FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Data Preprocessing Strategies in Cancer Stage Prediction

Ana Maria Moreira

DISSERTAÇÃO

Mestrado em Engenharia e Ciência de Dados

Supervisor: Nuno Moniz

Co-Supervisor: Hélder Oliveira

September 23, 2022

Data Preprocessing Strategies in Cancer Stage Prediction

Ana Maria Moreira

Mestrado em Engenharia e Ciência de Dados

Approved in oral examination by the committee:

Chair: Carlos Soares
External Examiner: Rita Ribeiro
Supervisor: Nuno Moniz

September 23, 2022

Abstract

Cancer is one of the leading causes of death worldwide and a public health problem. If detected early and treated effectively, the chances of survival increase. The most common types of cancer are breast and lung. The data used in this master thesis is from the National Lung Screening Trial, which compares two ways of detecting lung cancer.

Medical imaging is an essential part of cancer diagnosis, and lung cancers exhibit substantial phenotypic differences that can be identified. The core of radiogenomics is extracting quantitative features from computed tomography (CT) images. These are inputs in a predictive model to classify mutation statuses for lung cancer patients.

The quality of medical data is often insufficient due to the difficulty of access to them and in terms of imbalanced class distributions in many real-world scenarios. Such imbalance often leads standard machine learning tools to focus their performance on the majority classes, while the minority class is the one associated with risky relevant cases. Imbalanced learning, the topic in machine learning that focuses on this type of scenario is one of its most challenging problems.

The most popular strategy in imbalanced learning is pre-processing methods, also known as resampling strategies. With the objective of maximising the utility of the available data for using it in predicting cancer status and characterisation, many data pre-processing strategies have appeared in the literature. This dissertation aims to identify these strategies and benchmark their impact in the context of cancer characterisation.

The conclusion reached is that SMOTE, although the most popular resampling technique used in the medical field, is not the best to use in the dataset used in this study. Also, we confirm that the best strategy must be selected based on the dataset and the learning algorithm used in each study.

Keywords: Imbalanced Learning, Resampling Techniques, Cancer Prediction, Lung Cancer

Resumo

O cancro é uma das principais causas de morte no mundo e um problema de saúde pública. Se for detetado precocemente e tratado de forma eficaz, as chances de sobrevivência aumentam. Os tipos de cancro mais comuns são da mama e do pulmão. Os dados utilizados nesta dissertação de mestrado são do *National Lung Screening Trial*, que compara duas formas de deteção do cancro de pulmão.

A imagem médica é uma parte essencial do diagnóstico do cancro, e os cancros de pulmão apresentam diferenças fenotípicas substanciais que podem ser identificadas. O essencial da radiogenómica é extrair características quantitativas de imagens de tomografia computadorizada (TC). Essas são entradas num modelo preditivo para classificar os estados de mutação para pacientes com cancro de pulmão.

A qualidade dos dados médicos é muitas vezes insuficiente devido à dificuldade de acesso aos mesmos e em termos de distribuições de classes desequilibradas em muitos cenários do mundo real. Esse desequilíbrio geralmente leva as ferramentas padrão de aprendizagem computacional a focar o seu desempenho nas classes majoritárias, enquanto a classe minoritária é aquela associada a casos relevantes de risco. Aprendizagem desequilibrada, o tópico em aprendizagem computacional que foca nesse tipo de cenário é um dos seus problemas mais desafiadores.

A estratégia mais popular na aprendizagem desequilibrada são os métodos de pré-processamento, também conhecidos como estratégias de reamostragem. Com o objetivo de maximizar a utilidade dos dados disponíveis para uso na previsão do estado e caracterização do cancro, muitas estratégias de pré-processamento de dados surgiram na literatura. Esta dissertação visa identificar estas estratégias e aferir o seu impacto no contexto da caracterização do cancro.

A conclusão a que se chega é que o SMOTE, embora seja a técnica de reamostragem mais popular utilizada na área médica, não é a melhor a ser utilizada no conjunto de dados utilizado neste estudo. Além disso, confirmamos que a melhor estratégia deve ser selecionada com base no conjunto de dados e no algoritmo de aprendizagem utilizado em cada estudo.

Keywords: Aprendizagem Desequilibrada, Técnicas de Reamostragem, Previsão de Cancro, Cancro de Pulmão

Acknowledgements

To my parents and brother for their support and patience over the years.

To all my friends for their patience in me not being present on many occasions and for always being present in my life.

To my colleagues who helped me achieve my academic goals with a good environment, teamwork and being good teachers when I needed them.

Finally, thank my supervisors and the team for their support throughout this process and their willingness to help me and give suggestions to improve my work.

Ana Maria Moreira

"One of the basic rules of the universe is that nothing is perfect.
Perfection simply doesn't exist...
Without imperfection, neither you nor I would exist."

Stephen Hawking

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a major public health problem that obstructs increasing life expectancy worldwide. According to World Health Organization (WHO), in 2019, cancer was the first or second leading cause of death before the age of 70 years in 112 of 183 countries (Figure 1.1). In 2020, an estimated 19.3 million new cancer cases and almost 10 million cancer deaths occurred worldwide. Female breast cancer has surpassed lung cancer as the most diagnosed cancer, and lung cancer remains the leading cause of cancer death [54].

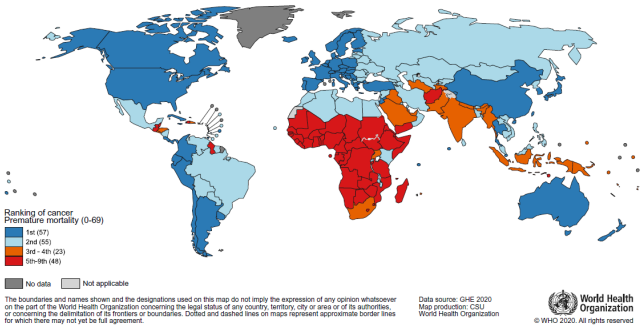

Figure 1.1: National Ranking of Cancer as a Cause of Death at Ages <70 Years in 2019 [54].

Risk factors may increase a person's chances of developing cancer. Some of these risks can be avoided, but others, such as growing older, aren't avoidable. Limiting exposure to the avoidable ones may lower the risk of developing certain types of cancer. The most-studied known risk factors for cancer are age, alcohol, cancer-causing substances, chronic inflammation, diet, hormones, immunosuppression, infectious agents, obesity, radiation, sunlight and tobacco [23].

