FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Multiple instance learning for lung cancer characterization in computed tomography scans

Julieta Pintado Jorge Frade

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Tânia Maria Pereira Lopes, PhD

Co-Supervisor: Hélder Filipe Pinto de Oliveira, PhD

Co-Supervisor: António Manuel Trigueiros da Silva Cunha, PhD

July 25, 2020

Multiple instance learning for lung cancer characterization in computed tomography scans

Julieta Pintado Jorge Frade

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. Rui Carlos Camacho de Sousa Ferreira da Silva
External Examiner: Prof. Petia Georgieva
Supervisor: Prof. Tânia Maria Pereira Lopes
Co-Supervisor: Prof. Hélder Filipe Pinto de Oliveira

July 25, 2020

Abstract

Lung cancer occurs when cells belonging to the lung grow out of control, disturbing regular cells and making it difficult for the body to function correctly. This type of cancer is the second most common in men and women.

In terms of treatments, the evolution of personalized medicine changed the therapeutic strategy from classical chemotherapy and radiotherapy to genetic modification targeted therapy. However, a meticulous tumor characterization is a fundamental requirement for these types of treatment.

Firstly, chest radiographs and computed tomography disclose lung cancer and later the diagnosis is validated by extorting lung tissue in a biopsy to characterize the tumor. However, a biopsy is associated with some issues including discomfort and pain for the patient and the probability of surgical difficulties, suggesting the importance to explore non-invasive methods.

Gene mutation predictive machine learning or deep learning models based on medical images obtain a vast amount of features from visual aspects to a model that can predict the mutated gene, presenting the benefits of being non-invasive, fast and easy to use. Beforehand, image features extracted from cancer nodules have been used to create predictive models for this problem. However, results from the literature hint that features from lung structures external to the nodule might be relevant to foretell the mutation status in lung cancer.

Multiple Instance Learning interprets the relationship between a label and the bag of instances responsible for it. Only the bag label is known, and the goal is to train the model to be able to classify new bags. This technique already proved to work well in the detection of diseases using medical images as bags.

This study aims to examine how Multiple Instance Learning can be applied to identify the presence of lung patterns in CT scans so that, in the future, these patterns help predict the mutated gene in lung cancer in a non-invasive and understandable way.

The detected lung patterns in this study were Emphysema, Satellite Nodules In Primary Lesion Lobe, Nodules In Contralateral Lung, Fibrosis and Ground Glass, being Fibrosis and Emphysema the ones with more outstanding results, reaching an AUC of 0.89 and 0.72, respectively.

Keywords: Multiple Instance Learning, Feature Engineering, Lung Cancer, Lung Patterns Detection

Resumo

Cancro do pulmão ocorre quando as células pertencentes ao pulmão se começam a desenvolver sem controlo, perturbando as células saudáveis e tornando difícil para o corpo operar corretamente. Este tipo de cancro é o segundo mais frequente tanto em homens como mulheres.

Em termos de tratamento, a evolução de medicina personalizada melhorarou a estratégia terapêutica das tradicionais quimioterapia e radioterapia para terapia direcionada à modificação genética. Contudo, o tumor e as suas imediações têm que ser meticulosamente caracterizadas para estes tipos de tratamento.

Cancro do pulmão é primeiramente detetado em radiografias e tomografias computorizadas do peito e, mais tarde, o diagnóstico é confirmado extraindo uma amostra de tecido do pulmão numa biópsia para caracterizar o tumor. Contudo, a biópsia está associada a alguns problemas incluindo desconforto e dor para o paciente, vários riscos clínicos e a possibilidade de complicações cirúrgicas, aumentando a urgência de procurar métodos não invasivos.

Modelos de machine learning para prever mutações de genes baseados em imagens médicas obtêm uma vasta quantia de características a partir de aspetos visuais para um modelo que consegue prever o gene mutado, apresentando os benefícios de ser não invasivo, rápido e fácil de usar. Anteriormente, características da imagem extraídas de nódulos de cancro foram usadas para criar modelos preditivos para este problema. Contudo, estudos recentes sugerem que características de estruturas do pulmão externas ao nódulo poderão ser relevantes para prever o estado da mutação no cancro do pulmão.

Multiple Instance Learning interpreta a relação entre uma etiqueta e um saco de instâncias por ela responsáveis. Só a etiqueta dos sacos é conhecida e o objetivo é treinar o modelo de modo a ser capaz de classificar novos sacos. Esta técnica já provou ser eficaz na deteção de doenças usando imagens médicas.

Este estudo pretence examinar como Multiple Instance Learning pode ser aplicado na identificação da presença de padrões pulmonares em scans CT para que, no futuro, estes padrões possam ajudar a prever o gene mutado no cancro do pulmão numa maneira não invasiva e compreensível.

Os padrões do pulmão detetados neste estudo foram Emphysema, Satellite Nodules In Primary Lesion Lobe, Nodules In Contralateral Lung, Fibrosis e Ground Glass, sendo que Fibrosis e Emphysema foram as deteções com melhores resultados, atingindo um AUC de 0.89 e 0.72, respetivamente.

Keywords: Multiple Instance Learning, Feature Engineering, Cancro do Pulmão, Deteção de Padrões do Pulmão

Acknowledgements

Firstly, I want to thank my supervisors, Tânia Pereira, Hélder Oliveira, and António Cunha, for demanding my best, encouraging me to not stop questioning the world, and showing me the right path to follow through this whole study.

I also want to thank my college friends that, even though they were not physically present, helped me exceed this period of my life and many others.

Foremost, I want to thank my parents for being my only and best companions for four months and providing me with everything and anything I needed to complete this task.

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

Julieta

"Into the light that spreads out before us,
Towards the future that no one knows of"

EXO

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the ROC Curve
CAD	Computer-Aided Diagnosis
CH-FD	Convex-Hull Fisher’s Discriminnat
CNNs	Convolutional Neural Networks
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
CXRs	Chest X-Rays
DL	Deep Learning
EGFR	Epidermal Growth Factor Receptor
EM-DD	Evolution Expectation Maximization Diverse Density
FD	Fisher’s Discriminnat
GGO	Ground-Glass Opacity
GLCM	Gray-Level Co-Occurrence Matrix
HU	Hounsfield Units
ILDs	Interstitial Lung Diseases
KNN	K-Nearest Neighbor
KDE	Kernel Density Estimation
KRAS	Kristen Rat Sarcoma Viral Oncogene Homolog
MAD	Median Absolute Deviation
ML	Machine Learning
MICA	Multiple Instance Classification Algorithm
MIL	Multiple Instance Learning
MissSVM	Multiple Instance Learning by Semi-Supervised SVM
MISVM	Multiple-Instance Support Vector Machines
MLP-NN	Multi Layer Perceptron Neural Networks
NSK	Normalized Set Kernel
NSCLC	Non-Small-Cell Lung Carcinoma
OR	Odds Ratio
PE	Pulmonary Embolism
PET	Positron-emission Tomography
PFS	Progression-free Survival
RF	Random Forest
ROC	Receiver Operating Characteristics
SB	Single Blob With No Neighbors
sbMIL	Sparse Balanced Multiple Instance Learning
SCLC	Small-Cell Carcinoma
SIL	Single-Instance Learning
sMIL	Sparse Multiple Instance Learning

SMOTE	Synthetic Minority Oversampling Technique
STK	Statistic Set Kernel
stMIL	Sparse Transductive Multiple Instance Learning
TKIs	Small-Molecule Tyrosine Kinase Inhibitors
WMD	Weighted Mean Difference
XGB	XGBoost

Chapter 1

Introduction

Lung cancer is the foremost determinant of cancer death amongst both men and women, killing a vaster number of people than colon, breast, and prostate cancers combined. This kind of cancer is the second most frequent in both men (losing to prostate cancer) and women (losing to breast cancer). For smokers, the chance of developing lung cancer is considerably higher than for non-smokers, however, if we consider both for an overall view, the risk of a man developing lung cancer is approximately 1 in 15, while for a woman, the probability is around 1 in 17. This disease is most common in older people, being 70 the average age of individuals when diagnosed [2].

Smoking is distinctly the foremost jeopardy agent for lung cancer, believed to cause approximately 80% of this disease mortality. Other danger factors individuals can avoid include secondhand smoke, exposure to radon and asbestos, taking certain dietary supplements and drinking water containing high levels of arsenic. Factors like previous radiation therapy to the lungs, air pollution, personal or family history of lung cancer are the risk factors that, unfortunately, people cannot avoid [2].

Non-Small-Cell Lung Carcinoma (NSCLC) serves roughly 80% to 85% of lung cancers. Its histological subtypes are Adenocarcinoma, Squamous Cell Carcinoma, and Large Cell Carcinoma, which start from various kinds of lung cells but are all filed under NSCLC due to their similarities in treatment and diagnosis. Considering individuals diagnosed between 2008 and 2014, people who have NSCLC present a 23% chance of 5-year-survival, which means that those who have NSCLC cancer are about 23% as likely as those who do not have that cancer to live for at least 5 years after being diagnosed. Small-Cell Carcinoma (SCLC) accounts for around 10% to 15% of all lung cancers and, considering people diagnosed between 2008 and 2014, individuals who have that type of lung cancer present a 6% chance of 5-year-survival. Less than 5% of all lung tumors are Lung Carcinoid Tumors [2]. Epidermal Growth Factor Receptor (*EGFR*) and Kristen Rat Sarcoma Viral Oncogene Homolog (*KRAS*) are the most frequently mutated genes that spring lung cancer of type Adenocarcinoma [45].

1.1 Motivation

Surgery, chemotherapy, radiotherapy and target therapies are the most prevalent treatments for Lung Cancer [2]. The growth of the latter has changed the therapeutic approach from conventional chemotherapy and radiotherapy to genetic modification targeted therapy [44]. The first targeted medicines for the treatment of lung cancer targeted the *EGFR* and were Small-Molecule Tyrosine Kinase Inhibitors (TKIs). In cases with mutated *EGFR*, treatments with targeted TKIs are related to higher radiographic response rates than regular chemotherapy and a more prolonged progression-free survival (PFS), which represents the length of time that a patient lives with the disease but it does not get worse [2] [35]. Nevertheless, in the case of *non-EGFR* mutated lung cancer, if some TKIs are administered, the subject will undergo a shorter PFS in contrast to chemotherapy [35].

Lung cancer is usually noticed on chest radiographs and computed tomography (CT) scans and, subsequently, the diagnosis is verified by extracting units of lung tissue in a biopsy to characterize the tumor. Nevertheless, a biopsy presents a lot of concerns. The main one is the hardship and ache that causes to the patient. Second of all, the quantity of tissue obtained may not be sufficient but the procedure cannot be repeated due to the likelihood of surgical complications. Since the tumor develops simultaneously with the disease, the biopsy becomes obsolete. Additionally, some tumors are hard or impracticable to reach. Finally, biopsies are prolonged and expensive. These problems establish a need to look for non-invasive techniques to identify and examine the tumor growth various times during the treatment [38].

Models based on medical images extract a large number of features from visual characteristics and can directly predict the mutated gene. Image biomarkers present the perks of being non-invasive, fast and easy to use, moreover medical images are accessible at a low price since they are a part of the clinical protocol [7]. Predicting gene mutation status by CT image analysis can help the experts to determine the best treatment for each patient, avoiding the biopsy [53].

A nodule is a cluster of tumoral cells and the Computer-Aided Diagnosis (CAD) models are usually based on it. However, previous results from the project “Lung Cancer Screening - A non-invasive methodology for early diagnosis” suggest that the most relevant information to predict the mutation status in lung cancer might be the combination of features from the nodule and other lung structures [45]. Semantic features from external structures to the nodule may give relevant information to the machine learning models and it may improve the accuracy of diagnosis.

1.2 Objectives

The goal of this dissertation is to investigate how Multiple Instance Learning can be applied to detect the presence of malicious lung patterns in CT scans. In future work, these detected lung patterns will be used to create predictive models that will be capable of foretelling more precisely *EGFR* and *KRAS* mutation status. Those models will have the advantage of being non-invasive and being understandable by physicians since known lung patterns are employed.

1.3 Contributions

The contributions of this dissertation include:

- A predictive model of Emphysema, Fibrosis, Ground Glass Component and Nodules presence using Multiple Instance Learning in lung CT scans;
- Elucidate how the decision of the Multiple Instance Learning bag generator influences the pattern detection result;
- Clarify how the election of the sampling technique affects the pattern detection result.

1.4 Document Structure

This document is broken into seven chapters. This chapter explains the motivation, identifies the objectives and lists the contributions of the project. The second chapter clarifies the medical concepts required to understand the problem. The Multiple Instance Learning chapter aims to explain this novel technique, as well as give some insights on how to perform its sampling on some of its models. The Literature Review chapter analysis the current state of the art and identifies what is missing. The Data Description and Preparation chapter describes the datasets used in the project, exposes the lung patterns to be detected selection criteria, describes the processing of the CT scans and explains the lung segmentation algorithm used. The sixth chapter describes the methodology that was employed, its results and the respective discussion. Lastly, the seventh chapter gives an overview of this dissertation conclusions.

Chapter 2

Lung Cancer

Lung cancer is a type of cancer that ignites in the lungs when cells develop uncontrollably, disturbing healthy cells and making it difficult for the body to function correctly. There are three main histological types of lung cancer: Non-Small-Cell Lung Carcinoma (80% to 85%), Small-Cell Carcinoma (10% and 15%) and Lung Carcinoid Tumors (less than 5%). NSCLC is the most studied one since it is the most frequent and it is divided into three histological subtypes: Adenocarcinoma, Squamous Cell Carcinoma, and Large Cell Carcinoma, being Adenocarcinoma the most frequent [2].

2.1 Mutated Genes

The most commonly mutated genes in Adenocarcinoma are *EGFR* and *KRAS* [45]. The cell surface receptor *EGFR* is responsible for cell growth and survival [24]. Its mutations promote *EGFR* permanent activation, which contributes to uncontrolled cell division [47]. The worldwide incidence of this mutation differs according to ethnicity, gender, and tobacco exposure [32]. *KRAS* mutations are a distinct cause of tumor growth. Although this is highly related to smoking, it is also detected in a substantial proportion of never-smokers [42].

The most frequent treatments for Lung Cancer are surgery, chemotherapy, and radiotherapy and, more recently, target therapies that depend on the mutated gene [2]. The evolution of the latest has improved the therapeutic strategy from traditional chemotherapy and radiotherapy to genetic modification targeted therapy. By determining their mutation status, it is possible to provide a targeted therapy for each patient. This separation can be called personalized medicine and permits distinguishing suitable treatment among chemotherapy, surgery, radiation and targeted therapy since lung cancer patients often exhibit diverging clinical results even with an identical tumor stage [27].

2.2 Biopsy

The universal practice to characterize the gene mutation is to obtain units of tumor tissue in a biopsy [45]. However, this method comes with several issues:

- Invasive surgical extraction that causes discomfort, pain, and risk for the patient;
- Multiple clinical risks that are intrinsic in surgeries and the probability of surgical complications;
- Some tumors are difficult or impossible to access;
- The amount of tissue extracted may not be enough for all the required tests;
- The tumor evolves along with the disease making the biopsy outcome obsolete;
- Biopsies are time-consuming and have a high financial cost.

These raise an urgency to seek for non-invasive methods to identify and observe the tumor evolution multiple times during the treatment [38]. This need is emphasized in the growth of the liquid biopsy approach to detect the mutated gene in cancer. Liquid biopsy is non-invasive since it employs blood samples to evaluate tumor genomics. Nonetheless, this procedure is not completely developed, not being applied without the traditional biopsy yet [25].

2.3 Medical Imaging

Medical imaging is one of the main agents that have enlightened medical science and treatment, being frequently employed in clinical method for oncologic diagnosis by computing the properties of human tissue noninvasively [4]. It offers the perks of being fast, easy to use and available at a low price [7]. Since imaging is already regularly repeated during treatment, it has the potential to supervise therapy continuously and control the rise and growth of the disease or its response to therapy [4].

CT is the strongest medical imaging approach for advanced discovery of tumors since it can distinguish several tissues depending on their varying gray levels [8]. It is regularly used and it's a vital part of all phases of cancer supervision, including prediction, screening, biopsy guidance for detection, treatment planning, treatment guidance, and treatment response evaluation [31].

Prior researches have shown that *EGFR* mutations are highly related to female patients and in Asian cultures [61]. However, deciding the type of treatment by clinical characteristics alone is incompetent [48]. Foretelling gene mutation status by CT imaging can improve the determination of the most suitable treatment for each subject, dodging biopsy [53]. Gene mutation predictive models established on medical images extort a vast number of features from visual characteristics to a model that can directly predict the mutated gene [7].

2.4 Visual Features

Studies have shown that the classification of different features of the lung CTs may give an additional understanding of the existence and development of tumors to predict prognosis, in other words, valuable knowledge to personalize lung cancer treatment [48] [56].

The CT feature extraction method can lean on the investigation of quantitative features, qualitative features or both. Lung characteristics captured in CT images are described as qualitative when using radiologist-defined semantic features and quantitative when using computer-derived "radiomic" features [51] [56].

Semantic features are regarded as qualitative considering they are scored according to the observations of experts. These characteristics can define a tumor's traits, internal structure, and external environment, being limited by what is noticeable by the eye [27] [56].

Some of the most common lung patterns are the presence of Emphysema, Fibrosis, Ground Glass Component, and Nodules. Emphysema involves the injury of the alveoli, being larger air sacs rather than multiple small ones. The quantity of oxygen that enters the bloodstream is reduced due to the decrease in the surface area of the lungs. Pulmonary Fibrosis takes place when lung tissue becomes damaged, scarred and thick, making it more challenging for the lungs to function correctly [3]. Ground Glass Component is a confined space of blurred lung opacity [28]. Finally, a Pulmonary Nodule is a round or oval regular shape enveloped in the lung parenchyma that has a defined edge and high density [33] [54].The visual manifestations of these patterns are shown in Figure 2.1.

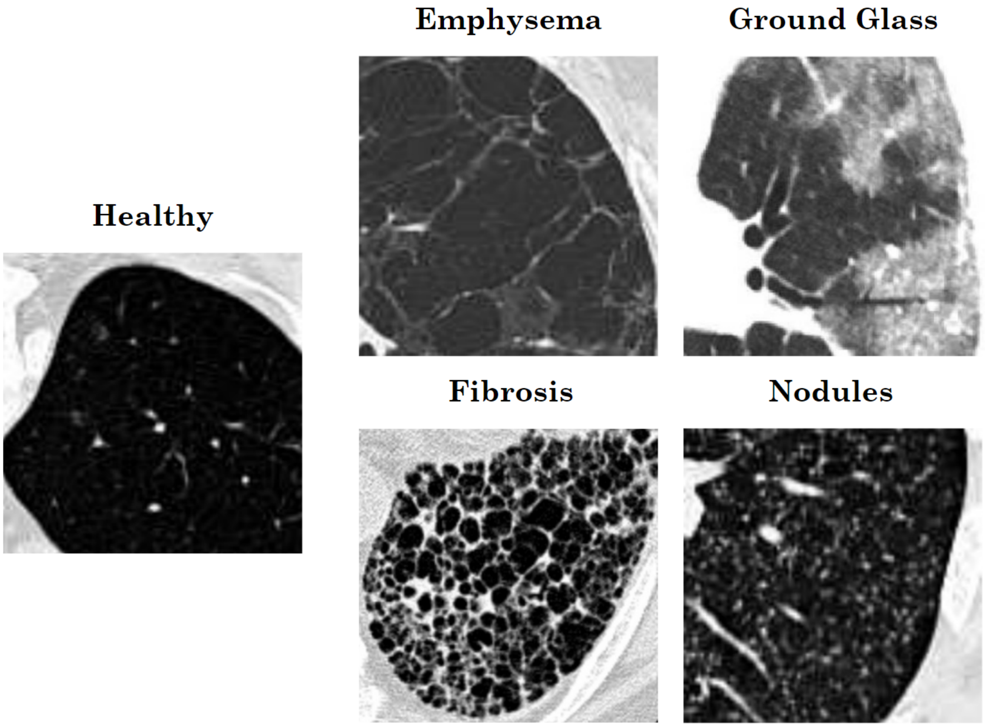


Figure 2.1: Visual manifestations of the most common lung patterns on CT scans. From Depeursinge et al. [17].

Lung features can be used by CADs to detect the gene mutation status. CADs are methods that support experts in the analysis of medical images, they can fuse components of artificial intelligence, computer vision, radiology, and pathology.

2.5 Summary

NSCLC is the most prevalent histological type of lung cancer and therefore the most investigated. The mutated gene dictates the treatment for lung cancer and biopsy is the current technique to identify it, however, this approach causes pain and discomfort to the patient.

Medical imaging is used repeatedly during cancer diagnosis and treatment and it reveals lung visual characteristics with no need for surgery. CT imaging is the most powerful type of medical imaging since it can differentiate human tissues and disclose quantitative and qualitative lung features.

Chapter 3

Multiple Instance Learning

The Multiple Instance Learning (MIL) strategy was described for the first time in 1997 by Dietterich et al. [18] and it was inspired by the Drug Activity Prediction problem. This problem aims to classify a molecule as "active" or "inactive" based on its binding with a particular protein. Each molecule can have many distinct shapes, being recognized as active if at least one of its forms is active, and inactive if all of its configurations are inactive [19]. This concept is further explained in Section 3.1, some methods to represent a MIL problem are shown in Section 3.2, a sampling technique is detailed in Section 3.3 and some of its models are described in Section 3.4.

3.1 The Concept

Since in some cases it is more fitting to classify the group rather than the individual but the latter can deliver more detailed information that the group as one cannot, MIL deciphers the relationship between a label and the set of instances responsible for it [6]. A bag is composed of various instances which are feature vectors. Solely the bag label is known, in other words, the instance classification stays unknown [29] [60]. The premise is that the bag label is positive if at least one of the instances in the bag is positive (the target concept), and negative if all the instances in the bag are negative (background instances). The intent is to train the model to be able to classify new bags that it has never seen before [19] [60]. This strategy can be regarded as a generalization of the Traditional Supervised Learning if we consider each example a bag with only one instance [6]. Figure 3.1 displays an example of a bag labeling process.

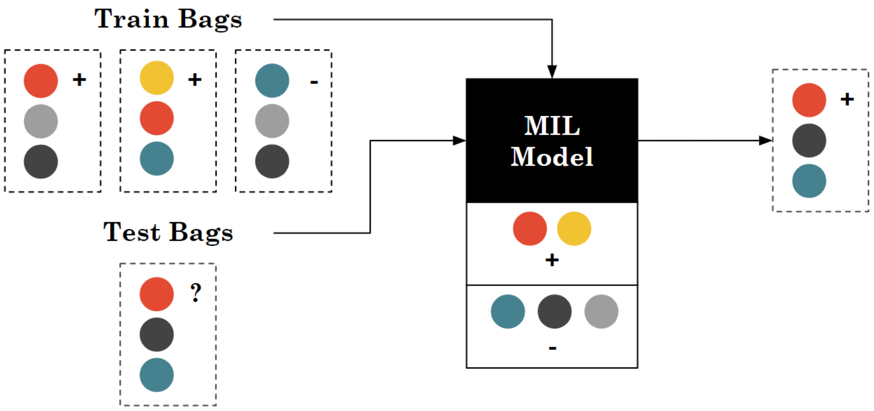


Figure 3.1: Illustration of a test bag classification employing the knowledge obtained by the train bags. The orange and yellow instances are in positive bags but not in negative bags so they were considered positive. Because the test bag contains at least one instance regarded as positive, this bag was labeled positive.

