.3019 ± 0.0874	0.5000 ± 0.2483	0.6931 ± 0.1206	65
Oversampling	Interpolation	0.6100 ± 0.0709	0.4334 ± 0.2269	0.4200 ± 0.1887	0.8000 ± 0.1285	85
	Linearly spaced	0.5921 ± 0.1061	0.4202 ± 0.2868	0.3000 ± 0.2408	0.8842 ± 0.0774	85

[(a)] Average training time for each random train-test split (in minutes).

When evaluating the effect of adding a manifold regularisation term to the discriminator loss, the results show that such a term is more advantageous when the strategy used for rescaling is the selection of linearly separated slices, as detailed in Table 6.3. While with the interpolation method, the model achieved worse performance, using a manifold regularisation with the former strategy increases the average AUC from 0.5974 ± 0.1060 to 0.6274 ± 0.1372. Again, these results concern only the utilisation of one labelled data collection.

Table 6.3: Performance results when considering the addition of a manifold regularisation term to the discriminator loss.

Image depth rescaling	AUC	Precision	Sensitivity	Specificity	Running time (*min*)[(a)]
Interpolation	0.6648 ± 0.1252	0.3668 ± 0.1338	0.6400 ± 0.2498	0.6895 ± 0.1401	84
Linearly spaced	0.6274 ± 0.1372	0.3495 ± 0.2158	0.4600 ± 0.2835	0.7947 ± 0.1011	84

[(a)] Average training time for each random train-test split (in minutes).

Lastly, different percentages for the NLST data were tested, as already detailed in Section 5.2.5. The results for the two models that achieved better results in the previous experiments are detailed in Tables 6.4 and 6.5. It should be noticed that a fully supervised model (using exclusively labelled data) is only presented in Table 6.4, since the addition of the manifold regularisation, as detailed in 5.2.2, was added upon the SSL approach. Additionally, Table 6.6 displays the corresponding average running times (in minutes) for these models.

Table 6.4: Performance results for different percentages of the unlabelled dataset used when considering a weighted loss function.

	Percentage unlabelled dataset used				
Metric	**100%**	**80%**	**60%**	**40%**	**0%**[(a)]
AUC	0.7011 ± 0.1143	0.6590 ± 0.1195	0.6515 ± 0.1107	0.6515 ± 0.1535	0.6316 ± 0.0629
Precision	0.4246 ± 0.1395	0.3883 ± 0.1885	0.3432 ± 0.1236	0.3680 ± 0.1529	0.2857 ± 0.0419
Sensitivity	0.6600 ± 0.2538	0.5600 ± 0.2653	0.5600 ± 0.2653	0.6200 ± 0.2441	0.8000 ± 0.1549
Specificity	0.7421 ± 0.1341	0.7579 ± 0.1582	0.7368 ± 0.1246	0.6895 ± 0.1341	0.4632 ± 0.1306

[(a)] Fully supervised model — baseline model.

As can be observed, reducing the amount of unlabelled data given as input to the models results in slightly worse performances. The variation between different percentages (80%, 60% and 40%) is almost inexistent when no manifold regularisation is applied, being small when this

Table 6.5: Performance results for different percentages of the unlabelled dataset used when considering a weighted loss function and a manifold regularisation term.

	Percentage unlabelled dataset used			
Metric	**100%**	**80%**	**60%**	**40%**
AUC	0.6648 ± 0.1252	0.6311 ± 0.1204	0.6179 ± 0.1054	0.6132 ± 0.1301
Precision	0.3668 ± 0.1338	0.3883 ± 0.1885	0.3480 ± 0.1077	0.3330 ± 0.1155
Sensitivity	0.6400 ± 0.2498	0.5600 ± 0.2653	0.5200 ± 0.2227	0.5000 ± 0.2720
Specificity	0.6895 ± 0.1401	0.7263 ± 0.1485	0.7258 ± 0.1228	0.7263 ± 0.1124

Table 6.6: Average running times (in minutes) for different percentages of the unlabelled dataset used when considering a weighted loss function with and without a manifold regularisation term (average training time for each random train-test split).

	Percentage unlabelled dataset used			
Models	**100%**	**80%**	**60%**	**40%**
Without manifold regularisation	65	63	55	38
With manifold regularisation	84	81	71	51

term is added. Nevertheless, utilising unlabelled data, even if in smaller amounts, provides better models than not using them at all.