The diagnosis and treatment of cancer were hampered by the COVID-19 pandemic. The restrictions and closures on health care resulted in delays in diagnosis and treatment that may lead to a short-term drop-in cancer incidence followed by an uptick in advanced stage disease and ultimately increased mortality [49].

Biopsies are the conventional way of collecting information about tumour genotyping by collecting tumour tissue and then characterising it using genomics approaches. This way of getting information is extremely invasive and painful for the patient, and the risk of complications increases by the repeated process. There are also some limitations with this process because some tumours are not homogeneous so the results may vary depending on which part of the tumour is biopsied. Medical imaging has the advantage of being less invasive, three-dimensional and provides information regarding the entire tumour. It is an essential part of cancer care and is crucial to cancer staging and diagnosis. The process that converts standard-of-care images into minable high-dimensional data is called radiomics. It can extract quantitative features from digital medical images that let us build models, link image features to the tumour's genomic profile, and identify alterations within tumour DNA, a field called Radiogenomics. So, radiogenomics has the capacity to identify the presence of relevant mutations [45, 58, 41].

1.1 Motivation

There is a lower chance of survival, higher difficulties linked with treatment and higher care expenses when cancer care is delayed or inaccessible. If effective approaches for identifying cancer earlier are developed, lives can be saved, and cancer care expenses are significantly reduced [38].

With patients being correctly diagnosed in the early stages, patients' quality of life can be improved, avoiding chemotherapy and radiotherapy. Even though these options are the spine of cancer treatment, their efficacies and applications are often vulnerable by their severe side effects, including cardiocytotoxicity, neurotoxicity, hepatotoxicity, alopecia, and others may further lead to late side effects in cancer patients [34].

Medical data is frequently hard to access, and the quality is an issue that needs an effort to improve. Privacy issues and the process of getting approval by ethical committees make medical data difficult to obtain. In many real medical scenarios, data with imbalanced class distributions are common. All of this leads us to work with a small number of cases/patients [31, 14].

1.2 Objectives

This dissertation aims to identify and benchmark data pre-processing strategies developed to deal with imbalanced data that could lead to improvements in the area and maximise the utility of the available data for using it in the prediction of cancer status and characterisation.

There is a significant gap between the current practice in image recognition applications within the scope of cancer prediction, using well-known data augmentation methods, and the methods available in areas such as imbalanced learning. This corresponds to a great opportunity to study and extend existing knowledge concerning the potential of state-of-the-art data pre-processing strategies in the context of this application.

1.3 Contributions

The contribution of this work is the following:

- Identify pre-processing strategies that can significantly improve the predictive performance of popular machine learning algorithms when confronted with imbalanced data in the context of radiomics data;
- Benchmark their performance and assess if there are any strategies that consistently perform better than others.

1.4 Structure

This document is organised into five chapters. The present chapter briefly introduces this dissertation's motivation, objectives, and expected contributions. Chapter 2 presents some concepts and contextualises the problem, including a description of cancer, more specifically about lung cancer, diagnosis, biopsies, mutations, target therapies, radiomics, and radiogenomics. Chapter 3 states the problems with imbalanced domains and describes the state-of-the-art methodologies that could be applied to help deal with imbalanced learning. Chapter 4 describes the dataset and explains the pre-processing steps taken to use it and improve the results of the models in the first section. The remaining sections describe the experimental work carried out on the obtained dataset and the different models that were tested. It also compares the results obtained for each model and explains the multiple parameters used. Chapter 5 provides a conclusion regarding the study completed in this dissertation and some future work considerations.

Chapter 2

Background

This chapter presents some concepts and contextualization of the problem. The first section gives an overview of cancer and then specifies the various types of lung cancer. Sections 2.2 and 2.3 analyse the diagnostic methods and their advantages and disadvantages. The mutated genes are addressed in Section 2.4 and the target therapies are introduced in Section 2.5. Finally, the topics of radiomics and radiogenomics are introduced in Section 2.6.

2.1 Cancer

The cells of our body are constantly growing, getting divided as they need and dying when they are old or abnormal. When this process is not going well and the abnormal and old cells keep dividing themselves and do not die, cancer starts. The name of the cancer is where it started to develop in the body, and if it metastasises to the other parts of the body, the name is still the same and doesn't change [61, 51].

There are two main types of cancer, hematologic, cancers of the blood cells (e.g. leukaemia, lymphoma, and multiple myeloma) and solid tumour, cancers of any of the body organs or tissues (e.g. breast, prostate, and lung cancers) [51].

Tumours are lumps or growths that can be cancer or not, in many cases they are not. The ones called benign are not cancer and cannot spread to other parts of the body, on the other hand, the ones which are cancer are called malignant and can spread all over the body [51, 61].

The stage of cancer can be characterized by the size of the cancer and if it is spread from the initial site. This is truly important to choose the better and adequate treatment for the patient. A lower stage (1 or 2) means that cancer has not spread too much or does not spread at all and a higher stage (3 or 4) means that it is spread to surrounding tissues or to other organs, to highlight that the highest is the stage 4 [47, 51].

2.1.1 Lung Cancer

The lungs, located in the chest, are sponge-like organs divided into sections, called lobes. The right lung has three lobes and is slightly bigger than the left lung, which has two lobes. When inhaled, the air is pulled through the trachea, the bronchi, and bronchioles. At the end of the bronchioles are the alveoli, tiny air sacs (Figure 2.1). Typically, lung cancer starts at the cells lining the bronchi, bronchioles or alveoli [22, 50].