There are two options to represent an image to give it as a training example to a Traditional Supervised Learning model with the goal to detect the existence of an object that only appears in a small area of the picture. The first option is to describe it as a unique feature vector instance as shown in Figure 3.2 A, however, the uniformization would dilute the information present on the relevant areas. On the other hand, the image can be split in multiple training examples classified with the image label as exhibited in Figure 3.2 B, but this would inject wrong information in the model since we would be saying some parts of the image have the said object when they do not. Luckily, the MIL approach (Figure 3.2 C) can provide detailed information without misleading the model, knowing that not every instance in a positive bag has the object but also knowing that every instance contains relevant information to help solve the problem [29] [60].

The formal definition of MIL is as follows. If χ stands for the instance space and considering a dataset $\{(X_1,y_1),..,(X_i,y_i),..,(X_N,y_N)\}$ with N training bags, where a bag is represented by $X_i = \{x_{i1},..,x_{ij},..,x_{in}\} \subseteq \chi$, its number of instances by n_i, and its label by $y_i \in Y = \{-1,+1\}$, the purpose is to classify new bags employing a learner. $x_{ij} \in \chi$ is an instance with d features and it is represented by $[x_{ij1},..,x_{ijl},..,x_{ijd}]$. X_i is a positive bag, hence $y_i = +1$, if there is an index $g \in \{1,..,n_i\}$ so that x_{ig} is a positive instance even though its value is unknown; otherwise X_i is a negative bag with $y_i = -1$ as its label.

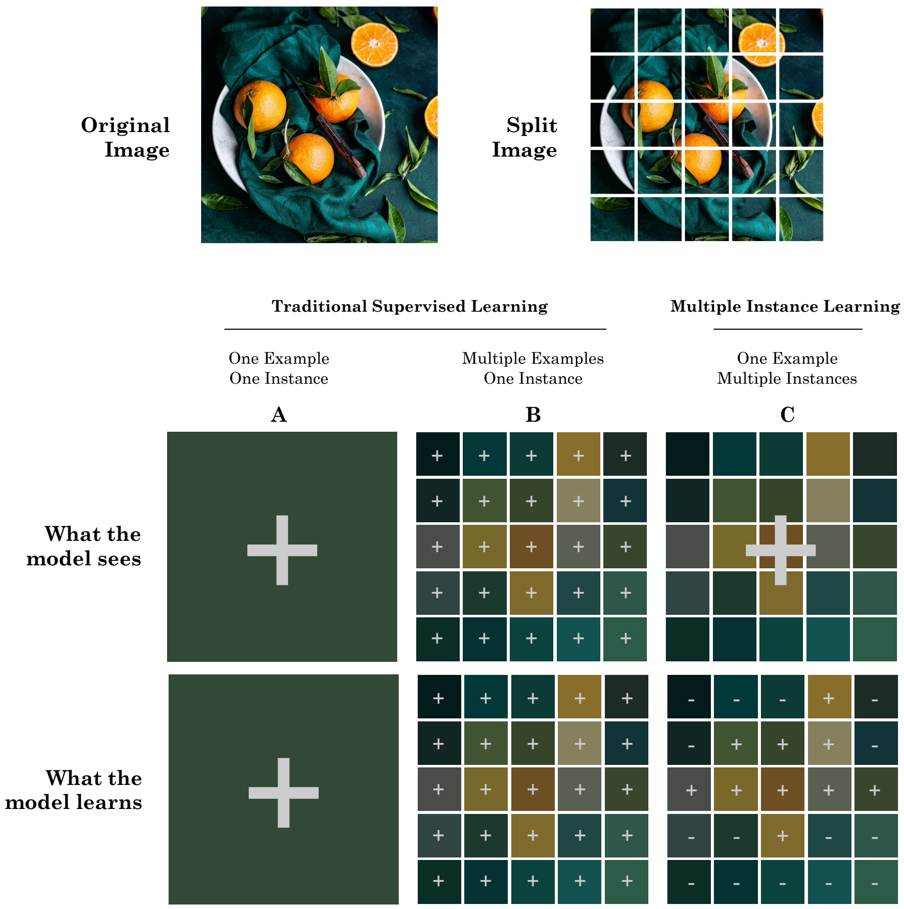

Original Image
Split Image
Traditional Supervised Learning
Multiple Instance Learning
One Example One Instance
Multiple Examples One Instance
One Example Multiple Instances
A
B
C
What the model sees
What the model learns

Figure 3.2: Illustration that presents the contrast in the learning of a model using Traditional Learning and Multiple Instance Learning. Considering an image to detect the orange fruit and that the features of each example are the mean RGB values, there are three options to represent it. Options A and B are Traditional Supervised Learning, while option C is Multiple Instance Learning. **(A) One Example One Instance**: Since a lot of detail was lost the model may not be accurate classifying other examples. **(B) Multiple Examples One Instance**: The model learned incorrectly that some examples have the desired object, which may decrease its future accuracy. **(C) One Example Multiple Instances**: The model knows that the object is present in the set but not in all of its instances, this increases the probability of better accuracy in the future.

3.2 Bags Generators

The Bag Generation step is fundamental in MIL and it can even be regarded as more relevant than the choice of the model for the quality of the final results [58]. The bag generator chooses how to split an image into regions, which are called instances, to constitute a bag [37] [55] [58]. It can also be responsible for the feature extraction function that holds the decision of how to represent the bag instances by a feature vector [55].

There are two types of bags generators: Segmentation and Non-Segmentation as illustrated with examples in Figure 3.3. The difference is that the Segmentation Bag Generators have in consideration the image semantic components, like its objects, to generate the instance, where the Non-Segmentation Bag Generators have a fixed strategy independent of the content of the image in question [55].

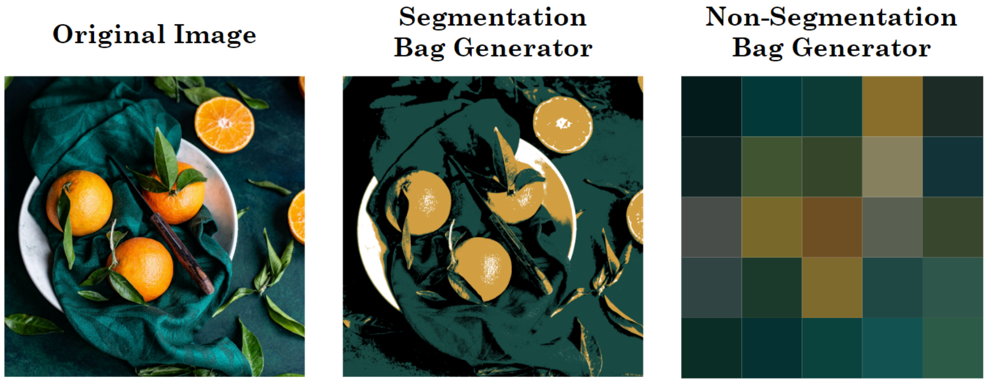


Figure 3.3: Example of a Segmentation Bag Generator and a Non-Segmentation Bag Generator. **Segmentation Bag Generator:** Uses the main colors of the image to segment its objects: the oranges, the white plate, the green fabric and the black background. Other images would have a different number of instances with distinct shapes. **Non-Segmentation Bag Generator:** Splits the image into 5×5 pixels. Other examples would have the same number of instances with the same shape.

If a perfect tool to recognize objects was available, a Segmentation Bag Generator would select the objects in the picture and the model would only have to calculate the union between the objects in the positive bags that are not present in the negative bags. On the other hand, if a model able to deal with billions of instances existed, a Non-Segmentation Bag Generator would only need to generate every possible pixel combination, and the model would certainly have all the needed information [37]. However, neither the perfect object recognition tool nor a model that can handle so much data exists, therefore it is required to conceive different bags generators to find the one that gets better results according to each problem.

Wei and Zhou [55] studied various bags generators and the Single Blob with no Neighbors (SB) proposed by Maron and Ratan [37] was among the ones with better results, even though it was one of the simpler ones. The methodology of this algorithm is illustrated in Figure 3.4. In this Non-Segmentation Bag Generator, the images are first resized and then divided into various not-overlapping sub-images, called blobs, with a 2×2 pixel size. Each instance matches a vector of the RGB values of its pixels: $[R1, G1, B1, R2, G2, B2, R3, G3, B3, R4, G4, B4]$ [37].

3.3 Sampling

In most machine learning problems, the classes are somehow unbalanced. This can bias the models to pick the class with the most examples classifying it with high precision, but ignore the minority

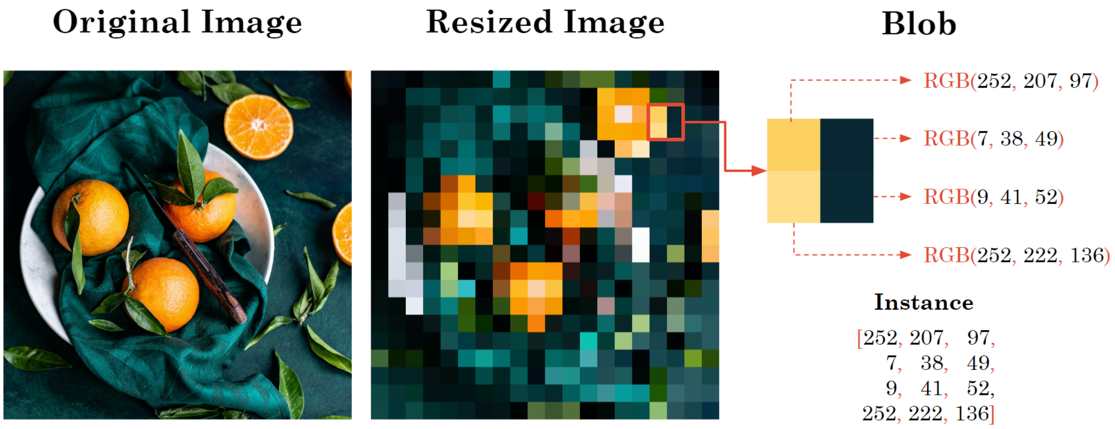


Figure 3.4: Representation of the Single Blob with No Neighbors Bag Generator [37] methodology. The original image is resized (in this example to 20 × 20 pixels) and later split into 2 × 2 pixels blobs. The features of the instance associated with each blob are the RBG values of its pixels.

class, which often is the most important [40]. To deal with unbalanced data there are two techniques: oversampling to increase the number of examples of the minority class and undersampling to decrease the number of samples of the majority class.

In Traditional Supervised Learning, random examples of the majority class can be deleted for undersampling. As an oversampling strategy, arbitrary instances of the minority class can be replicated, however, this strategy can increase the ambiguity in the dataset and result in overfitting. As an alternative solution to the duplication, Chawla et al. [13] developed the Synthetic Minority Oversampling Technique (SMOTE). This strategy randomly selects one example of the minority class, A, and one random k-nearest neighbor of that chosen example, B, creating new instances in the line that connects the two elected examples, A and B [13]. However, the SMOTE technique is hard to execute in MIL problems since it is intended to operate with individual examples and not sets of instances. There are some oversampling implementations designed to work with MIL, however, the majority does not consider if the instances are positives or negative and that can diminish the quality of the used data [40].

Mera et al. [40] proposed an oversampling technique that acknowledges the concept of a positive bag having at least one positive instance. The suggested method tries to model the negative instances in the negative bags with Kernel Density Estimation (KDE) to help find the most positive instances in the positive bags as shown in Figure 3.5. To generate the positive instance of the new bag, this algorithm picks the most positive instance in two random positive bags, x^+_{1j} and x^+_{2j}, and calculates a new instance employing the expression $x^+_{1j} + (x^+_{2j} - x^+_{1j}) \times \alpha$ in which α is a random value between 0 and 1. On the other hand, to create negative examples for the new set, different negative samples are selected and each one is added to the most negative instance from an elected bag. This addition is computed using the previous mathematical statement. The pseudocode of this method is displayed in Algorithm 1.

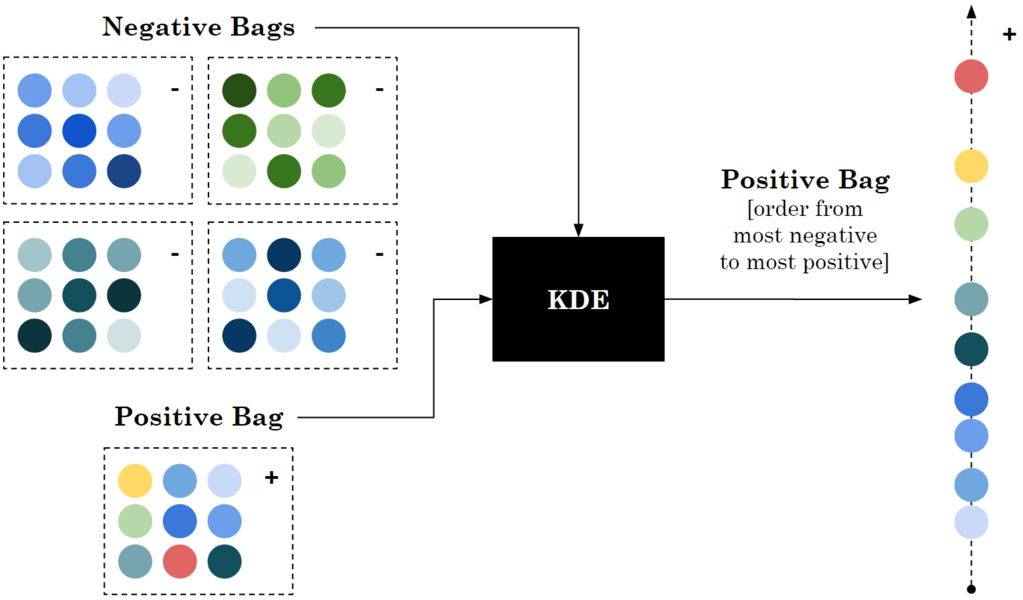


Figure 3.5: Representation of the KDE purpose in the selection of the most positive instance in a bag. This method maps the negative bags to learn what makes an instance negative. When a bag is given as an input, the system gives a score, ranking its instances from the most negative to the least negative, in other words, it sets on top the most positive instance.

Algorithm 1 Pseudocode of the Mera et al. [40] algorithm to oversample bags. Adapted from [40].

```
P ← average instances per bags
N ← number of synthetic bags to be generated

for i = 1 to N do
    $B^i_{new}$ ← Create a new empty bag with positive label

    //Generate a new positive instance for $B^i_{new}$ :
    $(B^+_1, B^+_2)$ ← Select 2 positive bags from the training data set at random
    $x^+_{1j}$ ← Select the most positive instance in $B^+_1$using KDE
    $x^+_{2j}$ ← Select the most positive instance in $B^+_2$using KDE
    $x_{i1} \leftarrow x^+_{1j} + (x^+_{2j} - x^+_{1j}) * \alpha$, where $\alpha \in [0,1]$ at random
    $append(x_{i1}, B^i_{new})$

    //Generate a new negative instance for $B^i_{new}$ :
    $x^-_{1j}$ ← Select the most negative instance in $B^+_1$using KDE
    for j = 2 to P do
        $x^-_{2j}$ ← Get a negative instance from $B^+_2$ at random
        $x_{ij} \leftarrow x^-_{1j} + (x^-_{2j} - x^-_{1j}) * \alpha$, where $\alpha \in [0,1]$ at random
        $append(x_{ij}, B^i_{new})$
    end for
end for
```

3.4 Models

Three sorts of models can solve MIL problems: Instance-Based Classifiers, Bag-Based Classifiers, and Hybrid Classifiers [19]. The first usually has in consideration the fundamental assumption that one positive bag contains at least one positive instance and employs this premise along with the bag label to classify its instances independently. After labeling the instances individually [12], it is simple to classify the bag based on the underlying MIL assumption. The Bag-Based Classifiers assume that the bags with the same label are similar and try to represent them with a single feature vector to find those similarities [15]. These approaches label the bags without classifying their instances first [12]. Lastly, Hybrid Techniques own traits from the other two varieties [19].

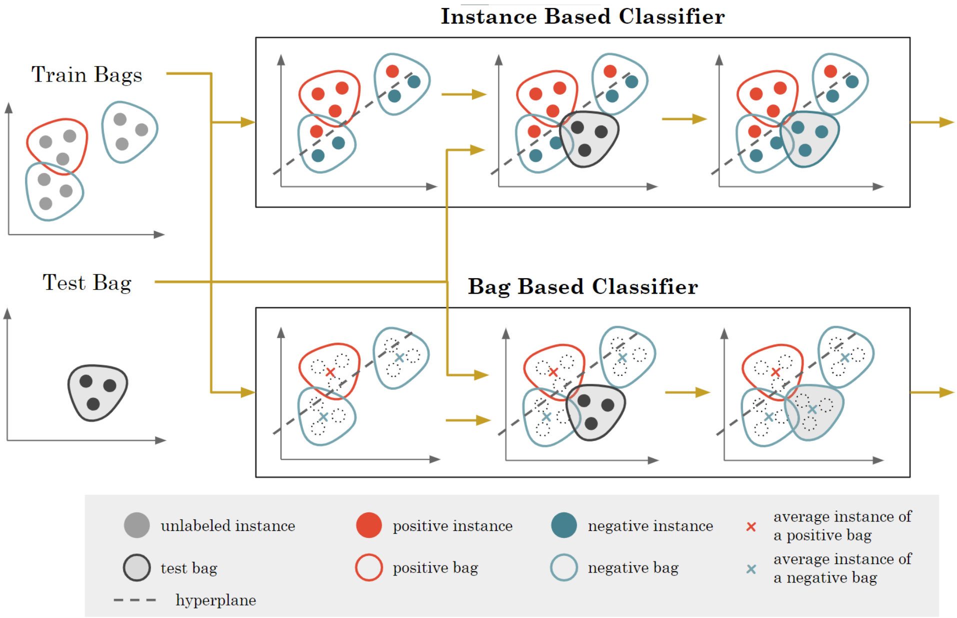


Figure 3.6: Classification example of an Instance-Based Classifier and a Bag Based-Classifier. In this example, both models are based on an SVM and therefore use a hyperplane to split the positive subjects from the negative ones. While the Instance-Based Classifier forms a hyperplane in order to have at least one positive instance in the positively labeled bags, the Bag-Based Classifier tries to design a hyperplane that splits the bags based on their instances average.

It was found two packages that implement some MIL models in Python (the chosen language to develop this work): miGraphPy [60] and MISVM (Multiple-Instance Support Vector Machines) [19]. Afterwards, a brief description of these packages implemented models are presented. MiGraph is implemented in the miGraphPy package and the rest of the models are implemented in the MISVM package. After redesigning the MIL dataset in its own way, every listed approach employs a traditional SVM model to solve the problem.

3.4.1 Instance-Based Classifiers

Single-Instance Learning (SIL) A naive approach that models a MIL problem in a Traditional Supervised Learning problem, giving the bag's label to all of its examples. Every positive instance will be accurately labeled, however, some negative instances will be incorrectly marked as positive as shown in Figure 3.2 B [10] [19].

mi-SVM A technique that tries to find a hyperplane that has at least one instance from every positive bag on one side and the remaining examples on the other side [5] [60].

MI-SVM Picks one instance called "witness" that works as a representative of every positive bag. Firstly, it chooses the example that is the average of every instance in that bag and tries to draw a hyperplane in that space. If there is a witness that is on the wrong side of the hyperplane, it is replaced by the instance from that bag that is placed the furthest from the hyperplane [5] [10] [19] [36].

Multiple Instance Classification Algorithm (MICA) Instead of picking a single witness, this strategy selects a random convex set of instances to work as witnesses in a positive bag. This group of instances does not necessarily have to be the furthest from the hyperplane [19] [36].

Multiple Instance Learning by Semi-Supervised SVM (MissSVM) Addresses the problem as a Semi-Supervised Learning Problem, classifying the instances in the negative bags and leaving unknown the labels of the positive bags instances. Subsequently, it applies a constraint that forces the existence of a positive instance in a positive bag. The optimization problem is then solved with the MI-SVM technique [19] [59] [60].

3.4.2 Bag-Based Classifiers

Normalized Set Kernel (NSK) Maps each bag to a feature vector, summing every instance in that bag and normalizing the result [10] [19] [23].

Statistic Set Kernel (STK) Converts each bag to a feature vector, in which each feature holds two variables: the maximum and the minimum value of that feature in every instance contained in the bag [10] [23].

3.4.3 Hybrid Classifiers

miGraph This method processes each bag as a whole but regards the similarities between its instances. Firstly, this approach creates a matrix for every bag that holds the distances among every pair of instances. This distance represents how comparable the two are. Then, it builds a graph that has an edge that connects two instances if their distance is less than a specific threshold. In the end, the graphs are compared to examine the similarities amongst bags [29] [60].

Sparse Multiple Instance Learning (sMIL) This technique uses a constraint to balance the classification considering that a smaller positive bag holds more information that a bigger positive bag [10] [19].

Sparse Transductive Multiple Instance Learning (stMIL) This method includes all the sMIL constraints but forces the instances of the positive bags to be outside the decision boundary [10] [19].

Sparse Balanced Multiple Instance Learning (sbMIL) If the density of positive instance, n, is known, the problem could be solved more efficiently. This parameter can be passed as an input to the model or calculated using a part of the dataset. This approach finds a solution with sMIL and the n instances of the positive bags that had the best score are classified as positive. This technique already includes stMIL constraints [10] [19].

3.5 Summary

This chapter aimed to describe the methodology of Multiple Instance Learning. The importance of the phase Bags Generation was explained, revealing in detail a Bag Generator in the literature. Furthermore, an algorithm that applies oversampling to MIL was exposed. Lastly, it was listed some models belonging to the three types of classifiers in this framework: Instance-Based Classifiers, Bag-Based Classifiers and Hybrid Classifiers.

Chapter 4

Literature Review

For this work, two areas in the literature were studied: lung structures characterization and diseases detection employing Multiple Instance Learning and predictive models for gene mutation status. The most relevant studies of each of these areas are described in Section 4.1 and Section 4.2 respectively. Section 4.3 makes an overview of these two areas, connecting them.

4.1 Lung Structures Characterization and Diseases Detection employing Multiple Instance Learning

As of yet, ten studies were found that employ Multiple Instance Learning to detect pathologies or characterize the lung in terms of image patterns. The studies were searched employing the query *("Multiple Instance Learning") AND ("Lung")* in the research databases IEEE Xplore [1] and PubMed [2]. The studies that were only focused on the lung cancer nodule were discarded, remaining ten publications that detected Emphysema, Tuberculosis, COPD (Chronic Obstructive Pulmonary Diseases), Pulmonary Embolism, or various lung patterns from 2007 to 2018. Table 4.1, at the end of this section, presents an overview of each paper in this area.

Orting et al. [43] studied MIL to predict the presence of Emphysema at the scan level and at the six lung regions level. Those regions were the Left Lower, Left Middle, Left Upper, Right Lower, Right Middle, and Right Upper region. A CT scan represents a bag, and its instances are 100 small possibly overlapping patches selected randomly. These instances are characterized by a feature vector of filter responses and by the indication of the region that includes it. A region is taken as positive if an expert found any trace of Emphysema in it, while a scan is considered positive if any of its lung regions is positive. Both scan level and region level detections had good results, having the upper right region the best results with an AUC (Area Under the ROC Curve) of 0.89. The best result at the scan level was an AUC of 0.82.

[1] https://ieeexplore.ieee.org/, last accessed on 02/06/20

[2] https://pubmed.ncbi.nlm.nih.gov/, last accessed on 02/06/20

Ramos et al. [49] proposed a method that employed MIL to detect distinct CT patterns in which labels were keywords extracted from the radiology report associated with the CT scan. The considered keywords were: Ground Glass, Honeycombing, Crazy Paving and Hyperlucency, Consolidation and Nodular Patterns. The features used to describe the patches were the mean Hounsfield Unit value, and the model used was based on the Evolution Expectation Maximization Diverse Density (EM-DD). The best result was related to detecting Honeycombing with an AUC of 0.78 but the results for Ground Glass, Crazy Paving, and Hyperlucency did not appear much worse.