As a result, the combination that achieved the best results was using a weighted loss function to tackle the imbalance of the labelled data, without adding a regularisation term, with a mean AUC of 0.7011 ± 0.1143, illustrated in the mean ROC curve of Figure 6.2.

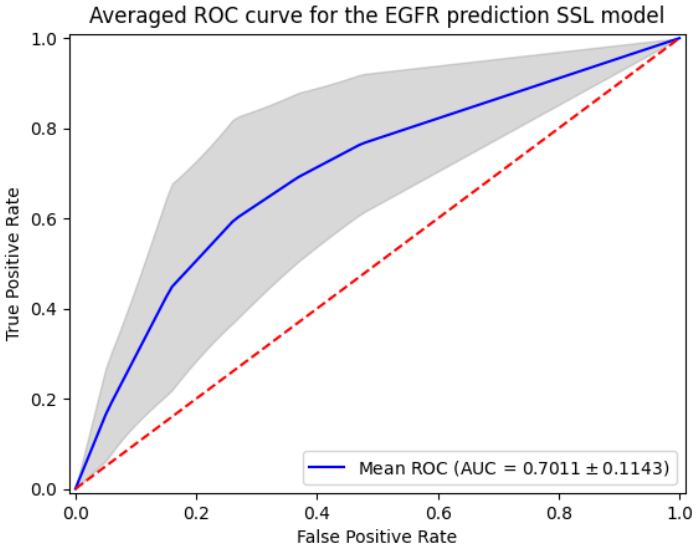


Figure 6.2: Averaged ROC curve for *EGFR* mutation status prediction with the best-performing SSL model. For each train-test split, the ROC curve is computed. The blue line represents the arithmetic average ROC curve, and the grey shaded area depicts the standard deviation. The red dashed line illustrates the ROC curve of an at-chance classifier.

It was not possible to build any acceptable model when including the UHCSJ dataset. When applying the two best aforementioned models with the new, extended labelled set, average AUCs of 0.5249 ± 0.0645 and 0.4921 ± 0.1005 were recorded with and without, respectively, the manifold regularisation term. Likewise, when a simple translation of the best networks architectures to a 2D approach was tested, the attained mean AUC was not able to go further than 0.5889 ± 0.1130.

6.2 Discussion

The results of this research display the difficulty in detecting relevant and significant features that could be related to *EGFR* mutation status. Conditioned by the limited amount of labelled data available, a more robust classification model was tried to be achieved by incorporating unlabelled data in a semi-supervised approach. Even with this extra amount of data, the task has proven to be quite challenging and susceptible to the train-test split variations, as can be noticed by the high standard deviation value across all experiments. This high variation can also be a consequence of the small number of included *EGFR* mutant patients (23 cases), being desirable to add extra samples of this class to verify if the variation would be reduced.

When tackling the imbalance present in the data, using a weighted loss function demonstrated to be a better approach when compared to oversampling the minority class. This has the additional benefit of reducing the required training time as the number of images is fewer. Undersampling the majority class was never an option in this work given that one is dealing with the problem of data scarcity and dismissing valuable data seems counterintuitive. The best-performing models were obtained when using only one data source as labelled data, and this can have some possible explanations:

- the UHCSJ dataset was only included in this project in a later stage. By that time, the tunning of the base model was almost concluded, lacking time to restart the entire process with the newly available data. Perhaps if the experimental process had been started with the three datasets, a better model that could incorporate more robustly all the data might have been obtained;
- the differences observed when comparing both labelled datasets (Section 4.1.4). Having an additional data collection with such distinctive nodules was expected to help improve the robustness of the developed model, increasing its generalisation capability. However, this did not happen. One must bear in mind that two different data sources were already being combined, and adding a third one, whose images were also acquired with different protocols, could not be helpful. Additionally, the effect of the considered mutation in distinct stages of cancer might be translated into different visual manifestations of the target variable (if such evidence exists). By combining distinct datasets, it is assumed that such manifestation is similar in terms of image patterns, though there is still no clear evidence that this happens.