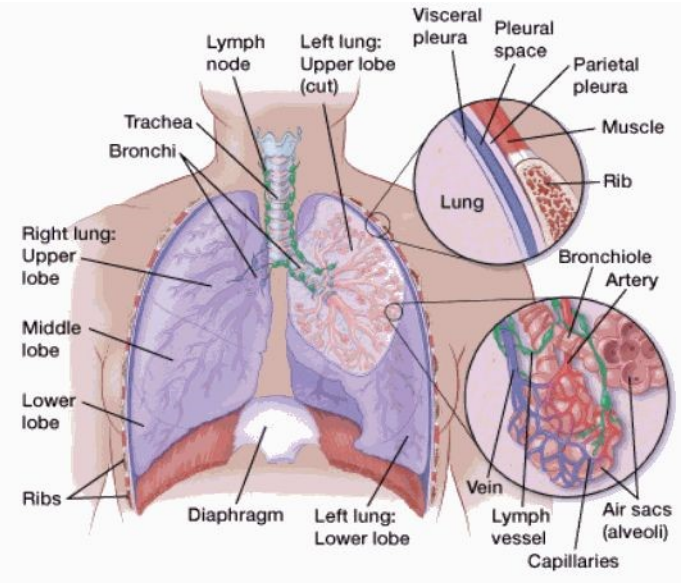

Figure 2.1: Lung Anatomy [50].

There are two main types of lung cancer [50, 49]:

- **Non-small cell lung cancer (NSCLC)** - about 80-85% of lung cancers, grouped together because of the similar treatment and prognoses although they are from different types of lung cells.

  The main subtypes of NSCLC are:

  - Adenocarcinoma: start in cells that normally secrete substances such as mucus;
  - Squamous cell carcinoma: start in squamous cells which are flat cells that line the inside of the airways in the lungs;
  - Large cell carcinoma: can appear in any part of the lung, tends to grow and spread quickly;
  - Other subtypes: are less common as adenosquamous carcinoma and sarcomatoid carcinoma.

- **Small cell lung cancer (SCLC)** - about 10-15% of lung cancers, tend to grow and spread faster than NSCLC.

- **Other types of lung tumours**
  - Lung carcinoid tumours: fewer than 5% of lung tumours and they tend to grow slowly;
  - Other lung tumours: adenoid cystic carcinomas, lymphomas and sarcomas;
  - Cancers that spread to the lungs: start in other organs and can sometimes metastasize to the lungs.

Lung cancer stands for nearly one-quarter of all cancer deaths, is caused directly in 82% of the cases by cigarette smoking, which translates, in 2021, to approximately 107 870 smoking-attributable lung cancer deaths and 3590 due to second-hand exposure. The survival rate is lowest for cancers like the pancreas, liver, oesophagus, and lung. About 57% of patients are diagnosed with a metastatic disease which reflects the low lung cancer survival rate. Quitting smoking after being diagnosed with lung cancer is strongly associated with substantial improvement in general survival and disease-free survival [49, 3].

2.2 Diagnosis

The objective of screening exams is to detect disease before the symptoms begin, at its earliest and most treatable stage. Screening tests might include lab tests, genetic tests, and imaging exams. Typically, they are available to all the population, however, an individual's requirements for a specific screening test are based on factors such as age, gender, and family history. Imaging exams are fundamental not only in the diagnosis but also in staging, treatment planning, postoperative surveillance, and response evaluation in the routine management of lung cancer [37, 32].

Contrasting with biopsies, it is less invasive and provides information regarding the entire tumour. Lung cancers exhibit strong phenotypic differences that can be identified by medical imaging. In lung cancer screening, typically is used low dose CT scanning, while diagnostic images are more frequently high quality and with contrast enhancement [45, 37, 58].

2.3 Biopsy

The biopsy is a medical procedure that involves taking a small sample of body tissue to be examined at the microscope to identify abnormal cells. This procedure can help diagnose a specific condition or, if the condition has already been diagnosed, the biopsy can be used to assess the severity and grade of that. For deciding on the most suitable treatment, evaluating how well a patient will respond to a particular treatment, and helping determine a patient's overall prognosis, this information can be very useful [46].

The biopsy can be sometimes inconclusive and may need to be repeated or other tests may be required to confirm the diagnosis. The tissue is often obtained from a portion of a heterogeneous tumour and may not be enough or representative. To fill this lack, multiple or sequential biopsies are not a solution because of logistical and financial barriers. Biopsies are also very intrusive and

can be painful for the patient, sometimes the location of the tumour can complicate the access and the collection of the tissue or, in some cases, make it impossible [44, 46].

2.4 Mutations

A change in the sequence of an organism's genome is called a mutation. Many different types of changes in sequences can result in mutations, they can vary from single-base pair alterations to megabase pair deletions, insertions, duplications, and inversions. In normal and abnormal biological processes, such as evolution, cancer, and the development of the immune system, mutations play an active role. Mutations are classified as somatic or germline mutations: the former are genomic alterations that are not transmitted to offspring, unlike the latter that are therefore heritable [18, 4].

The most frequently mutated genes in lung cancer are Epidermal Growth Factor Receptor (EGFR) and Kristen Rat Sarcoma Viral Oncogene Homolog (KRAS). These mutations are somatic oncogenic and 15 to 50% of NSCLC patients from never-smokers have the mutated EGFR present. For another hand, KRAS is related to smoking patients, with only 5 to 10% of KRAS-mutant lung cancers arising in non-smokers [41, 48].

The permanent activation of EGFR is promoted by its mutations who contribute to uncontrolled cell division since is responsible for cell growth and survival. With clinically approved treatments, this biomarker is now considered a robust prognostic pointer in lung cancer, upgrading chances of exploring treatment approaches that rely on the individual's genetic profile [19, 42, 25]. Unlike EGFR, KRAS has no inhibitors as an approved therapy and due to the biochemistry complexity has shown to be more difficult to target [59, 15]. In a traditional way, the tissues extracted by the biopsy, are molecularly tested and the oncogene mutation status is assessed. Recently, have been developed more automatic and less invasive techniques, like computer-aided diagnosis (CAD) based on CT analysis, this results in a decreasing risk to the patients and improves the accuracy of the diagnosis [17, 10].

2.5 Target Therapies

As researchers learn more about the DNA mutations and proteins that lead to cancer, they are better at designing treatments that target these proteins. Target therapies are a type of cancer treatment that targets proteins that control how cancer cells grow, divide, and spread. Most of the target therapies are small-molecule drugs or monoclonal antibodies. Testing cancer for targets that could help the patient and the doctor to choose the most adequate treatment is called biomarker testing. Concerning targeted therapy, tumour mutations have consistent predictive value and, in fact, guide the medical decision of treatment [24, 41].

2.6 Summary

Radiomics is a field of study that aims to the extraction of quantitative features from medical images. In the context of lung cancer, there has been considerable interest in the use of radiomics, with the objective of, though minimizing the time weight imposed upon radiologists, maximizing sensitivity and specificity [32, 58].