Gang et al. [22] proposed a technique to annotate CT slices using MIL. Firstly, this method segmented both lungs, representing each one an instance described by statistics features and Gabor wavelet features. The pretended labels were Spot Density Increased, Plaque Increased in Density, Cavity Empty, Nodular Masses, and were detected employing the Diverse Density MIL model considering retrieval relevance feedback. Detecting Nodular Masses had the best precision of 0.7.

Melendez et al. [39] built a new MIL model intending to detect tuberculosis. This new model was based on the famous miSVM and was designed to improve the estimation of positive instance and decrease the cost of each iteration. The scan CXRs (Chest X-Rays) was considered the bag and its instances were circular 32 pixels in a grid characterized by the intensity distribution of each patch and its position. The usage of MIL surpassed the supervised methods and the new model outdid miSVM, reaching an AUC of 0.91. Figure 4.1 compares the ground-truth of a scan and the areas classified with tuberculosis by each model tested in this study.

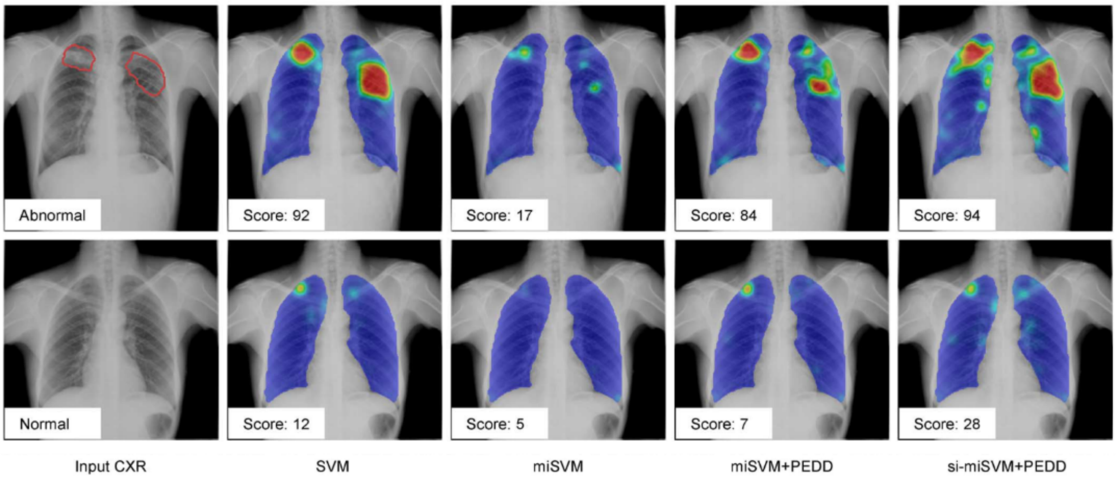


Figure 4.1: Heatmaps representing the results of a tuberculosis detection model (second to fifth columns). The oiginal scans are the first column which have the ground-truth outlined in red. On the heat maps, warm colors indicate abnormality. From Melendez et al. [39].

Pino Peña et al. [46] implemented a methodology that detects Chronic Obstructive Pulmonary Disease (COPD) and quantifies the Emphysema without annotations on the High Resolution Computed Tomography (HRCT). The bags were each scan, the instances were patches randomly selected from the lung area and were defined by the co-occurrence matrices features and Gaussian derivatives features. The authors choose the miSVM and MILES models to train and test. This work achieved an AUC of 1 in the COPD classification that may be justified by the severe stage of the dataset cases. As for the Emphysema identification, the chosen areas by the models were moderately related to the manual annotations. The overview of this paper methodology is shown in Figure 4.2

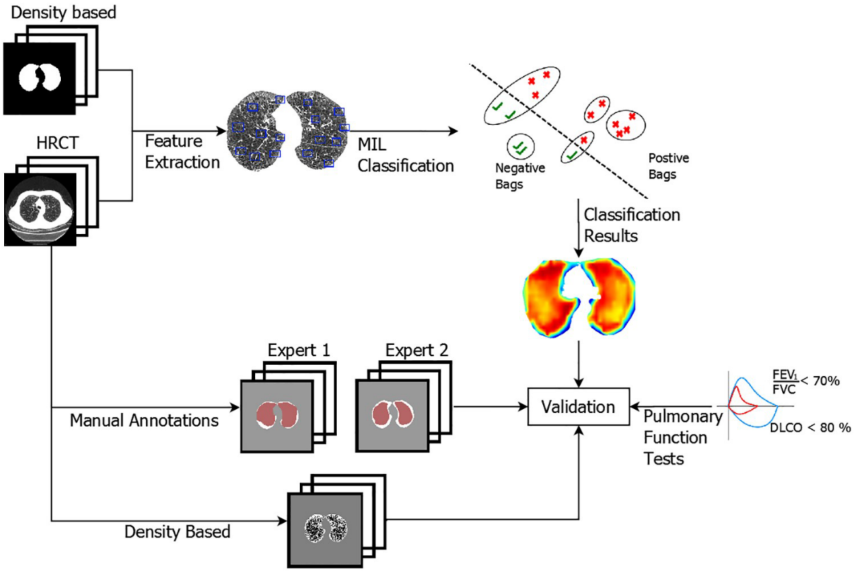


Figure 4.2: Methodology of a system that detects COPD and identifies Emphysema regions. This system uses scans HRCT and extracts features based on density. The results provided by a MIL model are validated with the manual annotations of experts and density based analysis. From Pino Peña et al. [46].

In 2014, Cheplygina et al. [15] studied the presence of COPD considering the distribution in the whole lung. Unlike the previous studies, this work regarded a single CT slice as a bag instead of the entire scan. The 50 randomly selected instances inside the lung were described with the response of 8 filters: Gaussian, gradient magnitude, Laplacian of Gaussian, first, second and third eigenvalue of the Hessian, Gaussian curvature and eigen magnitude. The authors made distinct experiments using the entire dataset or half of it, concluding that employing half of the data does not worsen the results significantly. The best result was achieved by an SVM based on the average of instances with an AUC of 0.742 that increased to 0.776 when using the full dataset. Four years later, Cheplygina et al. [16] studied the classification of COPD using MIL once again, but this time considering multiple datasets with heterogeneous distributions from different scanners, protocols, and centers. Each scan was a bag that contained 50 randomly selected possible overlapping patches that were characterized by Gaussian scale space features. This method assigned weights to the training bags based on their similarities and employed the naive model SimpleMIL achieving an AUC of 0.969.

Bi and Liang [9] developed a method to find areas with suspicion of Pulmonary Embolism in Computed Tomography Angiography images considering geodesic distances between candidate regions. This work aimed to identify at least one region in each positive example and not identify

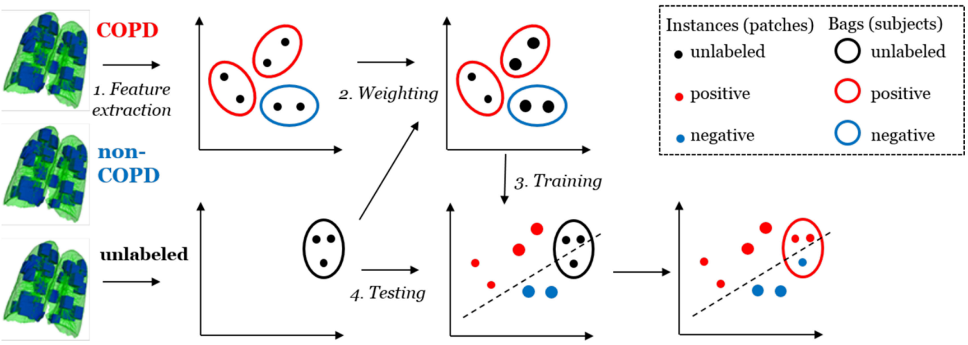


Figure 4.3: Overview of a MIL methodology to detect COPD considering weights. Firstly the features are extracted from the trainset scans and the instances are weighted. The model is trained, establishing a hyperplane for the set. The new bag is placed in the trained model and the instances are classified according to its weights and the previously fixed hyperplane. From Cheplygina et al. [16].

every positive instance. This technique achieved a sensitivity of 81%.

Dundar et al. [20] proposed a Convex Hull representation for MIL, shown in Figure 4.4, to detect Pulmonary Embolism (PE). This study considered positive the instances near the area annotated manually by experts. This work considered only positive bags since the goal was to detect positive instances and not classify the bag. Four models were tested: Fisher's Discriminant (FD), CH-FD (the Convex Hull proposed in this work), EM-DD, and IDAPR. The used features considered the intensity distribution, the neighbors' distribution, and the shape. The developed model was the most efficient since it can delete the majority of MIL combinations and can find the optimal global solution.

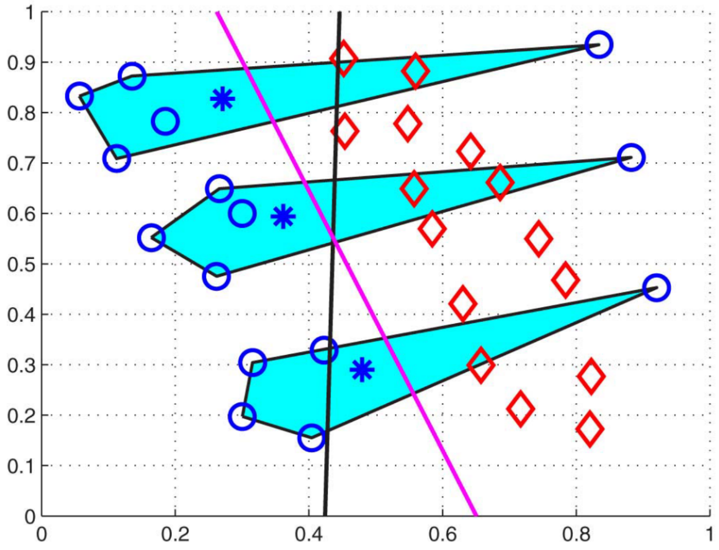


Figure 4.4: Convex Hull representation for Multiple Instance Learning. Circles - Positive classes; Diamonds - Negative classes; Polyhedrons - Convex hulls for three positive bags; Starts - Points to represent each bag. Grey line - SVM hyperplane; Pink line - Hyperplane by the proposed algorithm. From Dundar et al. [20].

Table 4.1: Overview of published studies regarding lung diseases detective models with Multiple Instance Learning.

Reference	Objective	# Exams	Labels	MIL Modelation	Models	Best Result
Orting et al. 2018 [43]	Use a multiple instance learning approach to predict both scan-level and region-level emphsema presence.	1800	Emphysema	**Bag:** The CT Scan **Instances:** 600 possibly overlapping random volumetric patches **Features:** Filter responses: Gaussian blur, gradient magnitude, eigenvalues of the Hessian, Laplacian of Gaussian, Gaussian curvature and the Frobenius norm of the Hessian	Not mentioned	AUC = **0.82**

Ramos et al. 2013 [49]	Use CT scans and keywords in the respective radiology reports to learn patterns.	1110	Hyperlucency (HL);
			Ground Glass (GG);
			Honeycombing (HC);
			Crazy Paving (CP);
			Consolidation (Cons.);
			Nodular Pattern (Nod.).
				Bag: The CT Scan
				Instances: Seed points based on local maxima or minima with a minimum distance of 5 voxels
				Features: Mean Hounsfield Unit
					EM-DD	(AUC =)
						HL - **0.71**;
						GG - **0.72**;
						HC - **0.78**;
						CP - **0.77**;
						Cons. - **0.52**;
						Nod. - **0.32**.

Gang et al. 2013 [22]	Propose a method of medical image semantic annotation based on multi-instance learning.	240	Mottled shadows of high density; Patchy shadows of high density; Cavity and hole; Nodular and masses.	**Bag:** One CT image **Instances:** The left and right lung **Features:** The gray and texture feature	EM-DD	(Precision =) Mottled shadows of high density - **0.625**; Patchy shadows of high density - **0.640**; Cavity and hole - **0.626**; Nodular and masses - **0.700**.

Melendez et al. 2015 [39]	Apply MIL to a CAD system for tuberculosis detection and propose an improved algorithm that overcomes miSVM's drawbacks related to positive instance underestimation and costly iteration.	2636	Tuberculosis	**Bag:** The CXR Scan
				Instances: Circular patches with a radius of 32 pixels on a grid with a spacing of 8 pixels
				Features: Based on the first four moments of the intensity distributions resulting after applying a multiscale local jet of second order
					k-NN;
					SVM;
					miSVM;
					miSVM+PEDD;
					si-miSVM+PEDD.
						(AUC =)
						Database 1:
						SVM - 0.88;
						Database 2:
						si-miSVM+PEDD - **0.86**;
						Database 3:
						si-miSVM+PEDD and si-miSVM - **0.91**.

Pino Peña et al. 2018 [46]	Build a classifiers that outputs a patient label indicating overall COPD diagnosis and local labels indicating the presence of Emphysema.	88	Chronic Obstructive Pulmonary Disease	**Bag:** The HRCT Scan **Instances:** Randomly selected 3D patches from inside the lungs **Features:** Co-occurrence matrices and Gaussian derivative features	miSVM; MILES.	(AUC =) miSVM - **1.0**

Reference	Objective	# Exams	Labels	MIL Modelation	Models	Best Result

Cheplygina et al. 2014 [15]	Investigate various MIL assumptions in the context of COPD.	200	Chronic Obstructive Pulmonary Disease	**Bag:** One CT image
				Instances: 50 ROIs samppled at random locations within the lungs
				Features: Filter responses: Gaussian, gradient magnitude, Laplacian of Gaussian, first, second and third eigenvalue of the Hessian, Gaussian curvature and eigen magnitude
					Simple logistic;
					Simple k-NN;
					miSVM;
					MILBoost;
					Citation k-NN;
					mean-inst SVM;
					extremes SVM;
					BoW SVM;
					MILES;
					meanmin SVM;
					meanmin k-NN;
					emd SVM;
					emd k-NN.
						(AUC =)
						mean-inst SVM - **0.776**

Cheplygina et al. 2018 [16]	Investigate classification of COPD in a multicenter dataset from different centers, different scanners, with heterogenous subject distributions	803	Chronic Obstructive Pulmonary Disease	**Bag:** The CT Scan **Instances:** 50 possibly overlapping volumtric ROIs of size 41*41*41 voxels extracted at random locations inside the lung mask. **Features:** Gaussian scale space features and compute eight filters: smoothed image, gradient magnitude, Laplacian of Gaussian, three eigenvalues of the Hessian, Gaussian curvature and eigen magnitude	SimpleMIL	AUC = **0.969**
Bi and Liang 2007 [9]	Propose a novel classification approach for automatically detecting pulmonary embolism from CTA images.	177	Pulmonary Embolism	**Bag:** Cluster of voxels **Instances:** Cluster of voxels **Features:** Voxel intensity distributions within the candidate, distributions in neighborhood of the candidate the 3D shape of the candidate and enclosing structures	Spatial MIL; SVM.	Sensitivity = **81%**

Dundar et al. 2008 [20]	Propose a framework for learning a Convex Hull representation of multiple instances that is significantly faster than existing MIL algorithms.	72	Pulmonary Embolism	**Bag:** A set of instances
				Instances: Candidates that are spatially close to the radiologist marked ground-truth
				Features: Voxel intensity distributions within the candidate, distributions in neighborhood of the candidate the 3D shape of the candidate and enclosing structures
					Fisher's Discriminnat;
					CH-FD;
					EM-DD;
					IDAPR.
						(AUC =)
						CH-FD - **0.86**

4.2 Predictive Models for Gene Mutation Status

This area can be divided into two subareas: predictive models for gene mutation status based on nodule features, presented in Subsection 4.2.1, and predictive models for gene mutation status based on both nodule features and lung features, explored in Subsection 4.2.2. In total, twenty studies were found after employing the query *("Gene Mutation Status") AND ("Prediction") AND ("Lung Cancer")* in the research databases IEEE Xplore [3] and PubMed [4] and excluding the ones that were not based on CT scans. These studies include semantic, radiomic, and deep learning features, which were the input of statistical, machine learning, or deep learning models. All of these studies are from 2017 to 2019 which shows how novel the investigation of this area is.

4.2.1 Based on Nodule Features

Thus far it was found six studies that would take into account features related to the nodule. Table 4.2 divides the articles per feature extraction and classification method, while Table 4.3, at the end of this subsection, gives an overview of each paper.

Table 4.2: Published studies regarding predictive models for gene mutation status based on nodule features organized by feature extraction and techniques used.

<table>
<tr><th colspan="2"></th><th colspan="5">Classification Methods</th></tr>
<tr><th colspan="2"></th><th>Statistical</th><th colspan="2">Machine Learning</th><th colspan="2">Deep Learning</th></tr>
<tr><th rowspan="3">Feature Extraction</th><th>Semantic</th><td>[61] [14]</td><td colspan="2">-</td><td colspan="2">-</td></tr>
<tr><th>Radiomic</th><td>-</td><td>[30]</td><td colspan="2" rowspan="2">[34]</td><td rowspan="2">[53] [57]</td></tr>
<tr><th>Automatic Feature Learning</th><td>-</td><td>-</td></tr>
</table>

Zou et al. [61] studied the *EGFR* mutation status of stage I/II lung adenocarcinoma in tumors with lesions <3 cm to know the correlation between *EGFR* mutation status, clinical features, and CT characteristics. To identify independent risk factors, it was used multiple logistic regression analyses. Zou et al. concluded that *EGFR* mutation appeared more frequently in women, never-smokers, and patients with a carcinoembryonic antigen level <2.6 ng/ml. However, papillary predominant adenocarcinomas, intermediate/low pathologic grade tumors, tumors in the upper lobe, and showing ground-glass opacity (GGO) or mixed GGO also had some level of correlation with *EGFR*. When it comes to independent risk factors, the multivariable analyses chose GGO, acinar or papillary predominant adenocarcinoma, and non-smoker.

Cheng et al.[14] examined the correlation between CT morphological features and the presence of *EGFR* mutations in NSCLC. The features obtained from the CT were ground-glass opacity (GGO) content, tumor size, cavitation, air-bronchogram, lobulation, and spiculation. Weighted mean difference (WMD) or inverse variance (IV) in the form of odds ratio (OR) was used to determine the association between the CT features and *EGFR* mutation. This study concluded that this gene mutation tended to exist in tumors with part-solid GGO as opposed to nonsolid GGO.

[3] `https://ieeexplore.ieee.org/`, last accessed on 02/06/20

[4] `https://pubmed.ncbi.nlm.nih.gov/`, last accessed on 02/06/20

Yet, features such as tumor size, cavitation, air-bronchogram, lobulation and spiculation weren't independently associated with *EGFR* mutations.

Koyasu et al. [30] aimed to develop two types of classifiers: one for predicting lung cancer histological subtype (adenocarcinoma vs. squamous cell carcinoma), and the second for predicting *EGFR* mutation status in adenocarcinoma (mutant vs. wild-type). It was used two machine learning algorithms: Random forest (RF) and XGBoost (XGB), a particular implementation of Gradient Tree Boosting. An overview of the model is presented in Figure 4.5.

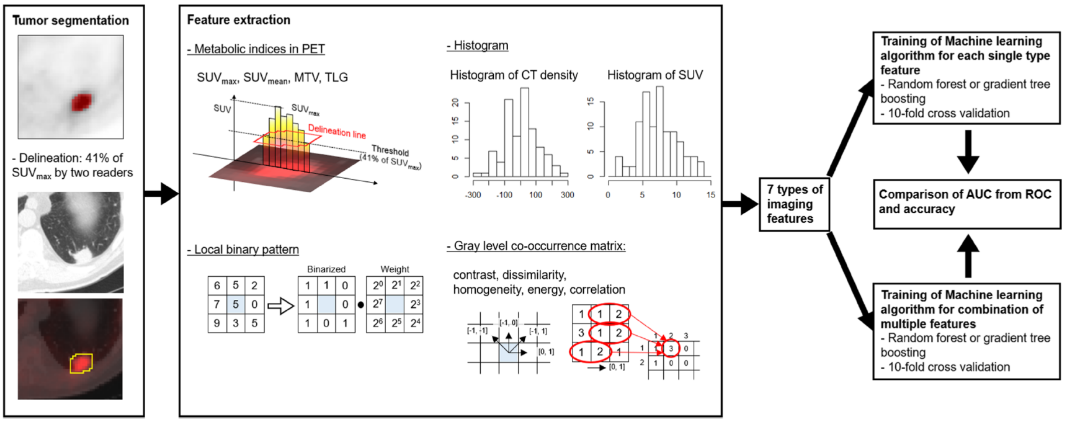

Tumor segmentation
- Delineation: 41% of SUVmax by two readers
Feature extraction
- Metabolic indices in PET
SUVmax, SUVmean, MTV, TLG
SUV
SUVmax
Delineation line
Threshold (41% of SUVmax)
- Histogram
Histogram of CT density
Histogram of SUV
0 5 10 15 20
-300 -100 100 300
0 5 10 15
0 5 10 15
- Local binary pattern
6 5 2
7 5 0
9 3 5
Binarized
1 1 0
1 0
1 0 1
Weight
2^0 2^1 2^2
2^7 2^3
2^6 2^5 2^4
- Gray level co-occurrence matrix:
contrast, dissimilarity, homogeneity, energy, correlation
[-1, -1]
[-1, 0]
[-1, 1]
[0, 1]
1 1 2
3 1 2
1 2 1
1 2 3
1 1 3 0
2 1 0 0
7 types of imaging features
Training of Machine learning algorithm for each single type feature
- Random forest or gradient tree boosting
- 10-fold cross validation
Comparison of AUC from ROC and accuracy
Training of Machine learning algorithm for combination of multiple features
- Random forest or gradient tree boosting
- 10-fold cross validation

Figure 4.5: Model that uses radiomic features as predict lung cancer histological subtype and to predict EGFR mutation status. From Koyasu et al. [30].

In Koyasu et al. study, XGB performed better than RF in both classification problems, having the AUC of 0.843 and 0.659 for histological subtype classification and *EGFR* classification respectively. Concerning the classification of *EGFR* mutation status using multiple types of imaging features, for RF, Bayesian optimization selected GLCM of CT and the histogram of PET and, for XGB, Bayesian optimization adopted Metabolic indices in PET, the histogram of PET, and GLCM of PET as the optimal combination of imaging features.

Li et al. [34] intended to investigate the capacity to detect *EGFR* mutations on CT images with lung adenocarcinoma applying radiomics and multi-level residual convolutionary neural networks. The study analyzed the predictive ability of both models in the same sample and investigated the feasibility of their combination as seen in Figure 4.6. The model that takes into account radiomics, CNNs and clinical information showed the highest AUC value of 0.834. The one regarding only CNNs was better than the radiomics model and did not express significant lack comparing with the model that considers radiomic, CNNs and clinical features or the model that considers radiomic and CNNs. The inclusion of clinical features did not increase the AUC of any of the other models.