When using the full unlabelled dataset, instead of only a portion of it, the model benefits more from the additional information provided by this data collection. Furthermore, the increased

performance is even more notorious when comparing the model with a fully supervised baseline, which uses only the labelled data as input. A direct comparison with related works concerning the effect of such variations cannot be made, given that, traditionally, the variation of the labelled-unlabelled data proportions is performed by adding labels to the desired part of the unannotated data and not by removing data without label, as was tested in this study.

The tested 2D approach, performed by simply adapting the best networks architectures to a 2D version, with little optimisation involved, attained worse results. This approach used the CT scan slice containing the largest nodule area instead of the entire volumetric region, maintaining the lung with nodule as the analysed ROI. Again, with more time available, better fine-tuning could be performed allowing for an improved comparison and possibly better results with this approach.

Regarding related works in *EGFR* assessment in lung cancer CT scans, to the best of gathered knowledge, no other research attempted a 3D deep learning approach using the entire lung volumetric region in a semi-supervised fashion. Furthermore, SSL methods are typically tested using extended labelled datasets and simply removing the label of a significant portion of the data, simulating the existence of a large unlabelled set. Approaches which combine different datasets in a similar way as the presented in this work are difficult to find. Additionally, no study was found combining the two primary datasets used in this task (NSCLC-Radiogenomics and NLST). Silva *et al.* [58] developed a DL model based on transfer learning methods using 2D CT scan slices from the NSCLC-Radiogenomics dataset. Utilising the analysis of the lung containing the nodule as ROI, a mean AUC of 0.68 ± 0.08 was achieved. Comparing the results, the developed SSL approach was able to slightly increase these figures, using the same labelled dataset, although increasing the variation as well.

The proposed model was not able to surpass state-of-the-art results in predicting *EGFR* mutation status. It would be interesting to overcome the current hardware limitations and test more complex architectures that could capture more accurately the complexity of the patterns that these mutations might induce.

Chapter 7

Conclusions

A personalised treatment plan presents the opportunity to improve lung cancer patient outcomes. In the era of precision medicine, the identification of driver mutations in lung cancer brought new treatment options and helped increase the overall survival rates. For this reason, a complete cancer characterisation is of utmost importance to choose the best treatment for each individual. Given the well-known heterogeneous nature of tumours, many times requiring several assessments with traditional techniques, the development of less invasive methods that could provide such a characterisation is crucial. This opens doors to the use of artificial intelligence, which is gaining ground in the medical field as the utilisation of images such as Computed Tomography scans has already proven to allow the detection of relevant patterns and relations. Despite the success of deep learning models when it comes to analysing such medical images, the lack of labelled data makes their development difficult.

This study aimed to provide an end-to-end lung cancer characterisation by analysing the entire volumetric region of the lung containing the nodule, using Computed Tomography images, in a semi-supervised approach. The method employed to integrate both label and unlabelled data consisted of a combination of a Variational Autoencoder and a Generative Adversarial Network. The best-performing classification model achieved a mean AUC of 0.7011 ± 0.1143. Despite not improving largely the performance results in this task, the utilisation of the additional unlabelled dataset brought more discriminative power to the model. This was further evidenced by the reduction in terms of performance when less unlabelled data was used, with the mean AUC dropping to 0.6316 ± 0.0629 when only labelled data was given as input.

It should be noticed that the best model was built only with 14% of the data containing label. Adding an unlabelled dataset, in an SSL fashion, improved the performance of the predictive deep learning model, allowing the development of a better-performing end-to-end model with a reduced amount of labelled data.

The developed work was constrained by hardware limitations which, when analysing images in a 3D perspective, may be cumbersome given the density of the networks involved. If this problem could be overcome, it would be helpful to explore, in the future, more complex architectures that might be able to capture more abstract patterns, given the demonstrated complexity of the

task. Furthermore, although features extracted from the unlabelled data to discriminate images had helped when classifying scans as mutant or wild-type, more samples from the minority class, which included only 23 images, would probably be needed for a more accurate model.

Albeit not reaching remarkable results, the developed work may be seen as a stepping-stone from which subsequent works can improve upon the highlighted limitations.

References

[1] Cancer history. https://www.news-medical.net/health/Cancer-History.aspx. [Online; accessed on 2022-01-08].