Aside from diagnosis, in the field of precision medicine, radiomics is also being used to predict prognosis and response to certain treatments. Some extracted features have been shown to identify genomic alterations within tumour DNA, this field is now called Radiogenomics. The presence of specific mutations and alterations can be identified by these features. So, radiogenomics is the capacity of radiomics to identify the presence or absence of clinically relevant mutations [58, 41].

The process workflow of radiomics or radiogenomics includes image acquisition and reconstruction, region of interest (ROI) segmentation, feature extraction and quantification, and building predictive prognostic models [58, 30]. This workflow is represented in the Figure 2.2.

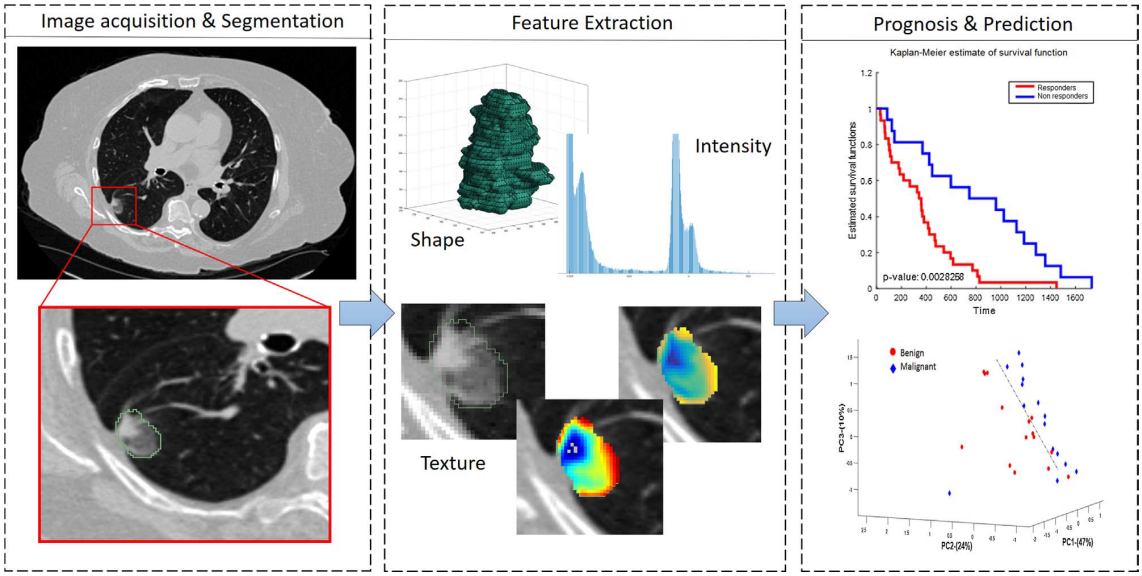

Figure 2.2: Overview of the Radiomic/Radiogenomic process workflow [58].

Chapter 3

Literature Review

This chapter is divided into two main subjects. The first part states the problems with imbalanced domains (Section 3.1) and the second compiles some relevant studies about strategies to deal with imbalanced learning (Section 3.2).

3.1 Imbalanced Domains

A classification data set with skewed class proportions is considered imbalanced. The classes that have a large proportion of the data set are called majority classes and those that have a smaller proportion are minority classes (Figure 3.1) [13, 28].

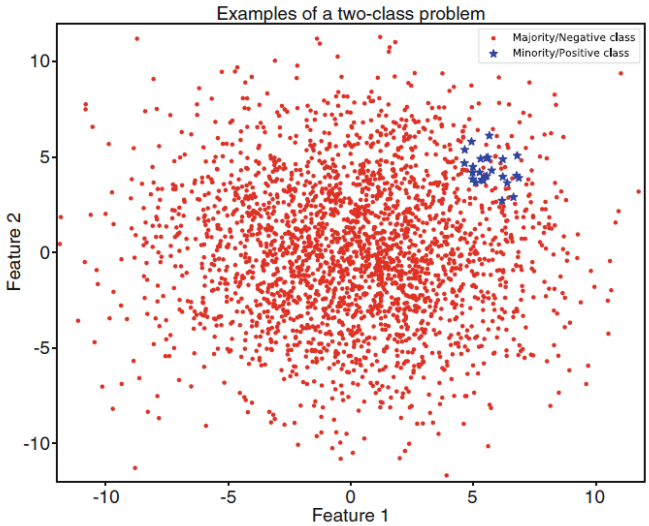

Figure 3.1: Example of a two-class imbalanced problem with ratio 1:100 [16].

In real-world predictive analytics, imbalanced domains are an important problem frequently arising. They are characterized by higher relevance being assigned to the performance on a subset of the target variable values and these most relevant values being underrepresented on the available data set [7]. Fundamentally, the performance of a classifier tends to focus on the performance of the majority classes in the imbalanced data set. However, the values of the minority classes are frequently associated with events that are highly relevant. These events might have different costs and benefits, that, when allied with the fewness of some of them on the available data, create serious difficulties for predictive modelling techniques [35, 28, 8]. This topic assumes importance in circumstances such as when the minority class affects the detection of rare cancer cells and an incorrect diagnosis may prove to be fatal [57].

Many solutions have appeared in the literature to respond to the inadequacy of the obtained models of imbalanced domains. Nevertheless, imbalanced learning faces two decisive challenges. The first one, the quantity of approaches proposed to deal with it. Has grown hugely, thus, the validation of a large set of methods is impractical. The second one, requires specialised knowledge, thus, its correct use by those without such level of experience is difficult [8, 35].

3.2 Strategies for handling imbalanced domains

The existing approaches to learn under imbalanced domains are grouped into four main categories (Figure 3.2) [8]:

- **Data Pre-processing** - changes on the data before the learning process takes place;
- **Special-purpose Learning Methods** - adjustments on the learning algorithms;
- **Prediction Post-processing** - alterations applied to the predictions of the learned models;
- **Hybrid Methods** - combine different types of strategies.