Wang et al. [53] introduces an end-to-end deep learning pipeline to predict *EGFR* mutation status in lung adenocarcinoma using CT. This method only demands the manually selected tumor region in a CT image without precise tumor boundary segmentation or human-defined features. The deep learning model performance achieved an AUC of 0.85 in the primary cohort and 0.81 in the independent validation cohort, demonstrating notable variations in *EGFR*-mutant and *EGFR*-wild type tumors, and it was able to find suspicious areas inside tumors. Even though it only

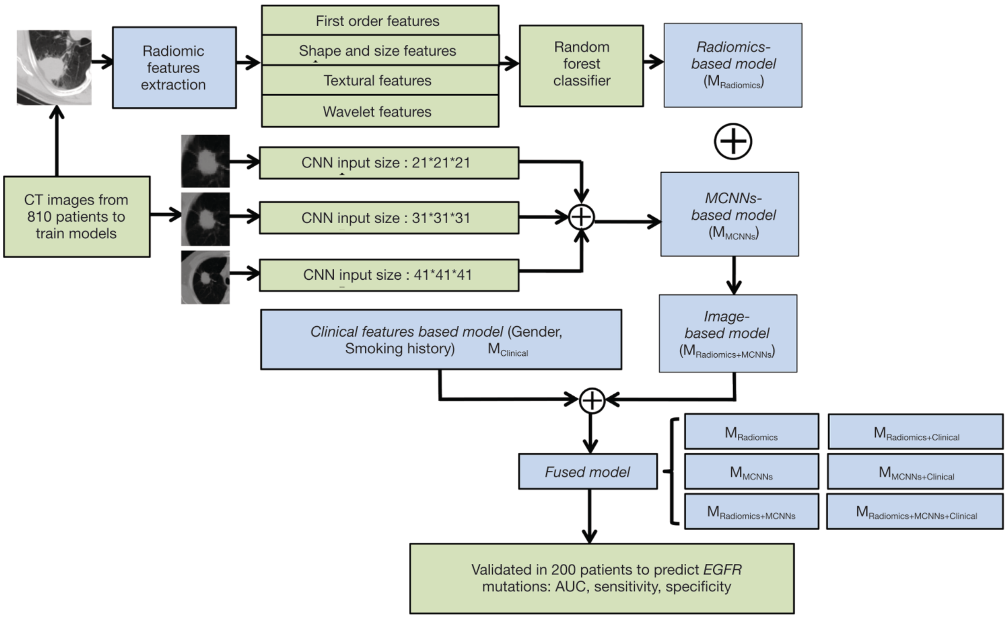

Radiomic features extraction
First order features
Shape and size features
Textural features
Wavelet features
Random forest classifier
Radiomics-based model ($M_{Radiomics}$)
CT images from 810 patients to train models
CNN input size : 21*21*21
CNN input size : 31*31*31
CNN input size : 41*41*41
MCNNs-based model (M_{MCNNs})
Clinical features based model (Gender, Smoking history) $M_{Clinical}$
Image-based model ($M_{Radiomics+MCNNs}$)
Fused model
$M_{Radiomics}$
$M_{Radiomics+Clinical}$
M_{MCNNs}
$M_{MCNNs+Clinical}$
$M_{Radiomics+MCNNs}$
$M_{Radiomics+MCNNs+Clinical}$
Validated in 200 patients to predict EGFR mutations: AUC, sensitivity, specificity

Figure 4.6: Framework with different combinations of a radiomic model, CNNs and clinical information to detect *EGFR* mutations on CT images. From Li et al. [34].

studied adenocarcinoma, the model shows good predictive value in other histological types with an AUC of 0.77.

Zhao et al. [57] produced a deep learning model based on 3D convolutional neural networks (CNNs) to automatically predict *EGFR*-mutant pulmonary adenocarcinoma in CT images. This method integrated modern advancements in deep supervised learning, such as dense connection and mixup training to decrease the chances of overfitting. This method was compatible with approximate locations of the nodules. An overview of the model is presented in Figure 4.7.

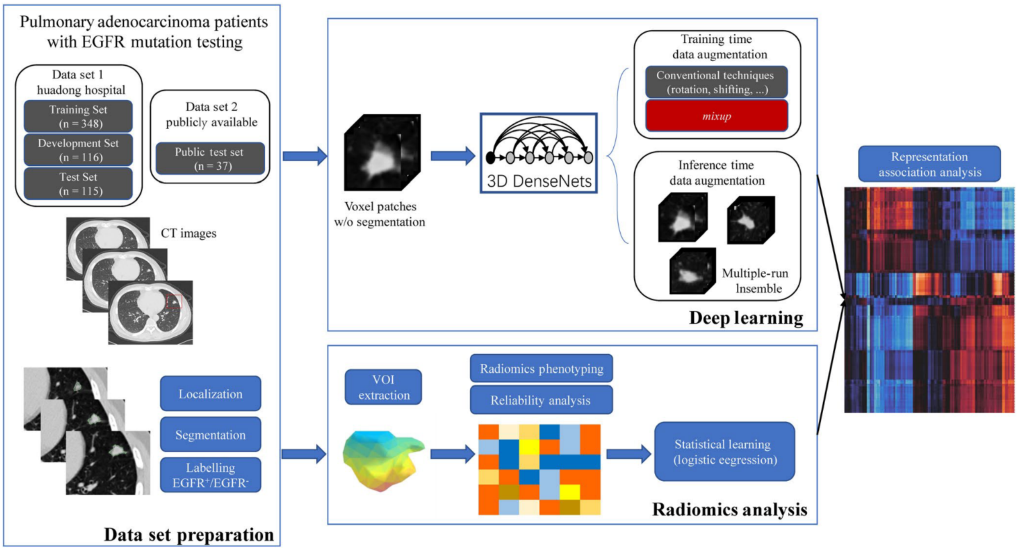

Pulmonary adenocarcinoma patients with EGFR mutation testing
Data set 1 huadong hospital
Training Set (n = 348)
Development Set (n = 116)
Test Set (n = 115)
Data set 2 publicly available
Public test set (n = 37)
CT images
Localization
Segmentation
Labelling EGFR⁺/EGFR⁻
Data set preparation
Voxel patches w/o segmentation
3D DenseNets
Training time data augmentation
Conventional techniques (rotation, shifting, ...)
mixup
Inference time data augmentation
Multiple-run lnsemble
Deep learning
VOI extraction
Radiomics phenotyping
Reliability analysis
Statistical learning (logistic eegression)
Radiomics analysis
Representation association analysis

Figure 4.7: Overview of a deep learning model based on 3D convolutional neural networks to automatically predict *EGFR* mutation using CT images. From Zhao et al. [57].

The model predicted *EGFR* mutation status with AUCs of 0.758 and 0.750 for the holdout test set and public test set, respectively. It was found strong correlations between features extracted by deep learning and radiomics features. However, some deep learning features were not associated with any radiomic features, suggesting additional information obtained by the 3D DenseNets.

Table 4.3: Overview of published studies regarding predictive models for gene mutation status based on nodule features.

Reference	Objectives	Methods	Results	#Patients	Relevant Features
Zou et al. 2017 [61]	Identify the relationship between *EGFR* mutation status, clinical features, and CT characteristics	Multivariable Analyses	AUC = **0.737**	171	**Clinical**: gender, smoking history, carcinoembryonic antigen level, pathologic grade **Nodule**: lobe, ground glass component
Cheng et al. 2017 [14]	Investigate the relationship between CT features and *EGFR* mutations	Weighted Mean Difference, Inverse Variance	OR = **0.49**	1097	**Nodule**: ground glass component
Koyasu et al. 2019 [30]	Develop radiomics approach for classifying histological subtypes and *EGFR* mutation status	XGBoost and Random Forest	AUC = **0.843** AUC = **0.659**	138	**Radiomic**: GLCM of CT, the histogram of PET Metabolic indices in PET, GLCM of PET
Li et al. 2018 [34]	Analyze the ability to detect *EGFR* mutations on chest CT images	Random Forest and CNNs	AUC = **0.834**	1010	Not mentioned
Wang et al. 2019 [53]	Develop an end-to-end pipeline that requires only the manually selected tumour region in a CT image	CNNs	AUC = **0.85**	844	Not mentioned
Zhao et al. 2019 [57]	Predict *EGFR* mutation in CT images	3D DenseNets	AUCs = **0.758**	879	Not mentioned

4.2.2 Based on Nodule Features and Lung Structures and Diseases

Thus far it was found four studies that would take into account at least one feature related to structures or disease external to the nodule. Table 4.4 divides the articles per feature extraction and classification method, while Table 4.5, at the end of this subsection, gives an overview of each paper.

Table 4.4: Published studies regarding predictive models for gene mutation status based on nodule features and lung structures and diseases organized by feature extraction and techniques used.

		Classification Methods			
		Statistical		**Machine Learning**	
Feature Extraction	**Semantic**	[11]	[51]	[45]	[24]
	Radiomic		-		-

Cao et al. [11] aimed to dissect the disparities in CT features between subjects who have *EGFR* mutations and those who have wild-type *EGFR* and develop a prediction tool based on principal component analysis. Accompanying with gender, smoking history, and GGO, adenocarcinomas with *EGFR* mutation were significantly associated with emphysema, TDR, and the diameter in the mediastinal window. The sensitivity and specificity for predicting exon 19 deletion mutation were 59.09 and 76.79%, respectively and the prediction score is calculated by:

$$\begin{aligned} 0.305 gender + 0.254 smokinghistory + 0.198 MaxDmediastinal + TDR 0.254 \\ + 0.280 GGO + 0.095 emphysema \end{aligned} \tag{4.1}$$

The sensitivity for predicting exon 21 missense mutation was 72.34, the specificity was 78.57%, and the prediction score can be determined by:

$$\begin{aligned} 0.354 gender + 0.291 smokinghistory + 0.410 MaxDmediastinal \\ + 0.408 MinDmediastinal \end{aligned} \tag{4.2}$$

Rizzo et al. [51] aimed to confirm the beforehand produced models for the prediction of *EGFR* and *KRAS* mutations with univariate analysis to study the connections of the studied features. This study proved a connection between *EGFR* mutation and internal air bronchogram, pleural retraction, emphysema, and lack of smoking with an AUC of 0.82 and an association between *KRAS* mutation and round shape, emphysema, and smoking with an AUC of 0.60. However, even though several features were related to each of the gene mutations, the AUC for the models considering only smoking was identical to that of the complete model for both genes.

Pinheiro et al. [45] aims to examine and discuss the connections between imaging phenotypes and lung cancer-related mutation status. For that, this study conducted high-dimensional data visualization and developed classifiers using gradient tree boosting, which help analyze the outcomes for *EGFR* and *KRAS* according to diverse combinations of input features. Radiomic and semantic features were regarded as the main types of input features. The semantic were divided into features that only describe the nodule, features that only describe structures external to the nodule and a hybrid between the previous two. The main conclusion of this study was that the separation of classes between mutated and wild type *EGFR* gene status is better when using hybrid semantic features, obtaining the best classification result with AUC of 0.746. This implies that the best way to address this problem is by mixing nodule-related features with features from other lung structures. Unfortunately, for *KRAS* there is no visible separation between classes with any type of input features.

Gevaert et al. [24] studied whether *EGFR* and *KRAS* mutation status are predictable employing semantic imaging data annotated by thoracic radiologists, developing the following decision tree for the aforementioned prediction.

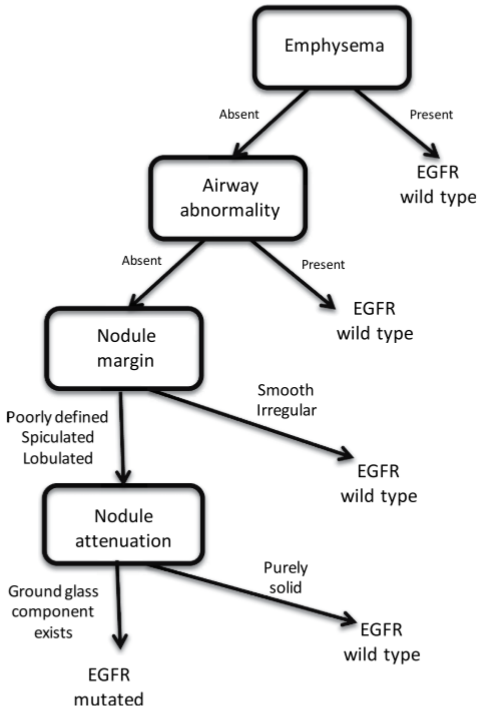


Figure 4.8: Decision tree to predict *EGFR* and *KRAS* mutation status using only semantic annotations. From Gevaert et al.[24].

This decision tree employed four features: emphysema, airway abnormality, the percentage of ground glass component and the type of tumor margin as show in Figure 4.8. The wild type status for *EGFR* is predicted by the appearance of either of the first two variables while the presence of any ground glass component indicates *EGFR* mutations. The AUC for predicting *EGFR* mutation status was 0.89, not improving when merging clinical data with the semantic image features. Like previous studies, *KRAS* mutation status was not connected with semantic image features. This study emphasizes the relevance of the lesion's appearance and its environment.

Table 4.5: Overview of published studies regarding predictive models for gene mutation status based on nodule features and lung structures and diseases.

Reference	Objectives	Methods	Results	#Patients	Relevant Features
Cao et al. 2018 [11]	Identify CT features that correlate with *EGFR* mutation status	Principal Component Analysis	Sensitivity = **72.34** Specificity = **78.57%**	156	**Clinical**: gender, smoking history **Radiomic**: diameter in the mediastinal window, tumor shadow disappearance rate **Nodule**: ground glass component **Structures and Diseases**: emphysema
Rizzo et al. 2019 [51]	Validate associations between radiological features and clinical features with *EGFR*/*KRAS* alterations.	Univariate Analysis	AUC = **0.82**	122	**Clinical**: smoking history **Nodule**: pleural retraction, shape, internal air bronchogram **Structures and Diseases**: emphysema
Pinheiro et al. 2019 [45]	Analyse the results for *EGFR* and *KRAS* biological markers according to different combinations of input features	Gradient Tree Boosting	AUC = **0.746**	211	**Clinical**: smoking history, gender **Nodule**: periphery, attenuation, air bronchogram, shape **Structures and Diseases**: emphysema, lung parencyma features
Gevaert et al. 2017 [24]	Investigated whether *EGFR* and *KRAS* mutation status can be predicted using imaging data.	Decision Tree	AUC = **0.89**	186	**Nodule**: ground glass component, tumor margin **Structures and Diseases**: emphysema, airway abnormality

4.3 Discussion

The best result for each lung characterization label studied in the literature (excluding the studies that do not reveal the AUC result) is presented in Figure 4.9. The label with the best results is Chronic Obstructive Pulmonary Disease. This may happen because this label covers multiple pathologies, which not only has more data available than the more specific diseases but also does not require the model to learn the patterns in so much detail since it does not need to differentiate pathologies. It also should be kept in mind that some of these works with exceptionally good results do not provide the datasets used, so the study cannot be reproduced and verified.

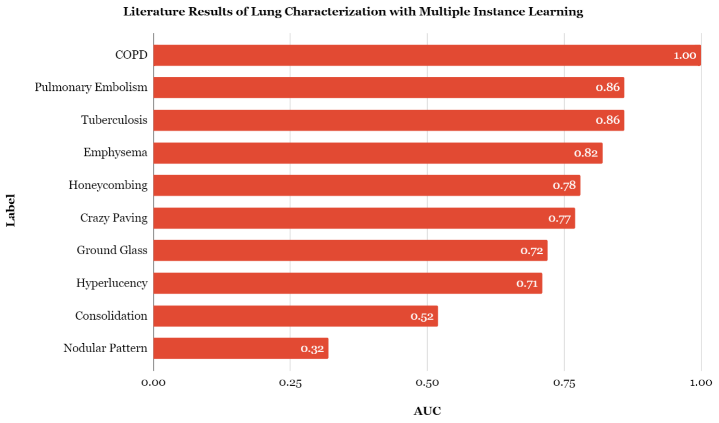


Figure 4.9: The best result for each lung characterization label studied in the literature. Only the studies that revealed the AUC were considered in this chart.

From the studies in Section 4.2 and the chart in Figure 4.10 it can be inferred that the most frequent and relevant features to predict *EGFR* are smoking history, gender, presence of emphysema, ground glass component and air bronchogram. The hardship to accurately predict *KRAS* mutation status can also be deduced from the literature.

A big part of the studies use radiomic and automatically learned features in their models. Those features cannot be interpreted by experts and therefore the results of the model cannot be confirmed by the human eye.

The study of the features external to the nodule is not well developed both its detection and its utilization in predicting gene mutation status. However, some studies mention the possibility of increasing the predicting results if these types of features are taking into account.

Considering the Figure 4.9 and Figure 4.10, one can conclude that the lung labels that are being detected with MIL are not the features used to predict *EGFR*. In the future, one should study the association between these two methodologies trying to detect with MIL the features proven relevant for the *EGFR* prediction or studying the influence of the labels that are being detected with MIL in predicting the gene.

Lastly, the systems presented in the literature whether are features detection models or mutation status prediction models, there is not an example that merges both.

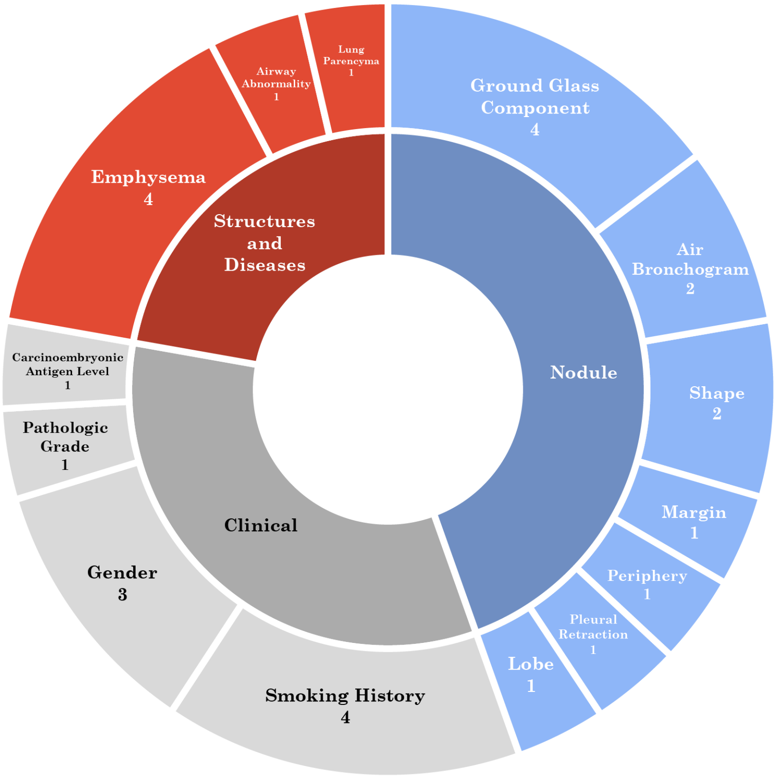


Figure 4.10: Most frequent and relevant features in the literature and the correspondent number of papers.

In conclusion, it is missing an end-to-end pipeline understandable by experts and that takes into consideration features of the whole lung to make the process more automatically and make the specialists' jobs easier.

4.4 Summary

The literature can be divided into lung structures characterization using Multiple Instance Learning and diseases detection plus predictive models for gene mutation. The latter can be based on nodule features or based on both nodule features and lung structures and diseases. Thus far, it wasn't found any study on a predictive model for gene mutations based solely on lung structures and diseases since this type of feature is not very well studied.

None of the studies is an end-to-end solution that from CT images can detect features both internal and external to the nodule which are seen by the human eye and considering those features predicts the gene mutation status.

Chapter 5

Data Description and Preparation

Interpreting the content of the used datasets and finding the areas they may differ is essential to understand the quality of the data that will be, posteriorly, given to the algorithm. This examination is performed in Section 5.1 by describing each one of the three datasets used in this work.

Due to the unbalance of some annotations of the distinct datasets, those had to be omitted from this experience. The exclusion criteria and the dismissed annotations are revealed in Subsection 5.2.1.

Even amongst the same database, the CT spacing and units can fluctuate depending on the equipment performing the CT exam. Hence, it is crucial to pre-process every scan in order to normalize the information and to be better understood by the used model. The pre-processing of the scans is disclosed in Subsection 5.2.2. Another core part before the development of the detection models is to obtain the lung segmentation as explained in Subsection 5.2.3.

5.1 Data Description

Three databases were employed in this project: one public, one from a formal collaboration with Portuguese CHSJ (Lung Cancer Screening - A non-invasive methodology for early diagnosis, which is a current FCT funded project), and one private. The first two datasets were not used in totality since some cases did not have semantic annotations related from experts. An overview of the three datasets is presented in Table 5.1 and subsequently described.

5.1.1 S. João Hospital Database

For this project, a collaboration was made with the S. João Hospital in Porto. This database has 141 samples, containing CT scans, patients' clinical information, and features related to the nodule and its external surroundings annotated by radiologists from late 2019 to early 2020.

Table 5.1: Overview of the datasets considered for the project

	Hospital de S. João Database	**Radiogenomic Dataset of Non-Small Cell Lung Cancer**	**ILDS Database**
Short Name	S. João DB	NSCLC Radiogenomics	ILD DB
#Total CT Scans	141	211	128
#Suitable CT Scans	30	190	128
Nodule Segmentation	X	X	-
Lung Segmentation	-	-	X
Semantic Annotations	X	X	X

Despite the availability of 141 CT scans, only 30 were selected since the remaining did not possess any semantic annotations yet. These 27 patients were composed of 18 males and 9 females, all with lung cancer and ages between 52 and 87 years old. Among them, 3 were smokers, 6 were former smokers, 3 never smoked, and the remaining 16 subjects had an unknown smoking history.

This dataset held annotations of three distinct radiologists for every case, hence the mode between the three was adopted.

5.1.2 Radiogenomic Dataset of Non-Small Cell Lung Cancer

Radiogenomic Dataset of Non-Small Cell Lung Cancer [7] holds knowledge about 211 patients collected from 2008 and 2012 at Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System. This information consists of CT images, experts' annotations and segmentation maps of the nodule in the correspondent CT images. Data about the mutated genes and patient clinical history are also present.

Only 190 lung cancer patients from this database were considered since solely these owned lung annotations. This subset consisted of 67 females and 123 males with ages between 24 and 87, of which 30 were current smokers, 117 were former smokers and 43 never smoked at all.

5.1.3 ILD Database

At last, ILDS Database [17] is focused on interstitial lung diseases (ILDs) and related lung tissue patterns. It consists of 128 University Hospitals of Geneva subjects that were diagnosed with at least one of the thirteen most common ILDs. As the previously mentioned datasets, it comprises CT images, radiological annotations about lung tissue patterns and clinical information. Besides, this is the only dataset that contains lung segmentation masks.

This dataset held only patients who were not diagnosed with lung cancer, among them 45 women and 79 men with ages between 11 and 93 years old. From all the subjects, 28 were current smokers, 25 were former smokers and 44 never smoked, the remaining 27 had an unknown smoking history.

5.2 Data Preparation

5.2.1 Selection of Lung Characteristics to Detect

Due to the small number of suitable cases in the São João Database and the fact that its annotations were based on the NSCLC-Radiogenomics [7] leading to a significant quantity of similarities, the two were merged and named SJ-NSCLC Database. The ILDS Database was considered in separate since the criteria used to annotate the exams may not be equivalent to the other two datasets.

A criteria based on class distribution was adopted to choose which lung annotations would be detected in each database. Table 5.2 reveals the lung annotation from each database, its class distribution, and whether they fulfilled the criteria or not. To dodge overfitting linked to the class with a higher number of cases, the health conditions that had a ratio of the class with most examples to the class with fewer examples higher than 0.25 were selected.

The SJ-NSCLC database presented six annotations related to the lung and, after applying the aforementioned criteria, three remained: Nodules In Contralateral Lung, Satellite Nodules In Primary Lesion Lobe, and Emphysema. The IDLS Database [17] proved to be significantly unbalanced since from ten lung annotations only two met the guidelines: Fibrosis and Ground Glass.

Table 5.2: Selection of lung annotations for each dataset based on the union of two criteria. Criteria - The ratio of the class with most examples to the class with fewer examples must be higher than 0.25.