[2] Cancer today - international agency for research on cancer. https://gco.iarc.fr/today/home. [Online; accessed on 2022-03-07].

[3] CT Image Module - DICOM Standard Browser. https://dicom.innolitics.com/ciods/ct-image/ct-image. [Online; accessed on 2022-08-015].

[4] Pydicom package for python. https://pypi.org/project/pydicom/. [Online; accessed on 2022-10-03].

[5] The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Online; accessed on 2021-10-30].

[6] Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine 365* (8 2011), 395–409. [Online]. Available: https://www.nejm.org/doi/full/10.1056/Nejmoa1102873.

[7] ABERLE, D. R. The national lung screening trial: Overview and study design. *Radiology. 258*, 1 (2010-11).

[8] AERTS, H. J., VELAZQUEZ, E. R., LEIJENAAR, R. T., PARMAR, C., GROSSMANN, P., CAVALHO, S., BUSSINK, J., MONSHOUWER, R., HAIBE-KAINS, B., RIETVELD, D., HOEBERS, F., RIETBERGEN, M. M., LEEMANS, C. R., DEKKER, A., QUACKENBUSH, J., GILLIES, R. J., AND LAMBIN, P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications 2014 5:1 5(1)* (June 2014), 1–9. [Online]. Available: http://doi.org/10.1038/ncomms5006.

[9] AL-AZZAM, N., AND SHATNAWI, I. Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. *Annals of Medicine and Surgery 62* (Feb 2021), 53–64. [Online]. Available: http://doi.org/10.1016/J.AMSU.2020.12.043.

[10] ALEKSAKHINA, S. N., AND IMYANITOV, E. N. Cancer therapy guided by mutation tests: Current status and perspectives. *International Journal of Molecular Sciences 22* (Oct 2021). [Online]. Available: http://doi.org/10.3390/IJMS222010931.

[11] BAKR, S., GEVAERT, O., ECHEGARAY, S., AYERS, K., ZHOU, M., SHAFIQ, M., ZHENG, H., BENSON, J. A., ZHANG, W., LEUNG, A. N., KADOCH, M., HOANG, C. D., SHRAGER, J., QUON, A., RUBIN, D. L., PLEVRITIS, S. K., AND NAPEL, S. A radiogenomic dataset of non-small cell lung cancer. *Scientific data 5* (2018). [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30325352.

[12] BELL, D., AND NADRLJANSKI, M. Computed tomography. *Radiopaedia.org* (Mar 2010). [Online]. Available: http://doi.org/10.53347/RID-9027.

[13] BODALAL, Z., TREBESCHI, S., NGUYEN-KIM, T. D. L., SCHATS, W., AND BEETS-TAN, R. Radiogenomics: bridging imaging and genomics. *Abdominal Radiology 2019 44:6 44* (May 2019), 1960–1984. [Online]. Available: http://doi.org/10.1007/S00261-019-02028-W.

[14] CHANG, K., BAI, H. X., ZHOU, H., SU, C., BI, W. L., AGBODZA, E., KAVOURIDIS, V. K., SENDERS, J. T., BOARO, A., BEERS, A., ZHANG, B., CAPELLINI, A., LIAO, W., SHEN, Q., LI, X., XIAO, B., CRYAN, J., RAMKISSOON, S., RAMKISSOON, L., LIGON, K., WEN, P. Y., BINDRA, R. S., WOO, J., ARNAOUT, O., GERSTNER, E. R., ZHANG, P. J., ROSEN, B. R., YANG, L., HUANG, R. Y., AND KALPATHY-CRAMER, J. Residual Convolutional Neural Network for the determination of IDH status in low- and high-grade gliomas from MR Imaging. *Clinical Cancer Research 24* (Mar 2018), 1073–1081. [Online]. Available: http://doi.org/10.1158/1078-0432.CCR-17-2236.

[15] CHAPELLE, O., SCHÖLKOPF, B., AND ZIEN, A., Eds. *Semi-Supervised Learning*. The MIT Press, 2006.

[16] CHOI, Y. S., BAE, S., CHANG, J. H., KANG, S.-G., KIM, H., KIM, J., RIM, T. H., CHOI, H., JAIN, R., AND LEE, S.-K. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro-Oncology 23* (2021), 304–313. [Online]. Available: http://doi.org/10.1093/neuonc/noaa177.