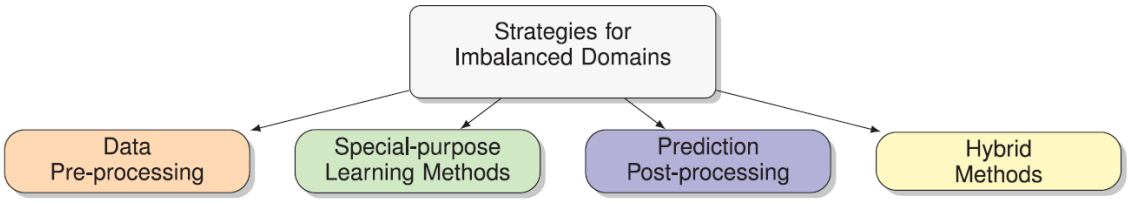

Figure 3.2: Main strategies for handling imbalanced domains [8].

3.2.1 Data Pre-processing

Data pre-processing methods, also known as resampling, are the most studied solution category. These work by changing the data's distribution via under or over sampling strategies, trying to clarify the decision limits between well and poorly represented classes [35].

Data pre-processing methods are applied in two different ways. The first way, **distribution change**, imbalanced data sets modified before learning, providing more balanced distributions, following user preferences, focusing on the cases that are more relevant to the user. The second one, **weighting the data space**, modifying the training set distribution using information regarding misclassification costs to avoid costly errors [35, 8].

The pre-processing methods have the advantage of being independent from the learning algorithm used. The selected models are biased to the user goals since data distribution was earlier changed to match these goals, thus, is expected that the models are more interpretable in terms of these goals. The main disadvantage of this methods is that it might be tough to relate the alterations in the data distributions with the information provided by the user, regarding the preference biases [35, 8].

Distribution change has three main approaches (Table 3.1): stratified sampling, synthesising new data, and combinations of the previous. In stratified sampling, the most common methods are under and oversampling [57, 35].

Table 3.1: Distribution Change Approaches

Stratified Sampling	Undersampling
	Oversampling
	Hybrid sampling
Synthesizing New Data	
Combination of Methods	

Sampling Methods are an easy and popular approach to balance the class distributions of the training data, they reduce learning time and faster execution once we get the balanced classes. Sampling techniques are better for imbalanced data using a decision tree where bagging and boosting do not improve decision tree performance [28].

A natural choice for answering the imbalanced class problem is resampling, and, at the same time, inter-class and intra-class diversity in the training data must be maintained. That throws a challenge regarding the need for careful and intelligent pruning of imbalanced datasets such that the representation from all classes is balanced. The cautious selection of a subset of samples that transmit class-discriminatory information would generate a balanced distribution that accomplishes a better learning performance [57].

Undersampling: Extracts a random set of samples from the majority class in order to balance the classes and the remaining samples are ignored. This method helps balance the class samples and makes the training phase faster. In random undersampling (RUS), since no intelligent technique or reasoning was applied for the selection of majority samples, the loss of information in the form of useful majority samples that were not included is higher [57, 28].

Undersampling reduces the number of examples considered normal, from the majority population to match the minority, and oversampling replicates examples of the under-represented cases, from the minority population to match majority. Both methods have disadvantages, undersampling has the risk of eliminating significant cases and oversampling the risk of overfitting due to

the replication of certain cases [57, 35].

Another tool to cut off the majority population is clustering. For the balance ensues, the number of clusters of the majority class is set equal to the minority population. Replacing the cluster centres with the nearest neighbours is a good option and is considered to be a more viable approach [33, 57].

Yu et al. [62] proposed heuristics undersampling method, for example, ACOSampling, while under-sampling does not delete the samples with useful information, rather it automatically extracts and saves them. An ensemble of EUSBoost was proposed by Krawczyk et al. [29] that contains the boosting idea for undersample evolution for each of the base classifiers. Level set active contours method yields an effective features extraction for better classification of the cancer symptoms for a clinical decision support system. Kang et al. [27] proposed another method that focuses on removing the noisy minority class samples, which difficult the performance in imbalanced data classification. A data gravitation classification model was proposed by Peng et al. [39], which is efficient for supervised learning methods to handle the issue of imbalanced data by undersampling, which shows an effective margin in sensitivity and specificity in results [28].

One of the important works in undersampling is that of Tomek [60], who defined Tomek links as a pair of samples belonging to two different classes whose distance is minimum. There are two ways to interpret this: both are doubtful examples or one of the samples is noise. The solution would be either to remove both samples or remove the sample that fits in the majority class. Tomek links are still used in several modern works in literature to balance their population [57].

Oversampling: Oversampling idea is to increase the size of the minority class to acquire balanced classes. Random oversampling selects the minority samples to be replicated in a randomized manner, some researchers prefer this method instead of the opposite since undersampling may lead to the loss of some important information [28, 57].

Being the most popular approach the minority oversampling technique (SMOTE), the synthetic samples are produced with the support of minority class samples. Some variants of SMOTE, for example, Borderline-SMOTE and Safe-Level-SMOTE, expand upon the original algorithm by also taking majority class neighbours into consideration. Borderline-SMOTE bounds oversampling to the sample's close class limits, although Safe-Level-SMOTE outlines safe areas to avoid over-sampling in overlapping or noise areas. The random walk oversampling approach generates samples to increment the number of samples in the minority class. The oversampling method for imbalanced text classification where to each minority class documented in the training class, a probabilistic function is assigned. Sigma Nearest Oversampling based on Convex Combination (SNOCC), which generates new samples and guarantees that generated new samples find the new nearest neighbours, can overcome the limitations of SMOTE [26, 28, 63].

SMOTE with a data cleaning method, Tomek links, was proposed to create better-defined class clusters. Thus, examples from both classes are removed instead of removing only the majority class examples that form Tomek links. Stefanowski proposed a new approach to selective preprocessing of imbalanced data, combining local over-sampling of the minority class with complex filtering examples from the majority classes. Cluster-SMOTE is a method that may be used as an

improvement over SMOTE, which applies unsupervised learning to partition datasets into regions that will enable SMOTE to deliver enhanced results [6, 53, 11].

The combination of sampling methods can stand as a good choice for improving the classifier performance dealing with imbalanced classes distributions [28].

Hybrid sampling: Methods that apply both re-sampling techniques to achieve balance in the data are called hybrid sampling. The techniques proposed are combining sampling methods undersampling and oversampling to handle the problem of imbalanced data. To create a balanced training data space, undersampling to remove the instances without comprising useful information, then oversampling is done to replicate existing instances. Thus, the proposed method decreases the chances of losing informative instances. Adding a vast number of synthetic samples to the training space leads to increasing classification performance [28].