Dataset	Lung Annotatioin	#Present	#Absent	Ratio	Selected
SJ-NSCLC	**Nodules In Contralateral Lung**	**54**	**164**	**0.33**	✓
	Nodules In Non-Lesion Lobe Same Lung	39	178	0.22	-
	Satellite Nodules In Primary Lesion Lobe	**58**	**159**	**0.36**	✓
	Emphysema	**113**	**106**	**0.94**	✓
	Fibrosis	25	194	0.13	-
	Bronchiectasis	18	113	0.16	-
ILDS	**Fibrosis**	**39**	**69**	**0.57**	✓
	Ground Glass	**37**	**71**	**0.52**	✓
	Consolidation	14	94	0.15	-
	Reticulation	10	98	0.10	-
	Emphysema	5	103	0.05	-
	Bronchiectasis	8	100	0.08	-
	Bronchial wall thickening	1	107	0.01	-
	Cysts	3	105	0.03	-
	Micronodules	16	92	0.17	-
	Macronodules	7	101	0.07	-

5.2.2 CT Images Pre-Processing

The CT scans use Hounsfield Units (HU) to represent the information. The Hounsfield Scale is a quantitative scale to describe radiodensity, which considers that, under standard conditions of temperature and pressure, the water's radiodensity is marked as 0 HU and the air's radiodensity is -1000 HU [26].

Occasionally, the original content of the CT comes in other units and it is required to convert these to HU using the Rescale Slope and the Resclape Intercept fields present in the metadata associated with the scan. These two metadata fields are defined by the hardware manufacturer [50]. The conversion is achieved by performing Equation 5.1.

$$HUValue = PixelValue \times RescaleSlope + RescaleIntercept \tag{5.1}$$

As previously said, the air's radiodensity is -1000 HU, therefore, the CT values below that threshold are meaningless. The values above 400 HU describe the bone, which makes them irrelevant to this lung problem. Hence, to conclude the processing associated with the HU units, all the CT values below -1000 HU were set to -1000 HU and the values above 400 HU were fixed to 400 HU [26].

Two other relevant CT metadata fields are Slice Thickness and Slice Spacing, both in millimeters. Slice Thickness is the distance through the CT slice and Slice Spacing holds the space between each x coordinates and each y coordinates in the slice, as illustrated in Figure 5.1. These fields can alternate due to differences in hardware, so it is vital to normalize them and rescale the image accordingly. Slice Thickness and Slice Spacing of every scan were set to 1 and [1;1] respectively. The algorithm 2 describes how this normalization was completed.

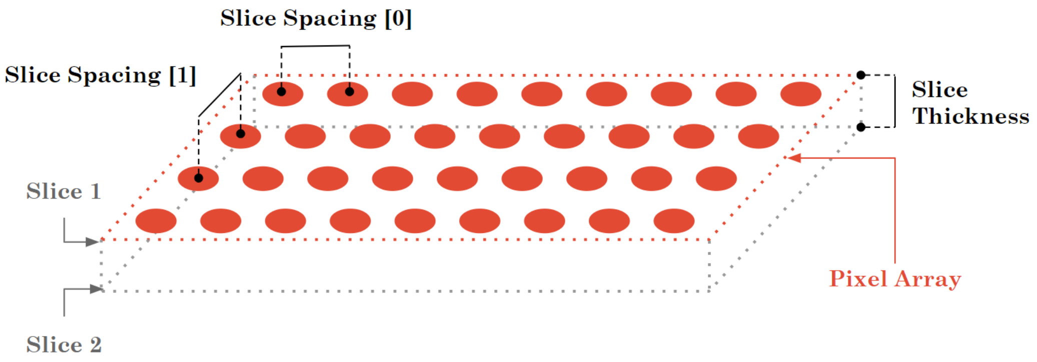


Figure 5.1: Illustration that represents the concepts of Slice Spacing and Slice Thickness in a CT slice.

Algorithm 2 Normalizing Spacing and Rescaling CT Image

```
$spacing \leftarrow append(slice\_spacing, slice\_thickness)$
$new\_spacing \leftarrow append(new\_slice\_spacing, new\_slice\_thickness)$

$resize\_factor \leftarrow spacing / new\_spacing$
$new\_shape \leftarrow \textbf{round}(image.shape \times resize\_factor)$
$real\_resize\_factor \leftarrow new\_shape / image.shape$

$image\_array \leftarrow rescale(image\_array, real\_resize\_factor)$

$new\_spacing \leftarrow spacing / real\_resize\_factor$
```

5.2.3 Lung Segmentation

Because the problem in hand is detecting lung pathologies, having the lung segmentation mask was fundamental. ILD Database dataset already held the lung segmentation masks drawn by experts, however, the lung masks of the remaining three had to be segmented by an adaptation of the Moreira Aresta [41] algorithm. Its pseudocode is presented in Algorithm 3 and the adjustments made are subsequently explained.

Algorithm 3 Pseudocode of the Moreira Aresta [41] algorithm to segment the lung

```
$candidates \leftarrow scan < -300$

$candidates \leftarrow \textbf{removeSmallCandidates}(candidates)$
$candidates \leftarrow \textbf{removeMarginCandidates}(candidates)$

$lungSegmentation \leftarrow candidates.\textbf{merge()}$
$lungSegmentation \leftarrow lungSegmentation.\textbf{MorphologicalCloseness()}$
$lungSegmentation \leftarrow lungSegmentation.\textbf{MorphologicalDilation()}$
```

When the histogram of a CT's HU values is plotted, like in Figure 5.2, two peaks are clearly observed. The most prominent peak corresponds to the lung parenchyma volume and the second one matches the fat and muscle around the lung.

The most reliable lung segmentation threshold is the HU value that coincides with the mean of these two peaks [21]. Though oftentimes that middle point is around -300 HU as applied in the original algorithm, that may not be true for every dataset or even every scan in the same dataset. Using a dynamic threshold dependent on the histogram peaks made the algorithm more robust as shown in Figure 5.3.

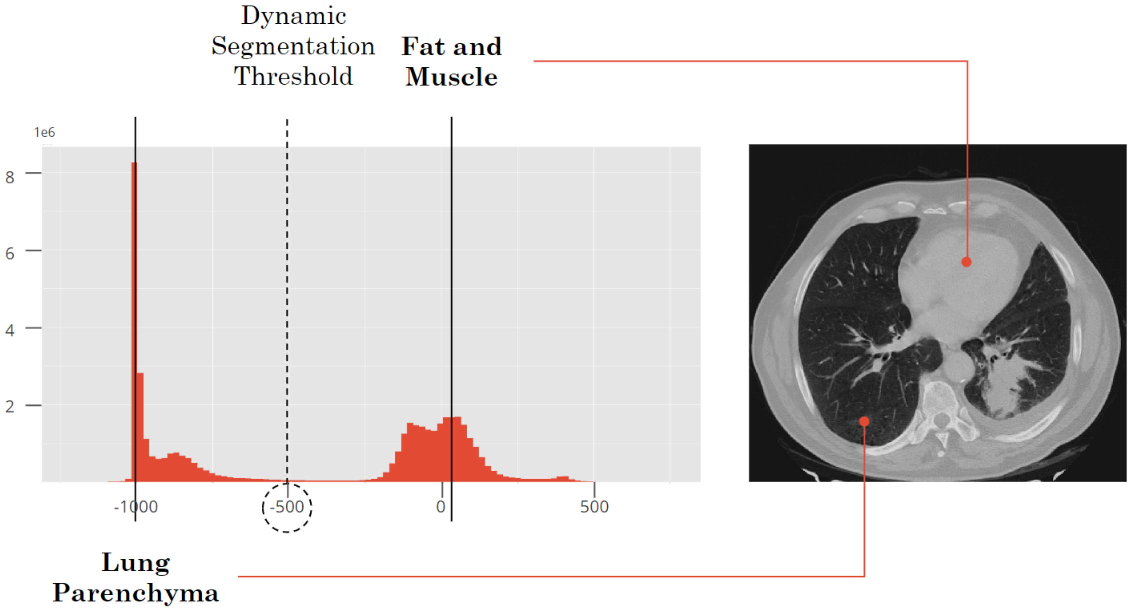


Figure 5.2: Relevant peaks identification on a CT scan HU values histogram. The fittest lung segmentation threshold is the midpoint between the higher peak, which represents the lung parenchyma, and the second higher peak, which represents fat and muscle. Adapted from Farag et al. [21].

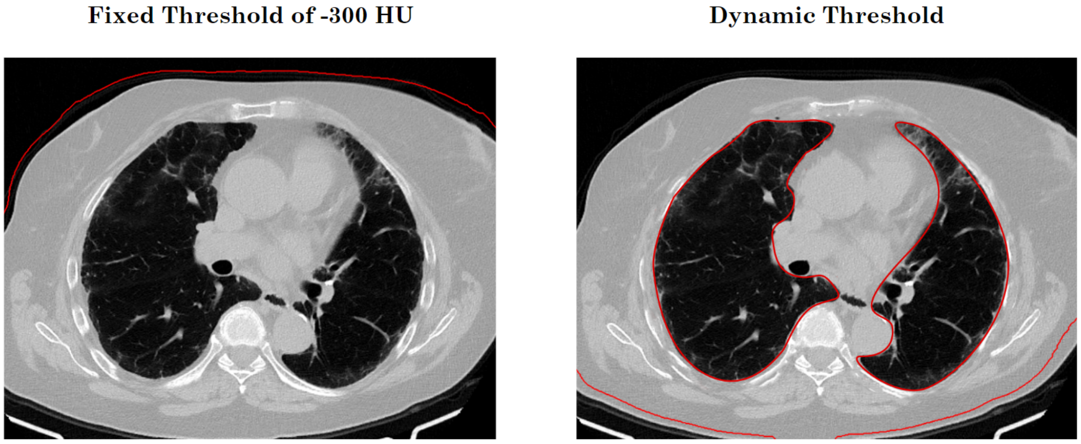


Figure 5.3: Results of a lung segmentation using a fixed threshold of -300 HU (left) and a dynamic threshold based on the peaks of the CT scan HU values histogram (right).

Depending on the scan, the threshold may cover regions that are not linked to the lung as exposed in the right segmentation of Figure 5.3. To mend this, the y coordinate of the lung candidate's centroid was taken into consideration. If the y coordinate of the candidate volume's centroid was above 39% or below 68% of the scan y width, that volume would be discarded. In Figure 5.4, the candidates and the final result using this amend are exposed. These values were empirically discovered using the datasets utilized in this work. In the future, a solution should be found to fix the problem without depending on the datasets.

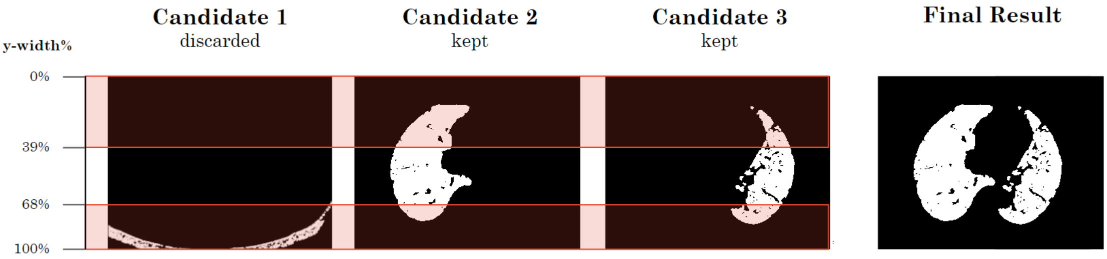


Figure 5.4: Discarded and kept analyzed volumes as well as the final result of a lung segmentation considering the *y* coordinate of the candidate volumes' centroid. Because the *y* coordinate of the Candidate 1's centroid laid between 68% and 100% of the *y* width, that candidate was discarded.

The last modification was the addition of the possibility to consider a given nodule segmentation. Sometimes the nodules are too close to the margin and the algorithm does not recognize it as part of the lung. The adapted version of the algorithm will merge the provided nodule mask to the lung segmentation. In Figure 5.5 the impact of this enhancement is displayed.

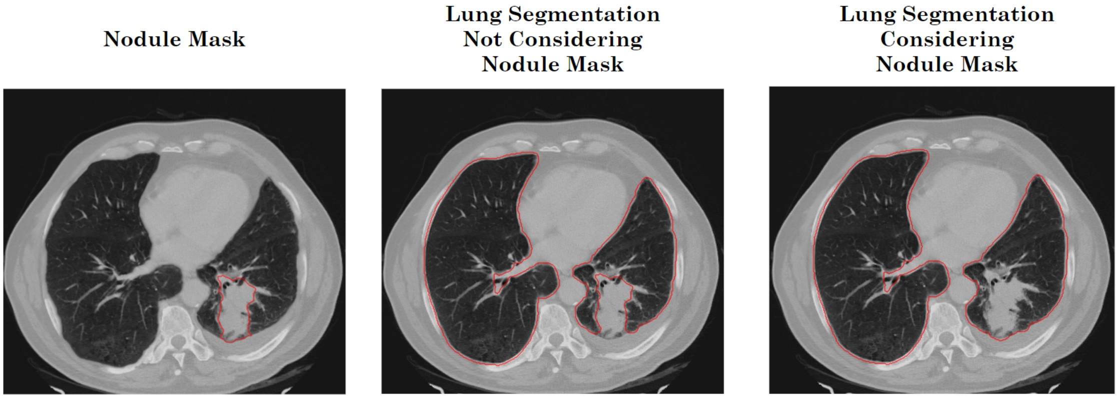


Figure 5.5: Nodule segmentation mask (first) and results of a lung segmentation not considering the nodule mask (second) and considering the nodule segmentation mask (third).

This lung segmentation algorithm has two main limitations. The first one is the use of a threshold as the algorithm basis since this will work poorly on scans from patients with severe state of lung diseases similar to the examples shown in Figure 2.1. The other limitation is the use of the location of the centroid empirically inferred from the datasets used since it will not be adaptable to other datasets.

5.3 Summary

Three datasets were studied to be employed in this work. All three present a different number of patients and were built in different locations, however, all have semantic annotations and CT images. Nonetheless, not all annotations were kept due to their unbalance. Pre-process the CT scans was fundamental to the task at hand, as well as the lung segmentation mask extraction.

Chapter 6

Lung Structures Characterization and Diseases Detection

After preprocessing the CTs, it is indispensable to generate bags that contain instances with features based on these exams to give as the input for the MIL models. Two types of bags generators were designed, *Radiomic Bag Generator* and *Hounsfield Units Bag Generator*, each one including some variations with different complexity levels. Due to some classes' unbalance, the Mera et al.[40] sampling technique for multiple instance bags was tested. A random search to find the best combination of the models' parameters was implemented to get the best possible results produced by that same model. Lastly, to make sure the results were accurate and to avoid overfitting, the technique k-fold cross-validation was adopted. This methodology is detailed in Section 6.1. The results of the experiment are shown in Section 6.2 and subsequently discussed in Section 6.3.

6.1 Methodology

To adapt the dataset images to a MIL problem, it was determined that a bag would represent one CT scan, distinct sections of the scan would be the bags' instances, and image features of those sub-regions would be the instances' features. However, there were almost infinite possibilities to divide the image into regions and represent the image features. Two bag generators were implemented, *Radiomic Bag Generator* and *Hounsfield Units Bag Generator*, and differed mainly on the complexity of the instances' shape and the features' nature.

The first bag generator was denominated *Radiomic Bag Generator* since the instances' features are the radiomic features of that region. Radiomic features aim to extract from a medical image a significant amount of quantitive features to infer clinical or pathological information about the patient. These features can be categorized into three classes: Histogram features (First-order statistics), Morphological features (Shape-based features), and Texture-based features

(Higher-order statistics).These radiomic features used in this work were extracted using the package PyRadiomics [52] and are fully listed in Appendix A. The literature established Radiomics can be employed in several lung problems to achieve high-grade results.

The instances were the intersection of the areas created by dividing the lung segmentation bounding box in $n \times n \times n$ (n being the number of parts in each axis) and the lung segmentation itself. Two variants of this bag generator were employed differing on the number of parts each axis of the lung segmentation bounding box was divided: *Radiomics-5* (Figure 6.1) and *Radiomics-10* (Figure 6.2) .

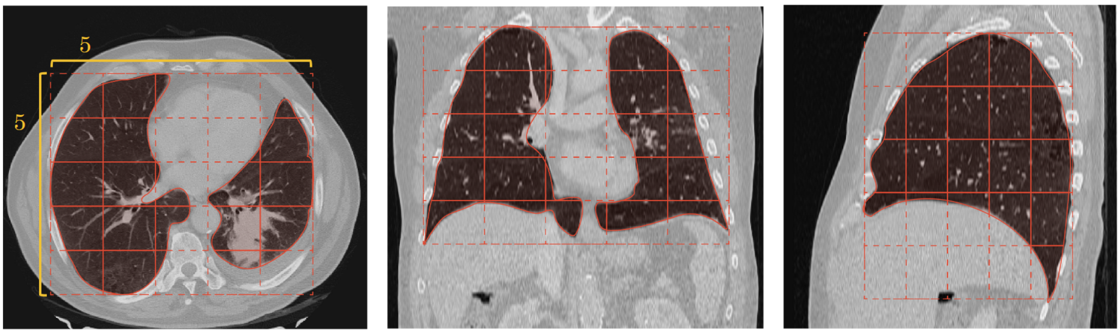


Figure 6.1: Axial, coronal, and sagittal views of a CT bag and the instances produced by the *Radiomics-5 Bag Generator*. The elected instances are filled with an orange translucent opacity and have a solid stroke, while the rejected ones have a dashed stroke.

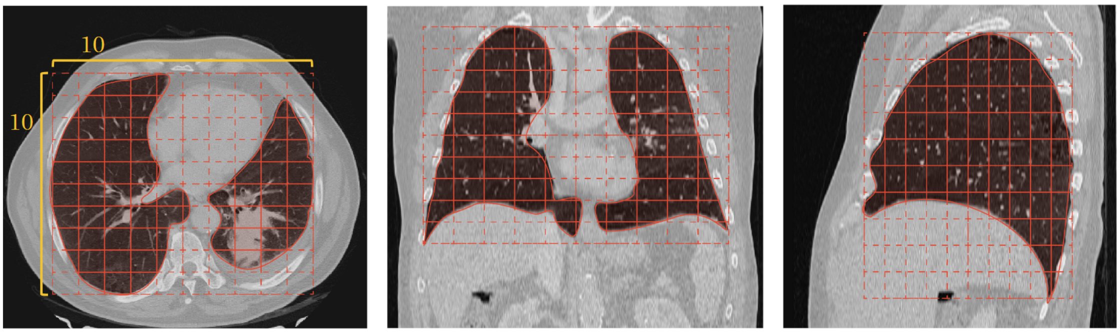


Figure 6.2: Axial, coronal, and sagittal views of a CT bag and the instances produced by the *Radiomics-10 Bag Generator*. The elected instances are filled with an orange translucent opacity and have a solid stroke, while the rejected ones have a dashed stroke.

The second bag generator was an adaptation of Single Blob with no Neighbors proposed by Maron and Ratan [37] and explained in Section 3.2. The first adaptation was using 3D instances and bags instead of 2D as the original SB. It was determined each scan would have $10 \times 10 \times 10$ instances composed of $2 \times 2 \times 2$ points, hence the image and the lung segmentation were resized to $20 \times 20 \times 20$ points. The second adjustment was, instead of using the value of red, blue, and green as the instance's features as in the original bag generator, it was used the HU of each instance's point since the RGB values are not available in a CT.

After creating the instances and respective features, three ways of creating bags were employed differing on the content of the corresponding instances. *HU-all* bags (Figure 6.3) were constituted by all the instances in the scan, *HU-some* bags (Figure 6.4) were composed by the instances that had at least a part that was lung and, finally, *HU-only* bags (Figure 6.5) were exclusively made of instances that were entirely lung. Note that, since the representations of Figure 6.3, Figure 6.4, and Figure 6.5 are in 2D, only four points can be seen in each instance, however in reality they are composed of eight points considering the CT scans are 3D.

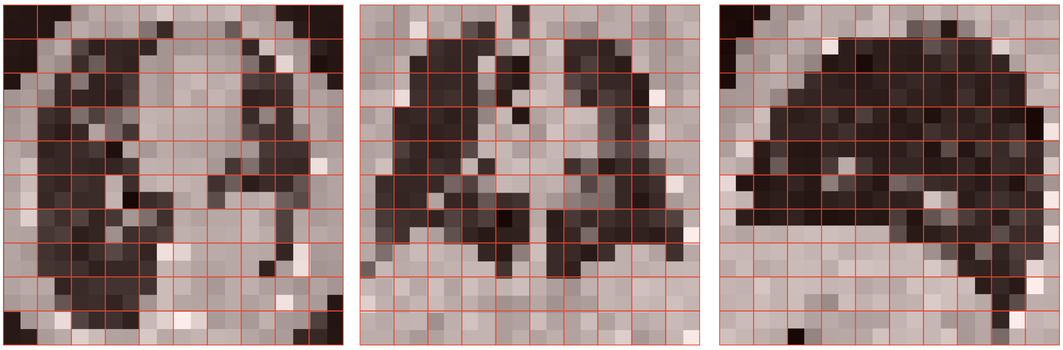


Figure 6.3: Axial, coronal, and sagittal views of a CT bag and the instances produced by the *HU-all Bag Generator*. All instances were selected.

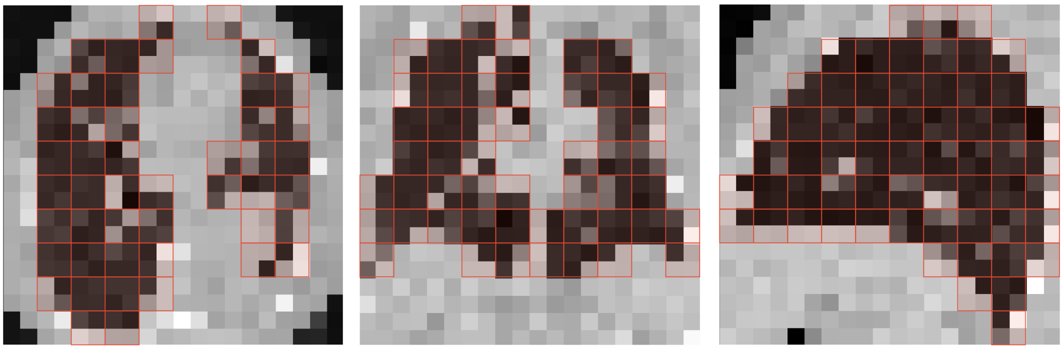


Figure 6.4: Axial, coronal, and sagittal views of a CT bag and the instances produced by *HU-some Bag Generator*. The instances that had at least a part that was lung were selected.

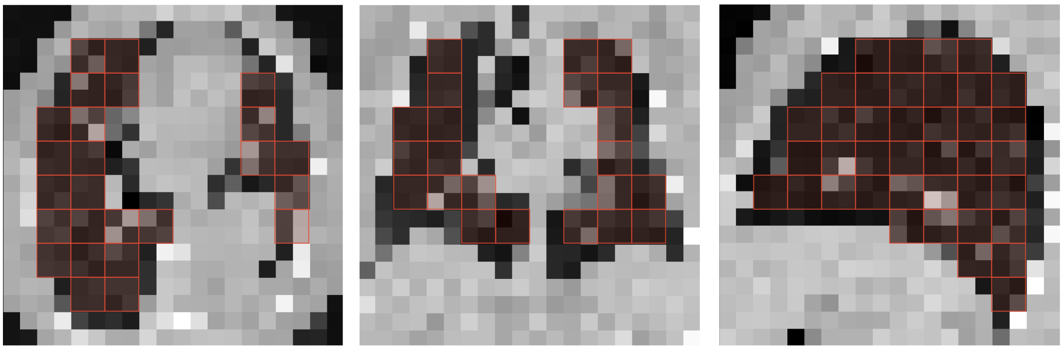


Figure 6.5: Axial, coronal, and sagittal views of a CT bag and the instances produced by the *HU-only Bag Generator*. Only instances that were entirely lung were selected.