[17] COVER, T. M., AND THOMAS, J. A. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[18] DAS, A., DEVARAMPATI, V. K., AND NAIR, M. S. Nas-sgan: A semi-supervised generative adversarial network model for atypia scoring of breast cancer histopathological images. *IEEE Journal of Biomedical and Health Informatics* (2021), 1–1. [Online]. Available: http://doi.org/10.1109/JBHI.2021.3131103.

[19] DIGUMARTHY, S. R., PADOLE, A. M., GULLO, R. L., SEQUIST, L. V., AND KALRA, M. K. Can CT radiomic analysis in NSCLC predict histology and EGFR mutation status? *Medicine 98(1)* (Jan. 2019), e13963. [Online]. Available: http://doi.org/10.1097/MD.0000000000013963.

[20] DOERSCH, C. Tutorial on variational autoencoders, 2016. [Online]. Available: https://arxiv.org/abs/1606.05908.

[21] ELLISON, G., ZHU, G., MOULIS, A., DEARDEN, S., SPEAKE, G., AND MCCORMACK, R. EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples. *Journal of Clinical Pathology 66* (Feb 2013), 79–89. [Online]. Available: http://doi.org/10.1136/JCLINPATH-2012-201194.

[22] FAGUET, G. B. A brief history of cancer: Age-old milestones underlying our current knowledge database. *International Journal of Cancer 136* (May 2015), 2022–2036. [Online]. Available: http://doi.org/10.1002/IJC.29134.

[23] FERREIRA, F. T., SOUSA, P., GALDRAN, A., SOUSA, M. R., AND CAMPILHO, A. End-to-end supervised lung lobe segmentation. In *2018 International Joint Conference on Neural Networks (IJCNN)* (2018), pp. 1–8. [Online]. Available: http://doi.org/10.1109/IJCNN.2018.8489677.

[24] GILLIES, R., BOELLARD, R., DEKKER, A., AND AERTS, H. J. W. L. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer 48(4)* (2012), 441–446. [Online]. Available: http://doi.org/10.1016/j.ejca.2011.11.036.

[25] GILLIES, R. J., KINAHAN, P. E., AND HRICAK, H. Radiomics: Images are more than pictures, they are data. *Radiology 278(2)* (Feb. 2016), 563–577. [Online]. Available: http://doi.org/10.1148/radiol.2015151169.

[26] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc.

[27] GOODFELLOW, I. J., BENGIO, Y., AND COURVILLE, A. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[28] HANAHAN, D., AND WEINBERG, R. A. The hallmarks of cancer. *Cell 100*, 1 (Jan. 2000), 57–70. [Online]. Available: https://doi.org/10.1016/S0092-8674(00)81683-9.

[29] HE, K., LIU, X., LI, M., LI, X., YANG, H., AND ZHANG, H. Noninvasive KRAS mutation estimation in colorectal cancer using a deep learning method based on CT imaging. *BMC Medical Imaging 20* (June 2020), 1–9. [Online]. Available: http://doi.org/10.1186/S12880-020-00457-4.

[30] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December* (Dec. 2015), 770–778. [Online]. Available: http://doi.org/10.1109/CVPR.2016.90.

[31] HUANG, G., LIU, Z., MAATEN, L. V. D., AND WEINBERGER, K. Q. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-January* (Aug. 2016), 2261–2269. [Online]. Available: http://doi.org/10.1109/CVPR.2017.243.

[32] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 07–09 Jul 2015), F. Bach and D. Blei, Eds., vol. 37 of *Proceedings of Machine Learning Research*, PMLR, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html.

[33] KAHN, C. E., CARRINO, J. A., FLYNN, M. J., PECK, D. J., AND HORII, S. C. Dicom and radiology: Past, present, and future. *Journal of the American College of Radiology 4* (2007), 652–657. [Online]. Available: http://doi.org/10.1016/j.jacr.2007.06.004.

[34] KIM, S., LEE, Y., ADHI TAMA, B., AND LEE, S. Reliability-enhanced camera lens module classification using semi-supervised regression method. *Applied Sciences 10* (May 2020), 3832. [Online]. Available: http://doi.org/10.3390/app10113832.