Most of the hybrid strategies use SMOTE as the oversampling technique. SMOTE-PSO was proposed by Hu et al. [21] and applied successfully for identifying malicious web domains and the undersampling part is by PSO. Susan et al. [56] proposed SMOTE-SSO which sequentially performs oversampling of the minority class and undersampling of the majority class. SMOTE-RSB proposed by Ramentol et al. [43] combines SMOTE with an editing method based on the Rough Set Theory. The synthetic samples generated by SMOTE were evaluated for their comparison with the majority class samples. Those synthetic samples whose similarity index was high were removed from the training set since they do not contribute to class-discriminatory information. SMOTE was combined with data gravitation-based classification (DGC) in Peng et al. [40], resulting in SMOTE-DGC, a physical-inspired classification model that fails under conditions of class imbalance. The method SSOMaj-SMOTE-SSOMin proposed by Susan et al. [55] contains cautious selection of the population of both classes, and only representative samples that carry discriminatory information are retained on both sides. An evolutionary algorithm (PSO) is used for the choice. The three-step data pruning process in SSOMaj-SMOTE-SSOMin constitutes undersampling the majority, oversampling the minority, and undersampling the oversampled minority. The results manifest an advance over the baseline undersampling and oversampling techniques [57].

All the methods mentioned above that involve hybridization of SMOTE with intelligent undersampling, help careful selection of both the majority and minority population, which is an improved method than careful selection of the majority samples alone [57].

One side selection method in which Tomek Links are used to decline the noisy and unreliable examples from the majority class, thus, it undersamples the majority class in an efficient mode. Then CNN is used to remove the samples that are distant from the decision borderline, thus, it keeps useful samples while undersampling the majority class [28].

The adaptive synthetic sampling approach (ADASYN) is proposed for transferring different weights to samples giving their level of difficulty while learning and synthetic data are generated. It demonstrates to stand an effective technique of handling imbalanced data in the succeeding ways: tends to decrease the condition of imbalance data where hyperplane continuously gets biased near the majority class and makes the classification hyperplane in an efficient way that it

automatically rests in the track of instances that are hard to learn [28].

Song et al. [52] proposed a study using oral cancer image data, applying data-level and algorithm-level approaches to the deep learning training process to improve the performance of the minority classes that were difficult to distinguish at the beginning. They combined both oversampling and undersampling and saw that the classifier's performance on the minority class improved in comparison with the use of data augmentation alone.

Naseriparsa et al. [36] proposed a study using a lung cancer dataset that used the combination of Principal component analysis (PCA) with SMOTE resampling method. SMOTE resampling method is applied to the dataset, and the accuracy in this condition has been increased. This increase is due to the use of SMOTE resampling right after the running of PCA. In this condition, SMOTE contributes to expanding the variety of sample domains and compensating for the loss of some information that occurred in applying PCA.

Corso et al. [12] presented a study with data that were lung cancer CT images, and since they noticed an imbalanced proportion of classes, they applied SMOTE technique. They didn't use or benchmark any other different strategy. Hasan et al. [20] published a study with data from Mammographic Images and as expected in this field, have imbalanced classes. To deal with it, they used SMOTE because of its popularity in dealing with imbalanced datasets and didn't try other techniques.

More recently, Aruna et al. [5] proposed a study that analyses the impact of resampling on supervised learning algorithms in identifying the types of lung cancer. This study was more complete because used five resampling methods: Random Oversampler, Adasyn, SMOTE, Svm-SMOTE and KMeans SMOTE. They conclude that Svm-SMOTE was the best resampling technique for the classifiers used in the study.

3.3 Summary

In this section, we discuss various approaches for dealing with imbalanced domains. Overall, the SMOTE strategy is the most used and some hybrid variations of the original SMOTE are highly used too. In a recent study [5], we can see the use of new techniques, but SMOTE is always present. Some surveyed works have shown that decreasing class imbalance in the training data with ROS meaningfully improves classification results. A study [9] found that plain ROS and RUS usually perform better than two-phase learning. Some differences in performance metrics and problem difficulty make it tough to relate approaches directly [26].

Chapter 4

Experimental Study

This chapter presents a detailed description of the dataset used in the model implementation with an explanation of some essential pre-processing steps necessary to improve model implementation and the final results (Section 4.1). Also describes the experimental work carried out on the obtained dataset (Section 4.2). Several different models were tested and compared for multiple grids of parameters specific to each model in question. Resampling strategies were applied to determine the best approaches to our problem.

4.1 Data Description

4.1.1 Dataset

The National Lung Screening Trial (NLST) was conducted by the American College of Radiology Imaging Network, a medical imaging research network focused on leading multicenter imaging medical trials, and the Lung Screening Study group, which the National Cancer Institute (NCI) established initially to inspect the viability of NLST [2].

The NLST compares two ways of detecting lung cancer, low-dose helical computed tomography (CT) and standard chest X-ray, on current or former heavy smokers aged 55 to 74. If they are former smokers had to quit within the previous 15 years. Participants were required to have a smoking history of 30 pack-years minimum, calculated by multiplying the average quantity of packs of cigarettes smoked per day by the number of years the participant has smoked. Each participant was randomly assigned to receive three annual screens using either low-dose helical CT or chest X-ray [1, 2].

NLST enrolled 53454 patients from August 2002 to April 2004, of which 26722 were assigned to low-dose CT and 26732 to screening with chest radiography. The screening took place from August 2002 through September 2007. This trial excluded persons who had earlier received a lung cancer diagnosis, had undergone chest CT 18 months before enrolment, had haemoptysis or had an inexplicable weight loss of more than 6.8 kg in the preceding year. Participants were requested

to submit to three screenings at 1-year intervals, with the first screening completed shortly after randomisation. Patients diagnosed with lung cancer were not offered following screening tests [1, 2].

4.1.2 Data Preparation

The dataset needs to be prepared to be used to train and test the model that will be developed. Only the clinical data will be used.

The first step was selecting only the participants eligible for the study on the dataset, then selecting only the CT participants and removing the patients with confirmed cancer diagnostic in study years more than two because these patients don't have imaging data. From the target variable, the status of lung cancer report, the objective was only to have confirmed lung cancer and not lung cancer. For that, the other values were removed.