To study the impact of the data unbalance in a MIL problem, experiments employing undersampling, the Mera et al. [40] oversampling algorithm for MIL explained in Section 3.3, or no sampling at all were conducted. Both sampling techniques were used so the ratio between the positive classes and negative classes on the train sets would be 1:1.

Making sure the experiment's results were the most accurate and not a consequence of an overfitted model was crucial, so a 5-Fold Stratified Cross-Validation was employed. For each fold, the best parameters concerning each model were sought using a Random Search Strategy with 20 iterations and the maximization of the AUC as the objective function. This Hyperparameter Tuning strategy was developed from scratch based on [1] as the used model were not compatible with the packages that held an Hyperparameter Tuning implementation. From the eleven distinct models described in Section 3.4, NSK was adopted for the experiments since bag-based classifiers are much more efficient than instance-based classifiers,and this classifier was proved one of the best in the literature [19]. Table 6.1 presents the values consider for the Hyperparamenter Tuning of the model. Figure 6.6 illustrates how the 5-Fold Cross-Validation, Sampling, Hyperparameter Tuning, and the adopted model interact with each other.

Table 6.1: List of values used for the Random Search in Hyperparameter Tuning in NSK Model

Parameter	Values
gamma	1, 0.1, 0.01, 0.001, 0.0001
C	0.1, 1, 10, 100, 1000
scale C	True, False
p	True, False
kernel	linear, quadratic, polynomial, rbf

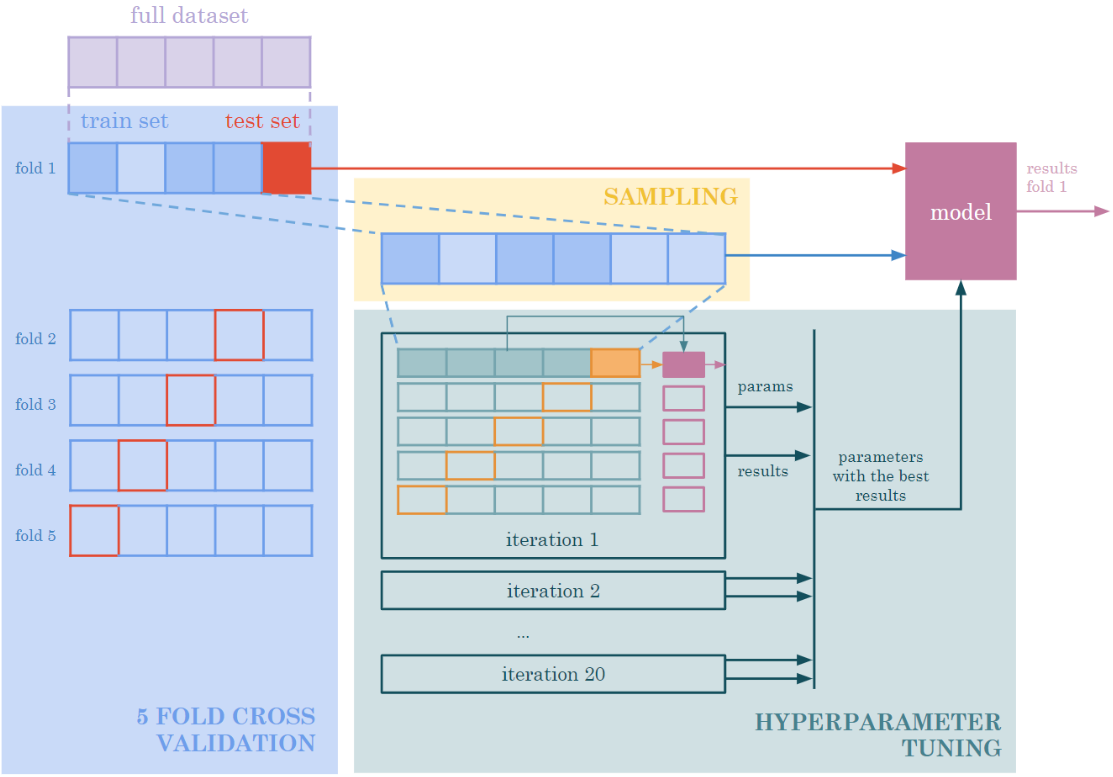


Figure 6.6: An illustration that shows how the 5-Fold Cross-Validation, Sampling, Hyperparameter Tuning, and the adopted model interact with each other. For each fold, the full dataset is split into a train set and a test set in a way that the later is different for every fold. Afterward, a Sampling technique is applied to the train set and the result is put through a 20 iteration Hyperparameter Tuning process. Each iteration is composed of a new 5-fold Cross-Validation. The model uses the train set and the parameters chosen by the Hyperparameter Tuning to label the test set, giving the results for the concerned fold.

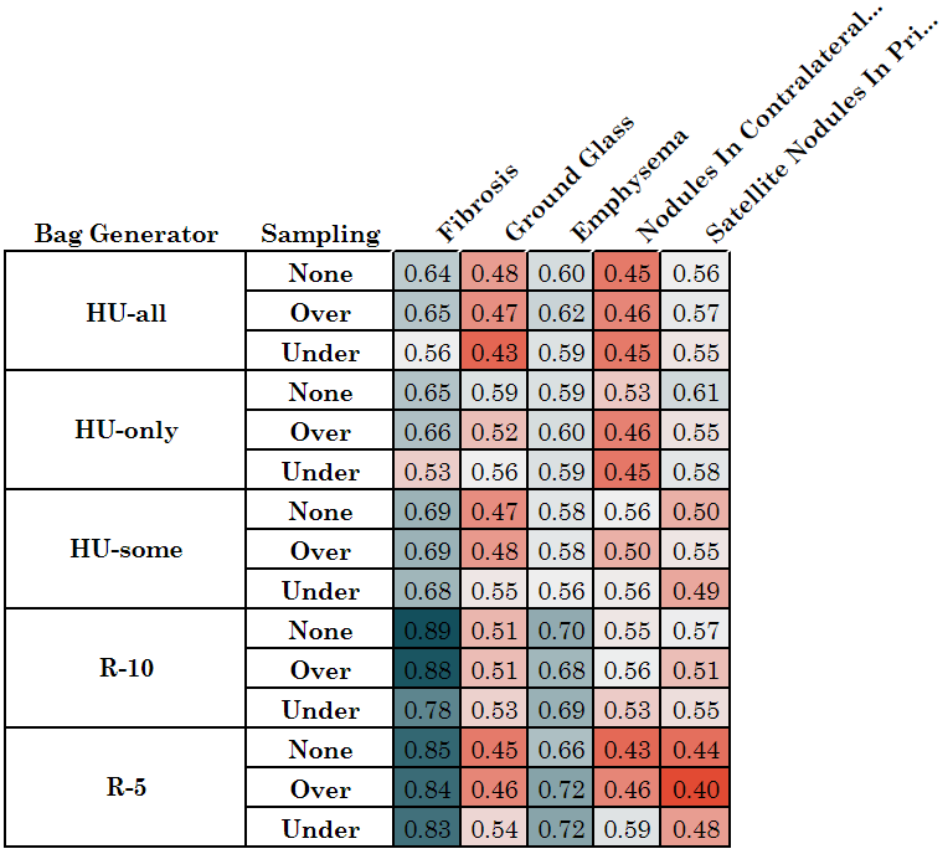
6.2 Results

Considering the 5 labels to detect (Emphysema, Satellite Nodules In Primary Lesion Lobe, Nodules In Contralateral Lung, Fibrosis and Ground Glass), the 5 bag generators designed (*Radiomics-5*, *Radiomics-10*, *HU-all*, *HU-some* and *HU-only*), and the 3 sampling procedures (none, oversampling and undersampling), 75 experiments were done in total. The full list of these experiments' results is presented in Appendix B. The 75 experiments are studied with distinct points of view in this section. Firstly, an overall analysis is made in Subsection 6.2.1, followed by a label analysis in Subsection 6.2.2, a sampling per label analysis in Subsection 6.2.3, a bag generator analysis in Subsection 6.2.4, and, finally, a sampling analysis in Subsection 6.2.5.

6.2.1 Overall Results Analysis

Table 6.2 exhibits the AUC of each experiment in a heatmap form, where the cold colors stand for the best results, and the warm colors are used for the worst outcomes. These results are in a range from 0.40 to 0.89, the smaller AUC corresponding to the detection of Satellite Nodules In Primary Lesion Lobe using the bag generator *Radiomics-5* with oversampling, and the best AUC matching the detection of Fibrosis with the bag generator *Radiomics-10* and with no sampling at all. With the colors help, it can be immediately concluded that Fibrosis owns the best results among labels, followed by Emphysema, and the Radiomics bags generators hold better AUCs than the Hounsfield bag generators. As for the sampling, no conclusions can be taken directly from this heatmap.

Table 6.2: Heatmap with the AUCs of the 75 experiments. Cold colors stand for the best results, and the warm colors are used for the worst outcomes.



Bag Generator	Sampling	Fibrosis	Ground Glass	Emphysema	Nodules In Contralateral...	Satellite Nodules In Pri...
HU-all	None	0.64	0.48	0.60	0.45	0.56
	Over	0.65	0.47	0.62	0.46	0.57
	Under	0.56	0.43	0.59	0.45	0.55
HU-only	None	0.65	0.59	0.59	0.53	0.61
	Over	0.66	0.52	0.60	0.46	0.55
	Under	0.53	0.56	0.59	0.45	0.58
HU-some	None	0.69	0.47	0.58	0.56	0.50
	Over	0.69	0.48	0.58	0.50	0.55
	Under	0.68	0.55	0.56	0.56	0.49
R-10	None	0.89	0.51	0.70	0.55	0.57
	Over	0.88	0.51	0.68	0.56	0.51
	Under	0.78	0.53	0.69	0.53	0.55
R-5	None	0.85	0.45	0.66	0.43	0.44
	Over	0.84	0.46	0.72	0.46	0.40
	Under	0.83	0.54	0.72	0.59	0.48

Table 6.3 presents the top 10 experiments with the best AUCs, which confirm the conclusions taken from the heatmap in Table 6.2. That is, the patterns with the best detection results are Fibrosis and Emphysema with the Radiomic bag generators. As for sampling, for the same label and the same bag generator, the use of no sampling surpasses oversampling, which outdoes undersampling. The best 10 AUCs are in a range from 0.89 to 0.69, and the corresponding accuracy has a range from 0.64 to 0.83.

Table 6.3: Top 10 experiments with the best AUCs.

Label	Bag Generator	Sampling	AUC	Accuracy	Precison	Recall	F1-Score	Time(s)
Fibrosis	**R-10**	**None**	0.89	0.83	0.82	0.82	0.82	3584
Fibrosis	**R-10**	**Over**	0.88	0.81	0.80	0.80	0.79	3183
Fibrosis	**R-5**	**None**	0.85	0.77	0.76	0.76	0.75	149
Fibrosis	**R-5**	**Over**	0.83	0.76	0.75	0.75	0.74	194
Fibrosis	**R-5**	**Under**	0.82	0.73	0.73	0.75	0.72	91
Fibrosis	**R-10**	**Under**	0.78	0.72	0.70	0.70	0.69	2441
Emphysema	**R-5**	**Over**	0.72	0.66	0.66	0.66	0.65	493
Emphysema	**R-5**	**Under**	0.72	0.64	0.65	0.64	0.64	424
Emphysema	**R-10**	**None**	0.71	0.69	0.69	0.69	0.68	20313
Emphysema	**R-10**	**Under**	0.69	0.69	0.70	0.69	0.69	17208

Table 6.4 shows the best results for each label considering AUC. Once more Fibrosis and Emphysema confirmed to be by far the labels with best results. The other three patterns have AUC higher than 0.5, indicating that the model can assist the detections. The non-use of sampling is more common than oversampling or undersampling. Curiously, the bag generator HU-only could exceed the Radiomics bag generators in the detection of Satellite Nodules In Primary Lesion Lobe and Ground Glass.

Table 6.4: Experiments of each label with the best AUC.

Label	Bag Generator	Sampling	Accuracy	AUC	Precison	Recall	F1-Score	Time(s)
Fibrosis	**R-10**	**None**	0.83	0.89	0.82	0.83	0.82	3584
Emphysema	**R-5**	**Over**	0.66	0.72	0.66	0.66	0.65	493
Satellite Nodules In Primary Lesion Lobe	**HU-only**	**None**	0.62	0.61	0.56	0.57	0.55	947
Ground Glass	**HU-only**	**None**	0.64	0.59	0.60	0.60	0.59	348
Nodules In Contralateral Lung	**R-5**	**Under**	0.55	0.59	0.55	0.56	0.51	120

6.2.2 Label Results Analysis

Table 6.5 shows the average results for each label, and Figure 6.7 reveals the plot comparing the three central metrics, AUC, Accuracy, and F1-Score. From these two resources, we can see that Fibrosis and Emphysema had great average results in every metric. On another hand, the F1-score of the other three labels was below 0.5, which may indicate that few combinations of bag generators and sampling techniques work as good as the experiments in Table 6.4 for the detection of these labels detection.

Table 6.5: Average results for each label.

Label	Average of AUC	Average of Accuracy	Average of Precison	Average of Recall	Average of F1-Score	Average of Time(s)
Fibrosis	0.72	0.70	0.68	0.68	0.67	2856
Ground Glass	0.50	0.56	0.50	0.51	0.50	3042
Emphysema	0.63	0.61	0.61	0.61	0.60	10199
Nodules In Contralateral Lung	0.50	0.59	0.50	0.50	0.47	6878
Satellite Nodules In Primary Lesion Lobe	0.53	0.59	0.52	0.53	0.49	12460

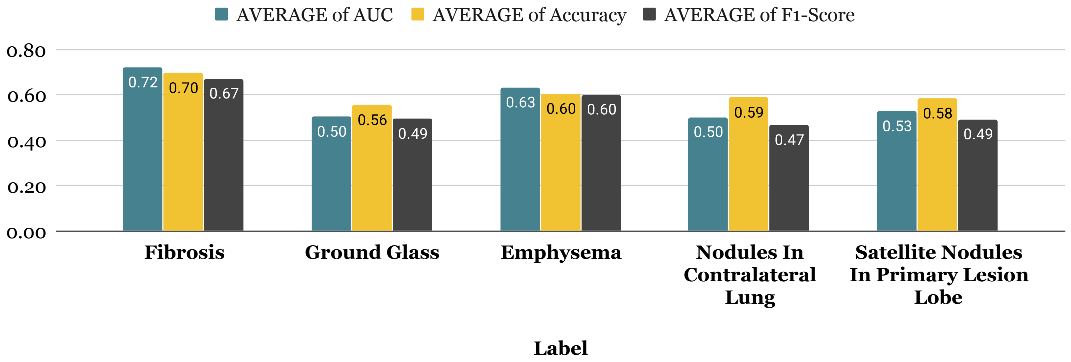


Figure 6.7: Plot comparing the average AUC, Accuracy, and F1-Score for each label.

Figure 6.8 employs the information of Table 6.5 and Table 6.6 to compare the number of training examples per fold in each class (solid lines) and the respective F1-Score (dashed lines), being the positive class represented by the blue lines and the negative class represented by the red lines. At first sight, it can be observed that when the number of training examples converges, which happens with Emphysema, the F1-Scores also converge. It can be due to the fact that, when there is a balance between classes, the model learns better to distinguish between each one. As expected, in the other labels the class with better F1-Score is the class with more training examples, the negative class.

Table 6.6: Average F1-Scores of each class and respective number of training examples per folder for each label.

Label	Average of F1-Score [positive]	Average of F1-Score [negative]	Average of #Training [positive]	Average of #Training [negative]
Fibrosis	0.59	0.75	39	47
Ground Glass	0.34	0.65	39	48
Emphysema	0.60	0.60	88.3	86.7
Nodules In Contralateral Lung	0.24	0.69	72.3	101.7
Satellite Nodules In Primary Lesion Lobe	0.30	0.69	73	100

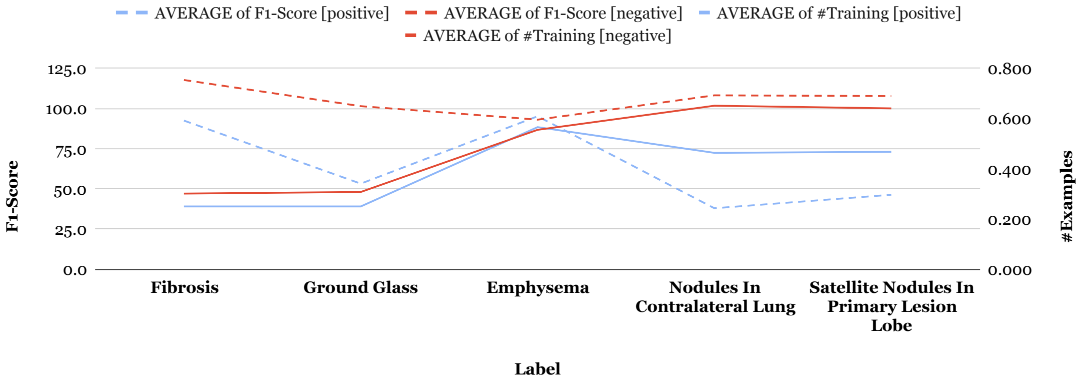


Figure 6.8: Plot that compares the number of training examples per fold in each class and the respective F1-Score. Blue lines - Positive class; Red lines - Negative class; Solid lines - Number of training examples; Dashed Lines - F1-Scores.

6.2.3 Sampling per Label Results Analysis

Table 6.7 shows the average results for each sampling technique per label and Figure 6.9 displays the corresponding plot comparing the AUC, Accuracy, and F1-Score. Considering the same label, the AUC does not vary much. In other words, the sampling technique does not significantly influence the final result. Looking at the two nodules labels (Satellite Nodules In Primary Lesion Lobe and Nodules In Contralateral Lung), the Accuracy increases when using oversampling, however, the other metrics do not. This shows how looking only at Accuracy can be misleading.

Table 6.7: Average results for each sampling technique per label.

Label	Sampling	Average of AUC	Average of Accuracy	Average of Recall	Average of Precison	Average of F1-Score	Average of Time(s)
Emphysema	**None**	0.63	0.60	0.60	0.60	0.60	10991
Emphysema	**Over**	0.64	0.61	0.61	0.61	0.60	8703
Emphysema	**Under**	0.63	0.61	0.61	0.61	0.60	10901
Fibrosis	**None**	0.74	0.71	0.70	0.69	0.69	3458
Fibrosis	**Over**	0.74	0.73	0.71	0.72	0.71	3132
Fibrosis	**Under**	0.68	0.64	0.64	0.64	0.62	1978
Ground Glass	**None**	0.50	0.56	0.51	0.49	0.49	3507
Ground Glass	**Over**	0.49	0.58	0.51	0.50	0.50	3757
Ground Glass	**Under**	0.52	0.53	0.52	0.52	0.50	1860
Nodules In Contralateral Lung	**None**	0.51	0.61	0.50	0.47	0.46	8033
Nodules In Contralateral Lung	**Over**	0.49	0.65	0.50	0.51	0.48	9704
Nodules In Contralateral Lung	**Under**	0.51	0.50	0.51	0.51	0.46	2895
Satellite Nodules In Primary Lesion Lobe	**None**	0.54	0.60	0.53	0.52	0.51	23853
Satellite Nodules In Primary Lesion Lobe	**Over**	0.52	0.64	0.52	0.51	0.50	10802
Satellite Nodules In Primary Lesion Lobe	**Under**	0.53	0.51	0.53	0.52	0.48	2724

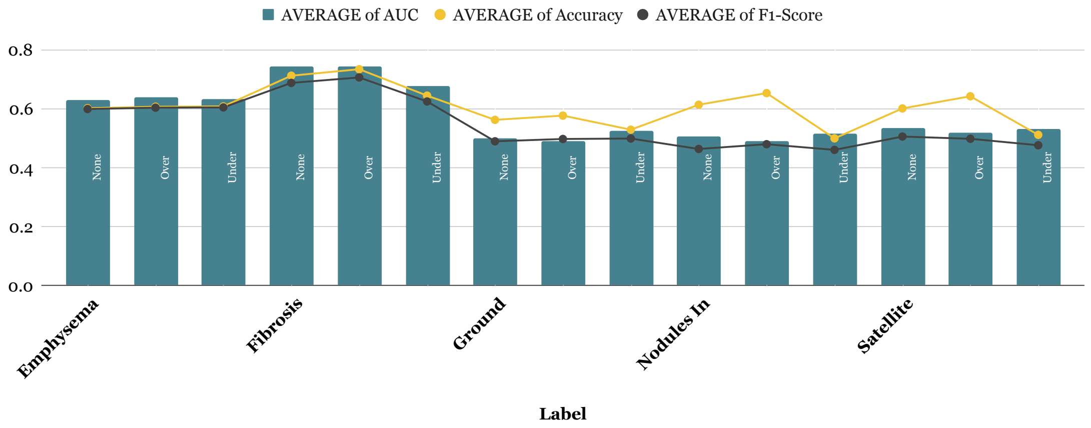


Figure 6.9: Plot comparing the average AUC, Accuracy, and F1-Score for each sampling technique per label.

The plot in Figure 6.10 uses the information from Table 6.7 and Table 6.8 to compare the training examples per fold of each sampling technique per label with the time the model took to train and test. As expected, a clear correlation can be seen in almost all labels. However, the time the model took to perform Satellite Nodules In Primary Lesion Lobe detections with no sampling technique is unusual given its number of training examples. A possible justification is the occurrence of a performance problem in the machine when it executed the respective experiments.

Table 6.8: Number of positive, negative and total training examples per fold for each sampling technique per label.

Label	Sampling	Average of #Training [positive]	Average of #Training [negative]	Average of #Training
Emphysema	**None**	90	85	175
Emphysema	**Over**	90	90	180
Emphysema	**Under**	85	85	170
Fibrosis	**None**	31	55	86
Fibrosis	**Over**	55	55	110
Fibrosis	**Under**	31	31	62
Ground Glass	**None**	30	57	87
Ground Glass	**Over**	57	57	114
Ground Glass	**Under**	30	30	60
Nodules In Contralateral Lung	**None**	43	131	174
Nodules In Contralateral Lung	**Over**	131	131	262
Nodules In Contralateral Lung	**Under**	43	43	86
Satellite Nodules In Primary Lesion Lobe	**None**	46	127	173
Satellite Nodules In Primary Lesion Lobe	**Over**	127	127	254
Satellite Nodules In Primary Lesion Lobe	**Under**	46	46	92

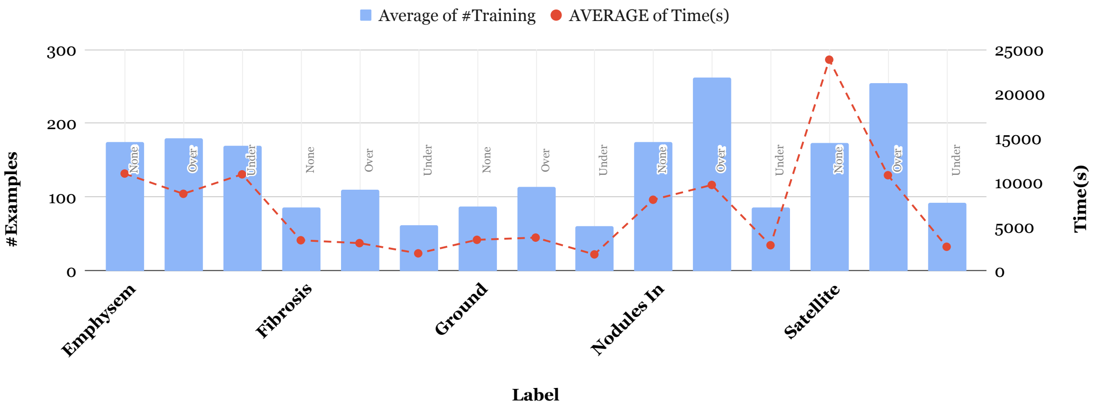


Figure 6.10: Plot comparing the total number of training examples per fold and the amount of time for each sampling technique per label.