[35] KINGMA, D. P., AND WELLING, M. Auto-encoding variational bayes, 2013. [Online]. Available: https://arxiv.org/abs/1312.6114.

[36] KOCAK, B., DURMAZ, E. S., ATES, E., AND ULUSAN, M. B. Radiogenomics in Clear Cell Renal Cell Carcinoma: Machine learning–based high-dimensional quantitative CT texture analysis in predicting PBRM1 mutation status. *https://doi.org/10.2214/AJR.18.20443 212* (Jan. 2019), W55–W63. [Online]. Available: http://doi.org/10.2214/AJR.18.20443.

[37] LARSEN, A. B. L., SØNDERBY, S. K., AND WINTHER, O. Autoencoding beyond pixels using a learned similarity metric. *CoRR abs/1512.09300* (2015). [Online]. Available: http://arxiv.org/abs/1512.09300.

[38] LEE, Y. T., TAN, Y. J., AND OON, C. E. Molecular targeted therapy: Treating cancer with specificity. *European Journal of Pharmacology 834* (2018), 188–196. [Online]. Available: http://doi.org/10.1016/j.ejphar.2018.07.034.

[39] MARTINS, R. A. P., AND SILVA, D. On teacher-student semi-supervised learning for chest X-ray image classification. [Online]. Available: http://doi.org/10.21528/CBIC2021-80.

[40] MONDELO-MACÍA, P., GARCÍA-GONZÁLEZ, J., LEÓN-MATEOS, L., CASTILLO-GARCÍA, A., LÓPEZ-LÓPEZ, R., MUINELO-ROMAY, L., AND DÍAZ-PEÑA, R. Current status and future perspectives of liquid biopsy in small cell lung cancer. *Biomedicines 9(1)* (Jan. 2021). [Online]. Available: http://doi.org/10.3390/biomedicines9010048.

[41] MORGADO, J., PEREIRA, T., SILVA, F., FREITAS, C., NEGRÃO, E., DE LIMA, B. F., DA SILVA, M. C., MADUREIRA, A. J., RAMOS, I., HESPANHOL, V., COSTA, J. L., CUNHA, A., AND OLIVEIRA, H. P. Machine learning and feature selection methods for EGFR mutation status prediction in lung cancer. *Applied Sciences 11*, 7 (2021). [Online]. Available: https://www.mdpi.com/2076-3417/11/7/3273.

[42] NOONE, A., HOWLADER, N., KRAPCHO, M., MILLER, D., BREST, A., YU, M., RUHL, J., TATALOVICH, Z., MARIOTTO, A., LEWIS, D., CHEN, H., FEUER, E., AND CRONIN, K. Seer cancer statistics review, 1975-2017, National Cancer Institute. https://seer.cancer.gov/csr/1975_2017/. [Online; accessed on 2022-01-05].

[43] NOORELDEEN, R., AND BACH, H. Current and future development in lung cancer diagnosis. *International Journal of Molecular Sciences 2021, Vol. 22, Page 8661 22* (Aug. 2021), 8661. [Online]. Available: https://www.mdpi.com/1422-0067/22/16/8661.

[44] PINHEIRO, G., PEREIRA, T., DIAS, C., FREITAS, C., HESPANHOL, V., COSTA, J. L., CUNHA, A., AND OLIVEIRA, H. P. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Scientific Reports 10* (Dec. 2020). [Online]. Available: http://doi.org/10.1038/s41598-020-60202-3.

[45] QIAN, C. N., MEI, Y., AND ZHANG, J. Cancer metastasis: issues and challenges. *Chinese Journal of Cancer 36* (2017), 38. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5379757.

[46] RADFORD, A., METZ, L., AND CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. [Online]. Available: https://arxiv.org/abs/1511.06434.

[47] RIELY, G. J., KRIS, M. G., ROSENBAUM, D., MARKS, J., LI, A., CHITALE, D. A., NAFA, K., RIEDEL, E. R., HSU, M., PAO, W., MILLER, V. A., AND LADANYI, M. Frequency and distinctive spectrum of kras mutations in never smokers with lung adenocarcinoma. *Clinical*

cancer research : an official journal of the American Association for Cancer Research 14 (Sept. 2008), 5731. [Online]. Available: http://doi.org/10.1158/1078-0432.CCR-08-0646.