The imputation with the average value was done for the missing values in numeric columns like height and weight. The imputation with mode was used instead for categorical and binary columns. Then, the variables were converted to binary for the categorical types with one-hot encoding.

The columns not needed for this study were removed, resulting in a dataset with mostly binary and some numeric columns. To deal with the numeric columns, a min-max scaler was used to normalise these data.

The obtained data set is summarised in the table below (Table 4.1).

Table 4.1: Dataset summary

Number of Cases	5985
Number of Features	87
Gender	Male (3603), Female (2382)
Cancer Report	Non Cancer (5275), Cancer (710)

Regarding the target variable, Figure 4.1 shows the distribution of the cancer report in the data set, where the percentage of non-cancer is 88,14%, clearly noticing the imbalance problem.

4.2 Implementation

All the experimental work was developed in python programming language version 3. The tested algorithms included Logistic Regression, Random Forest and XGBoost. The libraries used for exploring the resampling strategies were imbalanced-learn and smote_variants, and a representative batch of resampling strategies was selected to be evaluated. The chosen to be used were the following methods: Random Undersampling (RUS), Random Oversampling (ROS), SMOTE, Borderline Smote, Adasyn, Cluster SMOTE, Stefanowski, Safe Level SMOTE, SMOTE Tomek-links and SMOTE PSO.

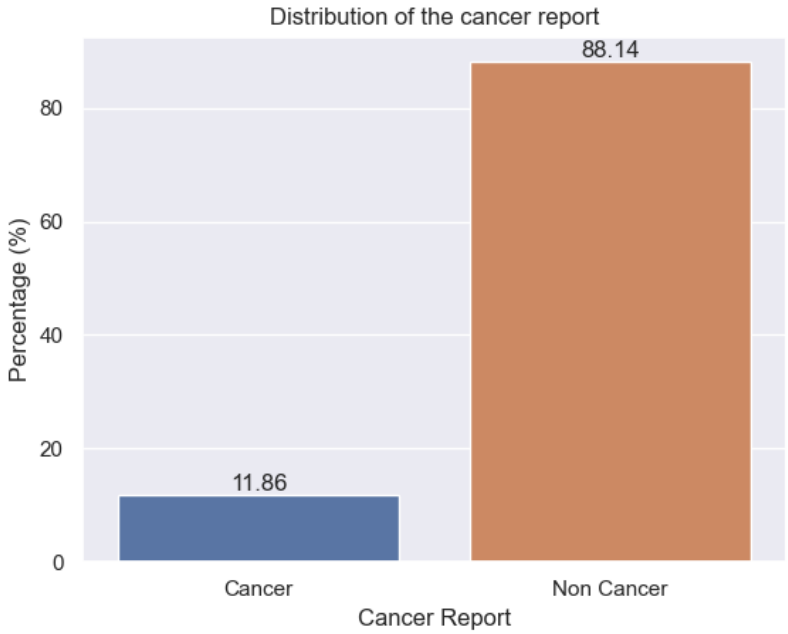

Figure 4.1: Distribution of the cancer report

Figure 4.2 represents the pipeline for models consisting of two phases: an optimisation phase where we optimise the model using grid-search, using 5-fold cross-validation (CV), and the classification phase, where the train/test process is repeated five times, randomising the split seed, for model validation. A split ratio of 80/20 % was used to train and test sets, respectively.

Each algorithm step is modular to facilitate using other learning algorithms or resampling strategies. So that in the future, as new solutions emerge, replacing one machine learning model or resampling strategy with another is more straightforward.

4.3 Hyperparameter Optimisation

To improve the performances of each model, the hyperparameters were tuned using Grid Search CV on the training data. Grid search CV is an approach that methodically builds and evaluates a model for each combination of algorithm parameters specified in a grid. Tables A.1, A.2 and A.3 present for Random Forest, Logistic Regression and XGBoost, respectively, the hyperparameters, optimal values and the search space used by the models for that classification problem.

For Random Forest, the hyperparameters are Max Depth (the maximum depth of the tree), Min Samples Split (the minimum number of samples required to split an internal node) and N estimators (the number of trees in the forest). For Logistic Regression, the hyperparameters are Penalty (specify the norm of the penalty: 'none' - no penalty is added; 'l2' - add an L2 penalty term and it is the default choice; 'l1' - add an L1 penalty term; 'elasticnet' - both L1 and L2 penalty terms are added), Solver (algorithm to use in the optimisation problem) and Max Iter (maximum

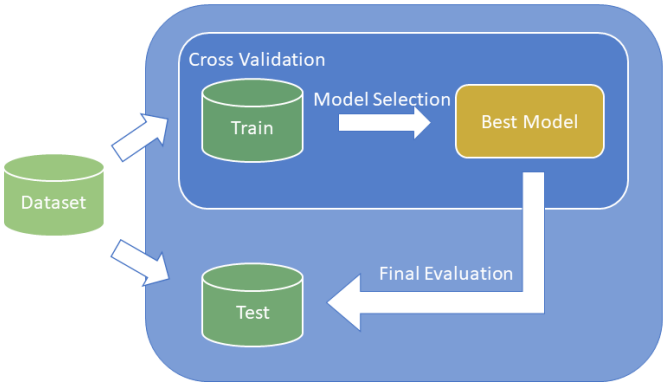

Figure 4.2: Implementation Pipeline

number of iterations taken for the solvers to converge). For XGBoost, the hyperparameters are Max Depth (the maximum depth of the tree) and N estimators (the number of trees in the forest). For the resampling strategies, the hyperparameters depend on the strategies but in most cases is the sampling information to resample the data set that corresponds to the desired ratio of the number of samples in the minority class over the number of samples in the majority class after resampling. On SMOTE, also have the k neighbors (number of nearest neighbours used to construct synthetic samples). On Stefanowski the strategy that can be 'weak_amp', 'weak_amp_relabel' or 'strong_amp'. And on SMOTE PSO, k (number of neighbors in nearest neighbors component), n pop (size of population) and num it (number of iterations).