6.2.4 Bag Generator Results Analysis

Table 6.2.4 shows the average results for each bag generator and Figure 6.11 compares the corresponding AUC, Accuracy, and F1-Score. The Radiomics bag generators appear to have a better performance as stated before.

Table 6.9: Average results for each bag generator.

Bag Generator	Average of AUC	Average of Accuracy	Average of Precison	Average of Recall	Average of F1-Score	Average of Time(s)
HU-all	0.54	0.54	0.53	0.54	0.51	18323
HU-only	0.56	0.61	0.55	0.56	0.54	671
HU-some	0.56	0.61	0.54	0.55	0.53	4841
R-10	0.63	0.64	0.60	0.60	0.58	11183
R-5	0.59	0.64	0.58	0.59	0.57	415

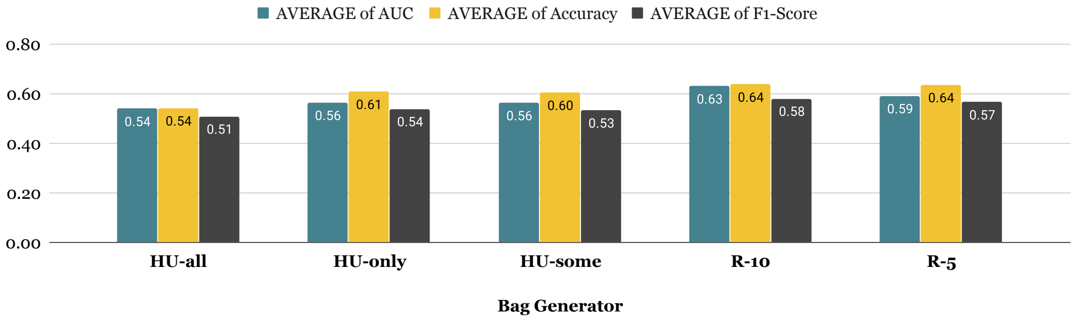


Figure 6.11: Plot comparing the average AUC, Accuracy, and F1-Score for each bag generator.

The plot in Figure 6.12 employs the information from Table and Table 6.10 to compare the average number of bags, instances, and features generated by each bag generator and the average time the model takes to finish the experiments. A correlation can be seen between the sum between instances and features, and the amount of time. Since the average of bags is the same in every bag generator conclusions cannot be made on how this variable influences the performance.

Table 6.10: Average number of bags, instances and features.

Bag Generator	Average of #Bags	Average of #Instances	Average of #Features
HU-all	174	1000	8
HU-only	174	170.54	8
HU-some	174	462.29	8
R-10	174	597.81	107
R-5	174	109.54	107

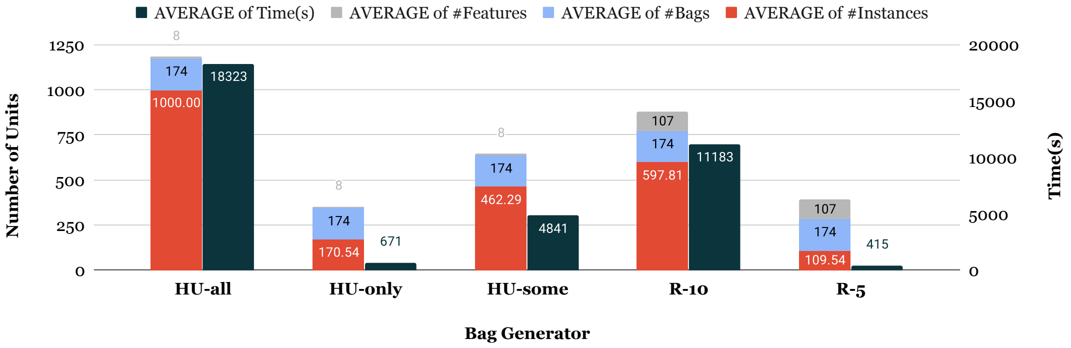


Figure 6.12: Plot comparing the average number of instances, features and bags with the amount of time for each bag generator.

6.2.5 Sampling Results Analysis

Table 6.11 shows the average results for each sampling technique, and Figure 6.13 compares the respective AUC, Accuracy, and F1-Score. The outcomes of each technique are not very different.

Table 6.11: Average results for each sampling technique.

Sampling	Average of AUC	Average of Accuracy	Average of Precison	Average of Recall	Average of F1-Score	Average of Time(s)
None	0.58	0.62	0.55	0.57	0.55	9968
Over	0.58	0.64	0.57	0.57	0.56	7220
Under	0.58	0.56	0.56	0.56	0.53	4072

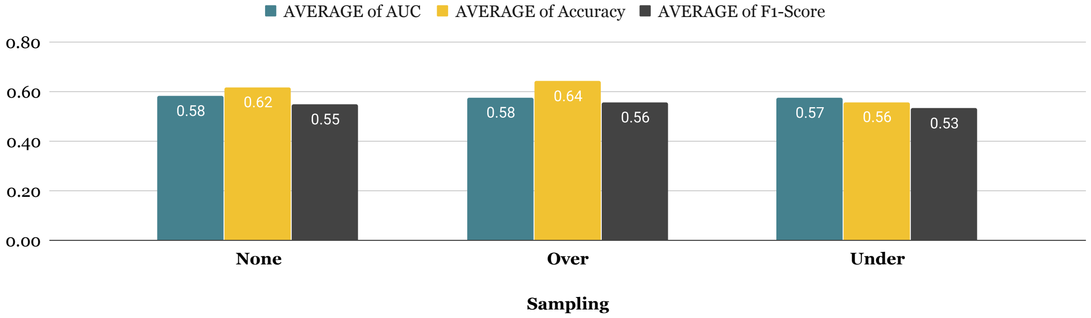


Figure 6.13: Plot comparing the average AUC, Accuracy, and F1-Score for each sampling technique.

The plots in Figure 6.14 use the information from Table 6.11 and Table 6.12 to compare the average number of positive and negative training examples per fold with the respective F1-Scores. When no sampling technique is employed, since the negative class is the one with more training examples, the F1-Score of this class is, as expected, higher. However, unlike the F1-Scores convergence that occurs in plot of Figure 6.9, these F1-scores stayed different even when the number of training examples was the same, in other words, when employing undersampling or oversampling. This may indicate that the bags generated in the oversampling are not similar to the real ones and the bags removed from the undersampling were essential for the detection.

Table 6.12: Average number of training examples per fold of each class and the respective F1-Scores.

Sampling	Average of #Training [negative]	Average of #Training [positive]	Average of F1-Score [negative]	Average of F1-Score [positive]
None	91	48	0.69	0.41
Over	92	92	0.72	0.40
Under	47	47	0.62	0.45

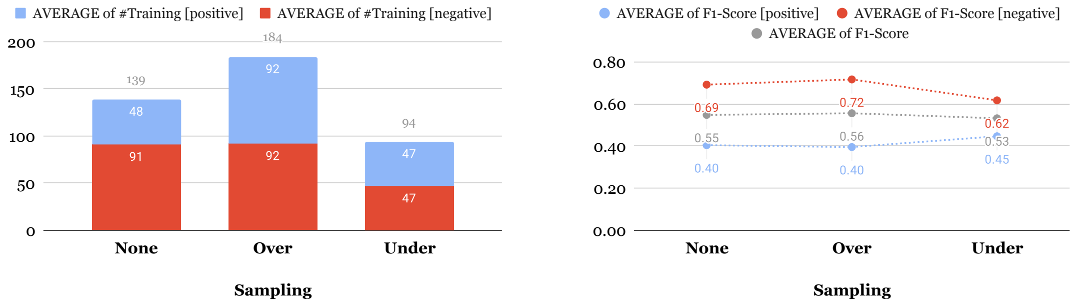


Figure 6.14: Plots comparing the number of training examples per fold in each class and the respective F1-Scores.

6.3 Discussion

The following list summarizes the main conclusions that can be drawn with this work's results and a respective examination.

- **The patterns with the best detection results are Fibrosis and Emphysema, reaching an AUC of 0.89 and 0.72, respectively** - This can be justified by the fact that the Fibrosis cases in the used dataset were in a severe stage. Moreover, the Emphysema is more of a spread out through the lungs than a localized disease unliked the other three lung patterns detected [16];

- ***Radiomics-10* is the bag generator with better results** - Even though it extracts the same radiomic features as the *Radiomics-5* bag generator, it also contains more and smaller instances, so more detailed information is given to the model. This comes with a performance expense since a problem generated by *Radiomics-10* takes much more time to solve than a problem represented by *Radiomics-5*;

- **Radiomics bags generators have better results than the Hounsfield Bag Generators** - The Radiomic bag generators extract many more image features than the Hounsfield Bag Generators that only look at the HU values of the instances;

- **The sampling technique does not significantly influence the final result** - If the results are not that different, no sampling technique should be used since no fake bags are created, as in oversampling, and no possible important ones are removed, as done in undersampling, being the problem more authentic.

6.4 Summary

Five bag generators were designed to create a MIL representation to detect some lung patters in CT scans. To validate the results a pipeline with three phases was created: 5-fold cross-validation, sampling, and hyperparameter tuning. The MIL model employed for the detection was NSK. The experiments with better results were the detection of Fibrosis and Emphysema, reaching an AUC of 0.89 and 0.72, respectively.

Chapter 7

Conclusions

Due to the high mortality caused by lung cancer and its late diagnosis, it is necessary to classify the gene mutation status to give the patient the best treatment possible. To avoid the pain caused by biopsy, the used method at the moment, it is necessary to develop non-invasive methods.

In spite of existing a lot of models in the literature that predict the status mutation, most of them only use the cancer nodule features. Some studies show that if some features external to the nodule are taken into consideration, the results of the models may improve. This studied resulted in the detection of multiple lung patterns to be later employed in gene mutation prediction.

Thus far this work was the first to use Multiple Instance Learning to detect Fibrosis in CT scans, reaching an outstanding AUC of 0.89. Considering the Emphysema detection, this study's result (AUC=0.72) is slightly below the literature (AUC=0.82), however, the existing study does not mention the MIL model used and, therefore, the results cannot be verified. Even though the results achieved for the other three lung patterns (Satellite Nodules In Primary Lesion Lobe (AUC=0.61), Nodules In Contralateral Lung (AUC=0.59) and Ground Glass (AUC=0.59) were not as exceptional, they hint that the detection of these visual patterns is possible and can be further studied. Five MIL bag generators were designed, among which *Radiomics-5* and *Radiomics-10* stand out. These should be applied to other MIL problems to verify if the good results still uphold.

As this study showed promising results, multiple research lines related to it can be conducted in the future, being some listed hereafter.

- **Detect the remaining lung patterns of the used datasets** - The detected lung patterns were selected based on a class balance criteria, which does not guarantee they would have better results than the remaining ones;

- **Implement more variations of the Radiomics bag generators** - The *Radiomics-10* bag generator proved to be better than the *Radiomcis-5*, which may indicate the detail quantity is important. In the future, a *Radiomics-15* or *Radiomics-20* could be implemented and tested to see if they could reach even better results. However, they would come with a high time cost associated;

- **Use a predetermined number of random instances instead of the whole set** - The subset of instances may be able to represent the whole CT scan. This is regularly used in the literature to decrease the amount of time the model needs to train and test;

- **Experiment with some bag generators of the literature** - This work employed bag generators designed especially for it. It should be tested if the literature bag generators can achieve better results for this particular problem than the ones designed;

- **Merge datasets** - Even though the datasets may have been annotated with different criteria, it would be interesting to see how the framework behaves training and testing with more than one dataset in each experiment;

- **Employ bigger datasets** - This would not only enhance the learning of the model but could also give the possibility to detect other lung patterns;

- **Explore different MIL models** - NSK was elected since it was proven to be one of the most efficient and with better results in the literature. However, given the problem, other models can have even higher results. Moreover, this would allow us to have a better understanding of how the model selection determine the outcomes;

- **Detect the lung cancer mutated gene** - Even though this work intended to detect semantic lung patterns that are understandable by experts for later predicting the mutated gene, it could be interesting to try to predict the mutated gene directly from the CT scans using this study's methodology.

Appendix A

List of Radiomic Features

First Order Statistics:

1. Energy
2. Total Energy
3. Entropy
4. Minimum
5. 10th percentile
6. 90th percentile
7. Maximum
8. Mean
9. Median
10. Interquartile Range
11. Range
12. Mean Absolute Deviation (MAD)
13. Robust Mean Absolute Deviation (rMAD)
14. Root Mean Squared (RMS)
15. Standard Deviation
16. Skewness
17. Kurtosis
18. Variance
19. Uniformity

Shape-based (3D):

1. Mesh Volume
2. Voxel Volume
3. Surface Area
4. Surface Area to Volume ratio
5. Sphericity
6. Compactness 1
7. Compactness 2
8. Spherical Disproportion
9. Maximum 3D diameter
10. Maximum 2D diameter (Slice)
11. Maximum 2D diameter (Column)
12. Maximum 2D diameter (Row)
13. Major Axis Length
14. Minor Axis Length
15. Least Axis Length
16. Elongation

17. Flatness

Gray Level Cooccurence Matrix:

1. Autocorrelation
2. Joint Average
3. Cluster Prominence
4. Cluster Shade
5. Cluster Tendency
6. Contrast
7. Correlation
8. Difference Average
9. Difference Entropy
10. Difference Variance
11. Joint Energy
12. Joint Entropy
13. Informational Measure of Correlation (IMC) 1
14. Informational Measure of Correlation (IMC) 2
15. Inverse Difference Moment (IDM)
16. Maximal Correlation Coefficient (MCC)
17. Inverse Difference Moment Normalized (IDMN)
18. Inverse Difference (ID)
19. Inverse Difference Normalized (IDN)
20. Inverse Variance
21. Maximum Probability
22. Sum Average
23. Sum Entropy
24. Sum of Squares

Gray Level Run Length Matrix:

1. Short Run Emphasis (SRE)
2. Long Run Emphasis (LRE)
3. Gray Level Non-Uniformity (GLN)
4. Gray Level Non-Uniformity Normalized (GLNN)
5. Run Length Non-Uniformity (RLN)
6. Run Length Non-Uniformity Normalized (RLNN)
7. Run Percentage (RP)
8. Gray Level Variance (GLV)
9. Run Variance (RV)
10. Run Entropy (RE)
11. Low Gray Level Run Emphasis (LGLRE)
12. High Gray Level Run Emphasis (HGLRE)
13. Short Run Low Gray Level Emphasis (SRLGLE)
14. Short Run High Gray Level Emphasis (SRHGLE)
15. Long Run Low Gray Level Emphasis (LRLGLE)
16. Long Run High Gray Level Emphasis (LRHGLE)

Gray Level Size Zone Matrix:

1. Small Area Emphasis (SAE)
2. Large Area Emphasis (LAE)

3. Gray Level Non-Uniformity (GLN)
4. Gray Level Non-Uniformity Normalized (GLNN)
5. Size-Zone Non-Uniformity (SZN)
6. Size-Zone Non-Uniformity Normalized (SZNN)
7. Zone Percentage (ZP)
8. Gray Level Variance (GLV)
9. Zone Variance (ZV)
10. Zone Entropy (ZE)
11. Low Gray Level Zone Emphasis (LGLZE)
12. High Gray Level Zone Emphasis (HGLZE)
13. Small Area Low Gray Level Emphasis (SALGLE)
14. Small Area High Gray Level Emphasis (SAHGLE)
15. Large Area Low Gray Level Emphasis (LALGLE)
16. Large Area High Gray Level Emphasis (LAHGLE)

Gray Level Dependence Matrix:

1. Small Dependence Emphasis (SDE)
2. Large Dependence Emphasis (LDE)
3. Gray Level Non-Uniformity (GLN)
4. Dependence Non-Uniformity (DN)
5. Dependence Non-Uniformity Normalized (DNN)
6. Gray Level Variance (GLV)
7. Dependence Variance (DV)
8. Dependence Entropy (DE)
9. Low Gray Level Emphasis (LGLE)
10. High Gray Level Emphasis (HGLE)
11. Small Dependence Low Gray Level Emphasis (SDLGLE)
12. Small Dependence High Gray Level Emphasis (SDHGLE)
13. Large Dependence Low Gray Level Emphasis (LDLGLE)
14. Large Dependence High Gray Level Emphasis (LDHGLE)

Neighbouring Gray Tone Difference Matrix:

1. Coarseness
2. Contrast
3. Cusyness
4. Complexity
5. Strength

Appendix B

List of Results

Table B.1: List of the 75 experiments' results.

Dataset	Bag Generator	Label	Sampling	AUC	Accuracy	Precison	Recall	F1-Score	F1-Score [positive]	F1-Score [negative]	Time(s)	#Training [positive]	#Training [negative]	#Instances	#Features
NSCLC_SJ	R-5	Emphysema	None	0.66	0.64	0.65	0.64	0.64	0.63	0.65	600	90	85	113.20	107
NSCLC_SJ	R-5	Emphysema	Under	0.72	0.64	0.65	0.64	0.64	0.62	0.65	424	85	85	113.20	107
NSCLC_SJ	R-5	Emphysema	Over	0.72	0.66	0.66	0.66	0.65	0.63	0.68	493	90	90	113.20	107
NSCLC_SJ	R-5	Satellite Nodules In Primary Lesion Lobe	None	0.44	0.55	0.48	0.48	0.46	0.67	0.24	610	46	127	113.20	107
NSCLC_SJ	R-5	Satellite Nodules In Primary Lesion Lobe	Under	0.48	0.50	0.49	0.49	0.45	0.61	0.30	96	46	46	113.20	107
NSCLC_SJ	R-5	Satellite Nodules In Primary Lesion Lobe	Over	0.40	0.64	0.46	0.48	0.47	0.76	0.17	1233	127	127	113.20	107
NSCLC_SJ	R-5	Nodules In Contralateral Lung	None	0.43	0.66	0.48	0.51	0.49	0.78	0.20	469	43	131	113.20	107
NSCLC_SJ	R-5	Nodules In Contralateral Lung	Under	0.59	0.55	0.55	0.56	0.51	0.64	0.38	120	43	43	113.20	107
NSCLC_SJ	R-5	Nodules In Contralateral Lung	Over	0.46	0.71	0.46	0.48	0.46	0.82	0.09	1217	131	131	113.20	107
NSCLC_SJ	R-10	Emphysema	None	0.70	0.68	0.69	0.69	0.68	0.69	0.68	20313	90	85	646.03	107
NSCLC_SJ	R-10	Emphysema	Under	0.69	0.69	0.69	0.69	0.69	0.66	0.71	17208	85	85	646.03	107
NSCLC_SJ	R-10	Emphysema	Over	0.68	0.65	0.65	0.65	0.65	0.63	0.66	13249	90	90	646.03	107
NSCLC_SJ	R-10	Satellite Nodules In Primary Lesion Lobe	None	0.57	0.63	0.51	0.52	0.51	0.75	0.28	14450	46	127	646.03	107
NSCLC_SJ	R-10	Satellite Nodules In Primary Lesion Lobe	Under	0.55	0.53	0.53	0.54	0.49	0.63	0.36	4331	46	46	646.03	107
NSCLC_SJ	R-10	Satellite Nodules In Primary Lesion Lobe	Over	0.51	0.72	0.53	0.54	0.53	0.82	0.23	28501	127	127	646.03	107
NSCLC_SJ	R-10	Nodules In Contralateral Lung	None	0.55	0.54	0.51	0.51	0.48	0.65	0.31	19272	43	131	646.03	107
NSCLC_SJ	R-10	Nodules In Contralateral Lung	Under	0.53	0.45	0.51	0.51	0.43	0.53	0.33	4608	43	43	646.03	107
NSCLC_SJ	R-10	Nodules In Contralateral Lung	Over	0.56	0.74	0.69	0.58	0.57	0.83	0.31	23515	131	131	646.03	107
NSCLC_SJ	HU-all	Emphysema	None	0.60	0.57	0.57	0.57	0.57	0.57	0.57	25538	90	85	1000.00	8
NSCLC_SJ	HU-all	Emphysema	Under	0.59	0.57	0.57	0.57	0.57	0.56	0.58	28434	85	85	1000.00	8
NSCLC_SJ	HU-all	Emphysema	Over	0.62	0.57	0.57	0.57	0.57	0.57	0.56	19954	90	90	1000.00	8
NSCLC_SJ	HU-all	Satellite Nodules In Primary Lesion Lobe	None	0.56	0.52	0.54	0.56	0.49	0.61	0.37	96731	46	127	1000.00	8
NSCLC_SJ	HU-all	Satellite Nodules In Primary Lesion Lobe	Under	0.55	0.52	0.54	0.55	0.49	0.61	0.37	7214	46	46	1000.00	8
NSCLC_SJ	HU-all	Satellite Nodules In Primary Lesion Lobe	Over	0.57	0.53	0.56	0.57	0.49	0.61	0.37	13528	127	127	1000.00	8
NSCLC_SJ	HU-all	Nodules In Contralateral Lung	None	0.45	0.44	0.47	0.45	0.41	0.54	0.28	11478	43	131	1000.00	8
NSCLC_SJ	HU-all	Nodules In Contralateral Lung	Under	0.45	0.44	0.46	0.45	0.41	0.54	0.28	7653	43	43	1000.00	8