[48] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M. S., BERG, A. C., AND FEI-FEI, L. Imagenet large scale visual recognition challenge. *CoRR abs/1409.0575* (2014). [Online]. Available: http://arxiv.org/abs/1409.0575.

[49] RUSSANO, M., NAPOLITANO, A., RIBELLI, G., IULIANI, M., SIMONETTI, S., CITARELLA, F., PANTANO, F., DELL'AQUILA, E., ANESI, C., SILVESTRIS, N., ARGENTIERO, A., SOLIMANDO, A., VINCENZI, B., TONINI, G., AND SANTINI, D. Liquid biopsy and tumor heterogeneity in metastatic solid tumors: The potentiality of blood samples. *Journal of Experimental and Clinical Cancer Research 39* (May 2020), 1–13. [Online]. Available: http://doi.org/10.1186/S13046-020-01601-2.

[50] SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., CHEUNG, V., RADFORD, A., CHEN, X., AND CHEN, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems* (2016), D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc. [Online]. Available: https://doi.org/10.48550/arXiv.1606.03498.

[51] SCHWARTZBERG, L., KIM, E. S., LIU, D., AND SCHRAG, D. Precision oncology: Who, how, what, when, and when not? *American Society of Clinical Oncology Educational Book*, 37 (2017), 160–169. [Online]. Available: http://doi.org/10.1200/EDBK_174176.

[52] SEEBACHER, N. A., STACY, A. E., PORTER, G. M., AND MERLOT, A. M. Clinical development of targeted and immune based anti-cancer therapies. *Journal of Experimental & Clinical Cancer Research 2019 38:1 38* (Apr. 2019), 1–39. [Online]. Available: http://doi.org/10.1186/S13046-019-1094-2.

[53] SHOLL, L. M., AND MINO-KENUDSON, M. Tumors of the lung and pleura. In *Diagnostic Histopathology of Tumors*, C. D. M. Fletcher, Ed., fifth ed. Elsevier, 2020, ch. 5, pp. 225–272.

[54] SHUI, L., REN, H., YANG, X., LI, J., CHEN, Z., YI, C., ZHU, H., AND SHUI, P. The era of radiogenomics in precision medicine: An emerging approach to support diagnosis, treatment decisions, and prognostication in oncology. *Frontiers in Oncology* (Jan. 2021), 3195.

[55] SIEGEL, R. L., MILLER, K. D., FUCHS, H. E., AND JEMAL, A. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians 71(1)* (Jan 2021), 7–33. [Online]. Available: http://doi.org/10.3322/CAAC.21654.

[56] SIEREN, J. C., OHNO, Y., KOYAMA, H., SUGIMURA, K., AND MCLENNAN, G. Recent technological and application developments in computed tomography and magnetic resonance imaging for improved pulmonary nodule detection and lung cancer staging. *Journal of Magnetic Resonance Imaging 32* (Dec. 2010), 1353–1369. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/jmri.223837.

[57] SILVA, F., PEREIRA, T., MORGADO, J., CUNHA, A., AND OLIVEIRA, H. P. The impact of interstitial diseases patterns on lung CT segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2021), pp. 2856–2859. [Online]. Available: https://ieeexplore.ieee.org/document/9630354.

[58] SILVA, F., PEREIRA, T., MORGADO, J., FRADE, J., MENDES, J., FREITAS, C., NEGRÃO, E., DE LIMA, B. F., SILVA, M. C. D., MADUREIRA, A. J., RAMOS, I., HESPANHOL, V., COSTA, J. L., CUNHA, A., AND OLIVEIRA, H. P. EGFR assessment in lung cancer CT images: Analysis of local and holistic regions of interest using deep unsupervised transfer learning. *IEEE Access 9* (2021), 58667–58676. [Online]. Available: https://ieeexplore.ieee.org/document/9393886.

[59] SINGH, G., MANJILA, S., SAKLA, N., TRUE, A., WARDEH, A. H., BEIG, N., VAYSBERG, A., MATTHEWS, J., PRASANNA, P., AND SPEKTOR, V. Radiomics and radiogenomics in gliomas: a contemporary update. *British Journal of Cancer 2021 125:5 125* (May 2021), 641–657. [Online]. Available: http://doi.org/10.1038/s41416-021-01387-w.