Dataset	Bag Generator	Label	Sampling	AUC	Accuracy	Precison	Recall	F1-Score	F1-Score [positive]	F1-Score [negative]	Time(s)	#Training [positive]	#Training [negative]	#Instances	#Features
NSCLC_SJ	**HU-all**	**Nodules In Contralateral Lung**	**Over**	0.46	0.45	0.47	0.45	0.41	0.54	0.28	14732	131	131	1000.00	8
NSCLC_SJ	**HU-some**	**Emphysema**	**None**	0.58	0.56	0.56	0.56	0.55	0.56	0.54	7547	90	85	462.31	8
NSCLC_SJ	**HU-some**	**Emphysema**	**Under**	0.56	0.56	0.56	0.56	0.56	0.57	0.55	7586	85	85	462.31	8
NSCLC_SJ	**HU-some**	**Emphysema**	**Over**	0.58	0.58	0.58	0.58	0.57	0.57	0.58	8989	90	90	462.31	8
NSCLC_SJ	**HU-some**	**Satellite Nodules In Primary Lesion Lobe**	**None**	0.50	0.68	0.50	0.53	0.51	0.79	0.24	6525	46	127	462.31	8
NSCLC_SJ	**HU-some**	**Satellite Nodules In Primary Lesion Lobe**	**Under**	0.49	0.48	0.50	0.50	0.45	0.57	0.32	1726	46	46	462.31	8
NSCLC_SJ	**HU-some**	**Satellite Nodules In Primary Lesion Lobe**	**Over**	0.55	0.66	0.50	0.50	0.49	0.78	0.20	9380	127	127	462.31	8
NSCLC_SJ	**HU-some**	**Nodules In Contralateral Lung**	**None**	0.56	0.67	0.48	0.51	0.49	0.79	0.20	7913	43	131	462.31	8
NSCLC_SJ	**HU-some**	**Nodules In Contralateral Lung**	**Under**	0.56	0.56	0.56	0.58	0.52	0.65	0.38	1837	43	43	462.31	8
NSCLC_SJ	**HU-some**	**Nodules In Contralateral Lung**	**Over**	0.50	0.70	0.46	0.50	0.47	0.82	0.12	7244	131	131	462.31	8
NSCLC_SJ	**HU-only**	**Emphysema**	**None**	0.59	0.55	0.55	0.55	0.55	0.53	0.56	957	90	85	171.74	8
NSCLC_SJ	**HU-only**	**Emphysema**	**Under**	0.59	0.57	0.57	0.57	0.57	0.56	0.57	855	85	85	171.74	8
NSCLC_SJ	**HU-only**	**Emphysema**	**Over**	0.60	0.58	0.58	0.58	0.58	0.57	0.58	831	90	90	171.74	8
NSCLC_SJ	**HU-only**	**Satellite Nodules In Primary Lesion Lobe**	**None**	0.61	0.62	0.56	0.57	0.55	0.73	0.37	947	46	127	171.74	8
NSCLC_SJ	**HU-only**	**Satellite Nodules In Primary Lesion Lobe**	**Under**	0.58	0.52	0.55	0.56	0.49	0.60	0.39	251	46	46	171.74	8
NSCLC_SJ	**HU-only**	**Satellite Nodules In Primary Lesion Lobe**	**Over**	0.55	0.66	0.50	0.52	0.51	0.78	0.24	1370	127	127	171.74	8
NSCLC_SJ	**HU-only**	**Nodules In Contralateral Lung**	**None**	0.53	0.75	0.40	0.49	0.44	0.85	0.03	1034	43	131	171.74	8
NSCLC_SJ	**HU-only**	**Nodules In Contralateral Lung**	**Under**	0.45	0.50	0.47	0.47	0.44	0.62	0.26	259	43	43	171.74	8
NSCLC_SJ	**HU-only**	**Nodules In Contralateral Lung**	**Over**	0.46	0.66	0.48	0.49	0.48	0.78	0.19	1814	131	131	171.74	8
ILDS	**R-5**	**Fibrosis**	**None**	0.85	0.77	0.76	0.76	0.75	0.82	0.69	149	31	55	104.06	107
ILDS	**R-5**	**Fibrosis**	**Under**	0.83	0.73	0.74	0.75	0.72	0.77	0.68	91	31	31	104.06	107
ILDS	**R-5**	**Fibrosis**	**Over**	0.84	0.76	0.75	0.75	0.74	0.81	0.68	194	55	55	104.06	107
ILDS	**R-5**	**Ground Glass**	**None**	0.45	0.62	0.51	0.53	0.52	0.73	0.30	150	30	57	104.06	107
ILDS	**R-5**	**Ground Glass**	**Under**	0.54	0.51	0.55	0.55	0.51	0.54	0.47	107	30	30	104.06	107
ILDS	**R-5**	**Ground Glass**	**Over**	0.46	0.59	0.50	0.50	0.49	0.71	0.27	273	57	57	104.06	107
ILDS	**R-10**	**Fibrosis**	**None**	0.89	0.83	0.82	0.83	0.82	0.87	0.77	3584	31	55	525.48	107
ILDS	**R-10**	**Fibrosis**	**Under**	0.78	0.72	0.70	0.70	0.69	0.77	0.61	2441	31	31	525.48	107
ILDS	**R-10**	**Fibrosis**	**Over**	0.88	0.81	0.80	0.81	0.79	0.84	0.75	3183	55	55	525.48	107
ILDS	**R-10**	**Ground Glass**	**None**	0.51	0.48	0.43	0.43	0.43	0.58	0.27	5938	30	57	525.48	107
ILDS	**R-10**	**Ground Glass**	**Under**	0.53	0.54	0.48	0.51	0.47	0.62	0.33	3022	30	30	525.48	107
ILDS	**R-10**	**Ground Glass**	**Over**	0.51	0.56	0.48	0.49	0.48	0.68	0.28	4129	57	57	525.48	107
ILDS	**HU-all**	**Fibrosis**	**None**	0.64	0.64	0.62	0.63	0.61	0.71	0.50	10610	31	55	1000.00	8
ILDS	**HU-all**	**Fibrosis**	**Under**	0.56	0.59	0.57	0.58	0.56	0.65	0.47	5839	31	31	1000.00	8
ILDS	**HU-all**	**Fibrosis**	**Over**	0.65	0.66	0.64	0.64	0.63	0.73	0.54	9357	55	55	1000.00	8
ILDS	**HU-all**	**Ground Glass**	**None**	0.48	0.59	0.49	0.54	0.51	0.69	0.32	8845	30	57	1000.00	8
ILDS	**HU-all**	**Ground Glass**	**Under**	0.43	0.45	0.45	0.45	0.43	0.51	0.35	4472	30	30	1000.00	8
ILDS	**HU-all**	**Ground Glass**	**Over**	0.47	0.59	0.48	0.51	0.49	0.70	0.27	10455	57	57	1000.00	8
ILDS	**HU-some**	**Fibrosis**	**None**	0.69	0.67	0.65	0.66	0.65	0.73	0.58	2645	31	55	462.27	8
ILDS	**HU-some**	**Fibrosis**	**Under**	0.68	0.59	0.59	0.59	0.57	0.64	0.50	1371	31	31	462.27	8
ILDS	**HU-some**	**Fibrosis**	**Over**	0.69	0.73	0.70	0.69	0.69	0.79	0.59	2576	55	55	462.27	8
ILDS	**HU-some**	**Ground Glass**	**None**	0.47	0.49	0.40	0.43	0.41	0.60	0.22	2256	30	57	462.27	8
ILDS	**HU-some**	**Ground Glass**	**Under**	0.55	0.59	0.56	0.56	0.56	0.67	0.45	1514	30	30	462.27	8
ILDS	**HU-some**	**Ground Glass**	**Over**	0.48	0.57	0.52	0.50	0.50	0.67	0.33	3511	57	57	462.27	8
ILDS	**HU-only**	**Fibrosis**	**None**	0.65	0.64	0.62	0.61	0.60	0.72	0.49	300	31	55	168.75	8
ILDS	**HU-only**	**Fibrosis**	**Under**	0.53	0.60	0.58	0.58	0.57	0.67	0.48	149	31	31	168.75	8
ILDS	**HU-only**	**Fibrosis**	**Over**	0.66	0.71	0.71	0.67	0.67	0.79	0.55	351	55	55	168.75	8
ILDS	**HU-only**	**Ground Glass**	**None**	0.59	0.64	0.60	0.60	0.59	0.73	0.45	348	30	57	168.75	8
ILDS	**HU-only**	**Ground Glass**	**Under**	0.56	0.55	0.53	0.53	0.52	0.62	0.42	183	30	30	168.75	8
ILDS	**HU-only**	**Ground Glass**	**Over**	0.52	0.58	0.53	0.54	0.53	0.68	0.38	419	57	57	168.75	8

References

[1] Intro to Model Tuning: Grid and Random Search | Kaggle. URL https://www.kaggle.com/willkoehrsen/intro-to-model-tuning-grid-and-random-search. Last accessed on 06/06/2020.

[2] Lung Cancer Guide | What You Need to Know. URL https://www.cancer.org/cancer/lung-cancer. Last accessed on 23/01/2020.

[3] Mayo Clinic: Diseases and Conditions. URL https://www.mayoclinic.org/diseases-conditions/. Last accessed on 04/02/2020.

[4] Hugo J.W.L. Aerts, Emmanuel Rios Velazquez, Ralph T.H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications, 5, jun 2014. ISSN 20411723. doi: 10.1038/ncomms5006.

[5] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support Vector Machines for Multi ple-Instance Learning. Technical report, 2003.

[6] Amina Asif, Wajid Arshad Abbasi, Farzeen Munir, Asa Ben-Hur, and Fayyaz ul Amir Afsar Minhas. pyLEMMINGS: Large Margin Multiple Instance Classification and Ranking for Bioinformatics Applications. nov 2017.

[7] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann N.C. Leung, Michael Kadoch, Chuong D. Hoang, Joseph Shrager, Andrew Quon, Daniel L. Rubin, Sylvia K. Plevritis, and Sandy Napel. Data descriptor: A radiogenomic dataset of non-small cell lung cancer. Scientific Data, 5, 2018. ISSN 20524463. doi: 10.1038/sdata.2018.202.

[8] C. Bhuvaneswari, P. Aruna, and D. Loganathan. A new fusion model for classifi cation of the lung diseases using genetic algorithm. Egyptian Informatics Journal, 15(2):69–77, 2014. ISSN 11108665. doi: 10.1016/j.eij.2014.05.001.

[9] Jinbo Bi and Jianming Liang. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In Proceedings of the IEEE Computer

Society Conference on Computer Vision and Pattern Recognition, 2007. ISBN 1424411807. doi: 10.1109/CVPR.2007.383141.

[10] Razvan C. Bunescu and Raymond J. Mooney. Multiple instance learning for sparse positive bags. In ACM International Conference Proceeding Series, volume 227, pages 105–112, 2007. doi: 10.1145/1273496.1273510.

[11] Yiyuan Cao, Haibo Xu, Meiyan Liao, Yanjuan Qu, Liying Xu, Dongyong Zhu, Bicheng Wang, and Sufang Tian. Associations between clinical data and computed tomography features in patients with epidermal growth factor receptor mutations in lung adenocarcinoma. International Journal of Clinical Oncology, 23(2):249–257, apr 2018. ISSN 14377772. doi: 10.1007/s10147-017-1197-8.

[12] Marc André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition, 77:329–353, may 2018. ISSN 00313203. doi: 10.1016/j.patcog.2017.10.009.

[13] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321–357, jan 2002. ISSN 10769757. doi: 10.1613/jair.953.

[14] Zenghui Cheng, Fei Shan, Yuesong Yang, Yuxin Shi, and Zhiyong Zhang. CT characteristics of non-small cell lung cancer with epidermal growth factor receptor mutation: A systematic review and meta-analysis. BMC Medical Imaging, 17(1), jan 2017. ISSN 14712342. doi: 10.1186/s12880-016-0175-3.

[15] Veronika Cheplygina, Lauge Sørensen, David M.J. Tax, Jesper Holst Pedersen, Marco Loog, and Marleen De Bruijne. Classification of COPD with multiple instance learning. In Proceedings - International Conference on Pattern Recognition, pages 1508–1513. Institute of Electrical and Electronics Engineers Inc., dec 2014. ISBN 9781479952083. doi: 10.1109/ICPR.2014.268.

[16] Veronika Cheplygina, Isabel Pino Pena, Jesper Holst Pedersen, David A. Lynch, Lauge Sorensen, and Marleen De Bruijne. Transfer Learning for Multicenter Classification of Chronic Obstructive Pulmonary Disease. IEEE Journal of Biomedical and Health Informatics, 22(5):1486–1496, sep 2018. ISSN 21682194. doi: 10.1109/JBHI.2017.2769800.

[17] Adrien Depeursinge, Alejandro Vargas, Alexandra Platon, Antoine Geissbuhler, Pierre Alexandre Poletti, and Henning Müller. Building a reference multimedia database for interstitial lung diseases. Computerized Medical Imaging and Graphics, 36(3):227–238, 2012. ISSN 08956111. doi: 10.1016/j.compmedimag.2011.07.003.

[18] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 89(1-2):31–71, jan 1997. ISSN 00043702. doi: 10.1016/s0004-3702(96)00034-3.

[19] Gary Doran and Soumya Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. In Machine Learning, volume 97, pages 79–102. Kluwer Academic Publishers, dec 2014. doi: 10.1007/s10994-013-5429-5.

[20] M. Murat Dundar, Glenn Fung, Balaji Krishnapuram, and R. Bharat Rao. Multiple-instance learning algorithms for computer-aided detection. IEEE Transactions on Biomedical Engineering, 55(3):1015–1021, mar 2008. ISSN 00189294. doi: 10.1109/TBME.2007.909544.

[21] Amal Farag, James Graham, and Aly Farag. Robust segmentation of lung tissue in chest CT scanning. In Proceedings - International Conference on Image Processing, ICIP, pages 2249–2252, 2010. ISBN 9781424479948. doi: 10.1109/ICIP.2010.5651233.

[22] Jia Gang, Feng Yuan, and Zheng Bing. Medical image semantic annotation based on MIL. In 2013 ICME International Conference on Complex Medical Engineering, CME 2013, pages 85–90, 2013. ISBN 9781467329699. doi: 10.1109/ICCME.2013.6548217.

[23] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, Alex J Smola AlexSmola, and anueduau Rsise. Multi-Instance Kernels. Technical report, 2002.

[24] Olivier Gevaert, Sebastian Echegaray, Amanda Khuong, Chuong D. Hoang, Joseph B. Shrager, Kirstin C. Jensen, Gerald J. Berry, H. Henry Guo, Charles Lau, Sylvia K. Plevritis, Daniel L. Rubin, Sandy Napel, and Ann N. Leung. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. Scientific Reports, 7, jan 2017. ISSN 20452322. doi: 10.1038/srep41674.

[25] J.W. Goldman, Z.S. Noor, J. Remon, B. Besse, and N. Rosenfeld. Are liquid biopsies a surrogate for tissue EGFR testing? Annals of Oncology, 29:i38–i46, jan 2018. ISSN 09237534. doi: 10.1093/annonc/mdx706.

[26] Adam O. Hebb and Andrew V. Poliakov. Imaging of deep brain stimulation leads using extended hounsfield unit CT. Stereotactic and Functional Neurosurgery, 87(3):155–160, jun 2009. ISSN 10116125. doi: 10.1159/000209296.

[27] Ahmed Hosny, Chintan Parmar, Thibaud P. Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J. Gillies, Raymond H. Mak, and Hugo J.W.L. Aerts. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Medicine, 15(11), nov 2018. ISSN 15491676. doi: 10.1371/journal.pmed.1002711.

[28] M. Infante, R. F. Lutman, S. Imparato, M. Di Rocco, G. L. Ceresoli, V. Torri, E. Morenghi, F. Minuti, S. Cavuto, E. Bottoni, F. Inzirillo, U. Cariboni, V. Errico, M. A. Incarbone, G. Ferraroli, G. Brambilla, M. Alloisio, and G. Ravasi. Differential diagnosis and management of

focal ground-glass opacities. European Respiratory Journal, 33(4):821–827, apr 2009. ISSN 09031936. doi: 10.1183/09031936.00047908.

[29] Melih Kandemir and Fred A. Hamprecht. Computer-aided diagnosis from weak supervision: A benchmarking study. Computerized Medical Imaging and Graphics, 42:44–50, jun 2015. ISSN 18790771. doi: 10.1016/j.compmedimag.2014.11.010.

[30] Sho Koyasu, Mizuho Nishio, Hiroyoshi Isoda, Yuji Nakamoto, and Kaori Togashi. Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT. Annals of Nuclear Medicine, 2019. ISSN 18646433. doi: 10.1007/s12149-019-01414-0.

[31] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G.P.M. Van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, André Dekker, and Hugo J.W.L. Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer, 48(4):441–446, mar 2012. ISSN 09598049. doi: 10.1016/j.ejca.2011.11.036.

[32] Letícia Ferro Leal, Flávia Escremim de Paula, Pedro De Marchi, Luciano de Souza Viana, Gustavo Dix Junqueira Pinto, Carolina Dias Carlos, Gustavo Noriz Berardinelli, José Elias Miziara, Carlos Maciel da Silva, Eduardo Caetano Albino Silva, Rui Pereira, Marco Antonio de Oliveira, Cristovam Scapulatempo-Neto, and Rui Manuel Reis. Mutational profile of Brazilian lung adenocarcinoma unveils association of EGFR mutations with high Asian ancestry and independent prognostic role of KRAS mutations. Scientific Reports, 9(1), dec 2019. ISSN 20452322. doi: 10.1038/s41598-019-39965-x.

[33] Cuifang Li, Shengdong Nie, Yuanjun Wang, and Xiwen Sun. Experimental investigation of fuzzy enhancement for nonsolid pulmonary nodules. In Proceedings - 2012 IEEE Symposium on Robotics and Applications, ISRA 2012, pages 756–759, 2012. ISBN 9781467322072. doi: 10.1109/ISRA.2012.6219301.

[34] Xiao Yang Li, Jun Feng Xiong, Tian Ying Jia, Tian Le Shen, Run Ping Hou, Jun Zhao, and Xiao Long Fu. Detection of epithelial growth factor receptor (EGFR) mutations on CT images of patients with lung adenocarcinoma using radiomics and/or multi-level residual convolutionary neural networks. Journal of Thoracic Disease, 10(12):6624–6635, dec 2018. ISSN 20776624. doi: 10.21037/jtd.2018.11.03.

[35] Ying Liu, Jongphil Kim, Fangyuan Qu, Shichang Liu, Hua Wang, Yoganand Balagurunathan, Zhaoxiang Ye, and Robert J. Gillies. CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. Radiology, 280(1):271–280, jul 2016. doi: 10.1148/radiol.2016151455.

[36] O L Mangasarian and E W Wild. Multiple Instance Classification via Successive Linear Programming. J Optim Theory Appl, 137(1):7–22, jun 2008. doi: 10.1007/s10957-007-9343-5.

[37] Oded Maron and Aparna Lakshmi Ratan. Multiple-Instance Learning for Natural Scene Classiication. 1998.

[38] José Marrugo-Ramírez, Mònica Mir, and Josep Samitier. Blood-based cancer biomarkers in liquid biopsy: A promising non-invasive alternative to tissue biopsy, oct 2018. ISSN 14220067.

[39] Jaime Melendez, Bram Van Ginneken, Pragnya Maduskar, Rick H.H.M. Philipsen, Klaus Reither, Marianne Breuninger, Ifedayo M.O. Adetifa, Rahmatulai Maane, Helen Ayles, and Clara I. Sánchez. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays. IEEE Transactions on Medical Imaging, 34(1): 179–192, jan 2015. ISSN 1558254X. doi: 10.1109/TMI.2014.2350539.

[40] Carlos Mera, Jose Arrieta, Mauricio Orozco-Alzate, and John Branch. A bag oversampling approach for class imbalance in multiple instance learning. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 9423, pages 724–731. Springer Verlag, 2015. ISBN 9783319257501. doi: 10.1007/978-3-319-25751-8_87.

[41] Guilherme Moreira Aresta. Detection of juxta-pleural lung nodules in computed tomography images. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 7 2016.

[42] Jarushka Naidoo and Alexander Drilon. Kras-mutant lung cancers in the era of targeted therapy. Advances in Experimental Medicine and Biology, 893:155–178, 2016. ISSN 22148019. doi: 10.1007/978-3-319-24223-1_8.

[43] Silas Nyboe Orting, Jens Petersen, Laura H. Thomsen, Mathilde M.W. Wille, and Marleen De Bruijne. Detecting emphysema with multiple instance learning. In Proceedings - International Symposium on Biomedical Imaging, volume 2018-April, pages 510–513. IEEE Computer Society, may 2018. ISBN 9781538636367. doi: 10.1109/ISBI.2018.8363627.

[44] Lais Osmani, Frederic Askin, Edward Gabrielson, and Qing Kay Li. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy, oct 2018. ISSN 10963650.

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

[46] Isabel Pino Peña, Veronika Cheplygina, Sofia Paschaloudi, Morten Vuust, Jesper Carl, Ulla Møller Weinreich, Lasse Riis Østergaard, and Marleen de Bruijne. Automatic emphysema detection using weakly labeled HRCT lung images. journals.plos.org, 13(10), oct 2018. doi: 10.1371/journal.pone.0205397.

[47] Endang Purba, Ei-ichiro Saita, and Ichiro Maruyama. Activation of the EGF Receptor by Ligand Binding and Oncogenic Mutations: The “Rotation Model”. Cells, 6(2):13, jun 2017. ISSN 2073-4409. doi: 10.3390/cells6020013.

[48] Xiaowei Qiu, Hang Yuan, and Bin Sima. Relationship between EGFR mutation and computed tomography characteristics of the lung in patients with lung adenocarcinoma. Thoracic Cancer, 10(2):170–174, feb 2019. ISSN 17597714. doi: 10.1111/1759-7714.12928.

[49] José Ramos, Thessa Kockelkorn, Bram Van Ginneken, Max A Viergever, Jan Grutters, Rui Ramos, and Aurélio Campilho. Learning Interstitial Lung Diseases CT Patterns from Reports Keywords. Technical report, 2013.

[50] Elmar Rendon-Gonzalez and Volodymyr Ponomaryov. Automatic Lung nodule segmentation and classification in CT images based on SVM. In 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves, MSMW 2016. Institute of Electrical and Electronics Engineers Inc., aug 2016. ISBN 9781509022663. doi: 10.1109/MSMW.2016.7537995.

[51] Stefania Rizzo, Sara Raimondi, Evelyn E.C. de Jong, Wouter van Elmpt, Francesca De Piano, Francesco Petrella, Vincenzo Bagnardi, Arthur Jochems, Massimo Bellomi, Anne Marie Dingemans, and Philippe Lambin. Genomics of non-small cell lung cancer (NSCLC): Association between CT-based imaging features and EGFR and K-RAS mutations in 122 patients—An external validation. European Journal of Radiology, 110:148–155, jan 2019. ISSN 18727727. doi: 10.1016/j.ejrad.2018.11.032.

[52] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. Cancer Research, 77(21):e104–e107, nov 2017. ISSN 15387445. doi: 10.1158/0008-5472.CAN-17-0339.

[53] Shuo Wang, Jingyun Shi, Zhaoxiang Ye, Di Dong, Dongdong Yu, Mu Zhou, Ying Liu, Olivier Gevaert, Kun Wang, Yongbei Zhu, Hongyu Zhou, Zhenyu Liu, and Jie Tian. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. European Respiratory Journal, 53(3), mar 2019. ISSN 13993003. doi: 10.1183/13993003.00986-2018.

[54] Zheng Wang, Hongshan Xu, and Meijun Sun. Deep Learning Based Nodule Detection from Pulmonary CT Images. In Proceedings - 2017 10th International Symposium on Computational Intelligence and Design, ISCID 2017, volume 2018-January, pages 370–373. Institute of Electrical and Electronics Engineers Inc., jan 2018. ISBN 9781538636749. doi: 10.1109/ISCID.2017.107.

[55] Xiu Shen Wei and Zhi Hua Zhou. An empirical study on image bag generators for multi-instance learning. Machine Learning, 105(2):155–198, nov 2016. ISSN 15730565. doi: 10.1007/s10994-016-5560-1.

[56] Stephen S.F. Yip, Ying Liu, Chintan Parmar, Qian Li, Shichang Liu, Fangyuan Qu, Zhaoxiang Ye, Robert J. Gillies, and Hugo J.W.L. Aerts. Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer. Scientific Reports, 7(1), dec 2017. ISSN 20452322. doi: 10.1038/s41598-017-02425-5.

[57] Wei Zhao, Jiancheng Yang, Bingbing Ni, Dexi Bi, Yingli Sun, Mengdi Xu, Xiaoxia Zhu, Cheng Li, Liang Jin, Pan Gao, Peijun Wang, Yanqing Hua, and Ming Li. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. Cancer Medicine, 8(7):3532–3543, jul 2019. ISSN 20457634. doi: 10.1002/cam4.2233.

[58] Zhi-Hua Zhou. Multi-Instance Learning: A Survey. Technical report.

[59] Zhi Hua Zhou and Jun Ming Xu. On the relation between multi-instance learning and semi-supervised learning. In ACM International Conference Proceeding Series, volume 227, pages 1167–1174, 2007. doi: 10.1145/1273496.1273643.

[60] Zhi Hua Zhou, Yu Yin Sun, and Yu Feng Li. Multi-instance learning by treating instances as non-I.I.D. samples. In ACM International Conference Proceeding Series, volume 382, pages 1–8, New York, New York, USA, 2009. ACM Press. ISBN 9781605585161. doi: 10.1145/1553374.1553534.

[61] Jiawei Zou, Tangfeng Lv, Suhua Zhu, Zhenfeng Lu, Qin Shen, Leilei Xia, Jie Wu, Yong Song, and Hongbing Liu. Computed tomography and clinical features associated with epidermal growth factor receptor mutation status in stage I/II lung adenocarcinoma. Thoracic Cancer, 8(3):260–270, may 2017. ISSN 17597714. doi: 10.1111/1759-7714.12436.