[60] SONG, Z., YANG, X., XU, Z., AND KING, I. Graph-based semi-supervised learning: A comprehensive review, 2021. [Online]. Available: http://arxiv.org/abs/2102.13303.

[61] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 56 (2014), 1929–1958. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html.

[62] SUN, W., TSENG, T.-L. B., ZHANG, J., AND QIAN, W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics 57* (2017), 4–9. [Online]. Available: http://doi.org/10.1016/j.compmedimag.2016.07.004.

[63] TOMASHEFSKI, J. F., AND FARVER, C. F. *Anatomy and Histology of the Lung*. Springer New York, New York, NY, 2008, pp. 20–48. [Online]. Available: https://doi.org/10.1007/978-0-387-68792-6_2.

[64] TRAVIS, W. D. Pathology of lung cancer. *Clin Chest Med 32(4)* (Dec 2011). [Online]. Available: http://doi.org/10.1016/j.ccm.2011.08.005.

[65] TRIVIZAKIS, E., PAPADAKIS, G. Z., SOUGLAKOS, I., PAPANIKOLAOU, N., KOUMAKIS, L., SPANDIDOS, D. A., TSATSAKIS, A., KARANTANAS, A. H., AND MARIAS, K. Artificial intelligence radiogenomics for advancing precision and effectiveness in oncologic care (Review). *International Journal of Oncology 57* (July 2020), 43–53. [Online]. Available: http://doi.org/10.3892/IJO.2020.5063.

[66] VAN ENGELEN, J. E., AND HOOS, H. H. A survey on semi-supervised learning. *Machine Learning 109* (Feb. 2020), 373–440. [Online]. Available: http://doi.org/10.1007/S10994-019-05855-6.

[67] WANG, S., SHI, J., YE, Z., DONG, D., YU, D., ZHOU, M., LIU, Y., GEVAERT, O., WANG, K., ZHU, Y., ZHOU, H., LIU, Z., AND TIAN, J. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *European Respiratory Journal 53* (Mar. 2019). [Online]. Available: http://doi.org/10.1183/13993003.00986-2018.

[68] WANG, S., YU, B., NG, C. C., MERCORELLA, B., SELINGER, C. I., O'TOOLE, S. A., AND COOPER, W. A. The suitability of small biopsy and cytology specimens for EGFR and other mutation testing in non-small cell lung cancer. *Translational Lung Cancer Research 4* (2015), 119. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384210/.

[69] XIE, Y., ZHANG, J., AND XIA, Y. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT. *Medical Image Analysis 57* (Oct. 2019), 237–248. [Online]. Available: http://doi.org/10.1016/J.MEDIA.2019.07.004.

[70] YALNIZ, I. Z., JÉGOU, H., CHEN, K., PALURI, M., AND MAHAJAN, D. Billion-scale semi-supervised learning for image classification, 2019. [Online]. Available: https://arxiv.org/abs/1905.00546v1.

[71] YANG, X., SONG, Z., KING, I., AND XU, Z. A survey on deep semi-supervised learning. [Online]. Available: https://arxiv.org/abs/2103.00550v2.

[72] YE, Z., HUANG, Y., KE, J., ZHU, X., LENG, S., AND LUO, H. Breakthrough in targeted therapy for non-small cell lung cancer. *Biomedicine & Pharmacotherapy 133* (2021), 111079. [Online]. Available: http://doi.org/10.1016/j.biopha.2020.111079.

[73] ZHANG, Y.-L., YUAN, J.-Q., WANG, K.-F., FU, X.-H., HAN, X.-R., THREAPLETON, D., YANG, Z.-Y., MAO, C., AND TANG, J.-L. The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis, 2016. [Online]. Available: http://doi.org/10.18632/oncotarget.12587.

[74] ZHENG, D. Review of radiomics and radiogenomics and big data in radiation oncology. *Journal of Applied Clinical Medical Physics 21* (Aug. 2020), 326–328. [Online]. Available: http://doi.org/10.1002/ACM2.12955.

[75] ZHENG, M. Classification and pathology of lung cancer. *Surgical Oncology Clinics of North America 25(3)* (2016), 447–468. [Online]. Available: http://doi.org/10.1016/j.soc.2016.02.003.