4.4 Evaluation Metrics

Choosing the evaluation metric very carefully when comparing models and their results is crucial. This can affect the conclusions taken. In this work, three metrics were selected because they are resistant when dealing with imbalanced datasets. The metrics are:

- F1-Score - combines the precision and recall of a classifier into a single metric by taking their harmonic mean:

$$F1 = \frac{2(Precision * Recall)}{Precision + Recall} \tag{4.1}$$

- Area Under ROC Curve (AUC) - Receiver Operating Characteristic (ROC) is a probability curve, and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes;

- Precision Recall AUC (PR AUC) - summarises the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.

4.5 Results and discussion

Table 4.2 presents the mean values of the results for the metrics used and the standard deviation (mean ± standard deviation).

Analysing Table 4.2, the results achieved for the different learning algorithms suggest the use of varying resampling strategies for each one, and it's unanimous the use of resampling strategies instead of not using. All the values have a low standard deviation.

Given Random Forest, the resampling strategies which obtain better results are Random Undersampling and Random Oversampling, with the mean ROC AUC, F1 score, and PR AUC being superior to the values for other strategies. Given Logistic Regression, the Random Oversampling, Stefanowski, and Borderline SMOTE have the best results. Given XGBoost, the best results are obtained by Random Undersampling, and the other strategies get a low F1 score value.

In Figure 4.3, we have four calibration plots for the best results of logistic regression implementation, the first without Resampling Strategy, the second with Random Oversampling, the third with Stefanowski and the last with Borderline SMOTE. The plots show that the results are far from the ideal line without a resampling strategy in some cases. With the resampling strategies, the results are closest in general, and all three resampling strategies have similar results.

The choice of a resampling strategy depends on the algorithm used and the available data for the study. In this study, we can see that for the same dataset, the use of a different learning algorithm can change the resampling strategy that gives the best results for the model. Also, the dataset influences that, so the approach used needs to be chosen based on benchmarking with different techniques that can be the best for our problem.

In the medical field, SMOTE is used as the principal technique to deal with imbalanced datasets, as we can state in the studies of Song et al. [52], Naseriparsa et al. [36], Corso et al. [12] and Hasan et al. [20]. Usually, they don't apply some techniques and benchmark them to see the most adequate to their data. Although SMOTE is the most popular strategy used in this field, in our study it is not the best for this dataset and for any of these learning algorithms and loses its place to other techniques. In fact, SMOTE and SMOTE PSO were the resampling strategies that take more time to compute, and the results are bad compared to the best results.

Other techniques can be studied and have better results for our data than the chosen ones. Still, it is impossible to explore all the available techniques in a short time because of the computational complexity involved and the limited time for this study. Even though, a good batch of techniques was implemented and we can state the best ones for each algorithm.

4.6 Summary

In this Chapter, different experiments were conducted regarding the class imbalance problem in the cancer classification problem. A batch of resampling strategies was selected to be used, and the results were compared to achieve the top for each one of the algorithms. One important conclusion from here is that using resampling strategies improves the model performance.

Table 4.2: Classification Results

	Random Forest			**Logistic Regression**			**XGBoost**		
Resampling Strategies	**ROC AUC**	**F1-Score**	**PR AUC**	**ROC AUC**	**F1-Score**	**PR AUC**	**ROC AUC**	**F1-Score**	**PR AUC**
No Resampling Strategy	0.700 ± 0.011	0.006 ± 0.008	0.271 ± 0.012	0.718 ± 0.023	0.068 ± 0.024	0.284 ± 0.014	0.635 ± 0.016	0.145 ± 0.022	0.206 ± 0.011
Random Undersampler	$\mathbf{0.722 \pm 0.017}$	0.334 ± 0.021	$\mathbf{0.302 \pm 0.024}$	0.715 ± 0.022	0.324 ± 0.022	0.267 ± 0.010	$\mathbf{0.657 \pm 0.014}$	$\mathbf{0.294 \pm 0.023}$	0.232 ± 0.018
Random Oversampler	0.716 ± 0.019	$\mathbf{0.344 \pm 0.022}$	0.290 ± 0.041	$\mathbf{0.723 \pm 0.020}$	0.345 ± 0.011	$\mathbf{0.294 \pm 0.009}$	0.642 ± 0.021	0.262 ± 0.014	0.217 ± 0.029
SMOTE	0.705 ± 0.022	0.318 ± 0.021	0.258 ± 0.028	0.709 ± 0.028	0.330 ± 0.020	0.279 ± 0.019	0.648 ± 0.019	0.143 ± 0.024	0.215 ± 0.011
Borderline SMOTE	0.709 ± 0.024	0.339 ± 0.014	0.267 ± 0.024	0.712 ± 0.026	$\mathbf{0.351 \pm 0.030}$	0.275 ± 0.019	0.644 ± 0.021	0.145 ± 0.035	0.216 ± 0.011
ADASYN	0.705 ± 0.025	0.328 ± 0.025	0.259 ± 0.032	0.704 ± 0.025	0.325 ± 0.018	0.279 ± 0.017	0.651 ± 0.016	0.143 ± 0.041	0.220 ± 0.013
Cluster SMOTE	0.699 ± 0.023	0.309 ± 0.029	0.255 ± 0.027	0.705 ± 0.028	0.331 ± 0.023	0.278 ± 0.022	0.654 ± 0.012	0.170 ± 0.039	$\mathbf{0.235 \pm 0.018}$
Stefanowski	0.717 ± 0.018	0.300 ± 0.024	0.298 ± 0.027	0.721 ± 0.020	0.347 ± 0.017	0.289 ± 0.015	0.645 ± 0.022	0.257 ± 0.021	0.216 ± 0.018
Safe Level Smote	0.688 ± 0.027	0.171 ± 0.037	0.237 ± 0.018	0.657 ± 0.041	0.285 ± 0.052	0.228 ± 0.035	0.638 ± 0.016	0.160 ± 0.020	0.208 ± 0.014
SMOTE Tomeklinks	0.706 ± 0.023	0.337 ± 0.024	0.254 ± 0.028	0.710 ± 0.027	0.332 ± 0.019	0.287 ± 0.019	0.641 ± 0.019	0.153 ± 0.024	0.222 ± 0.008
SMOTE PSO	0.695 ± 0.015	0.153 ± 0.014	0.229 ± 0.019	0.681 ± 0.043	0.296 ± 0.025	0.257 ± 0.030	0.647 ± 0.023	0.158 ± 0.018	0.221 ± 0.015

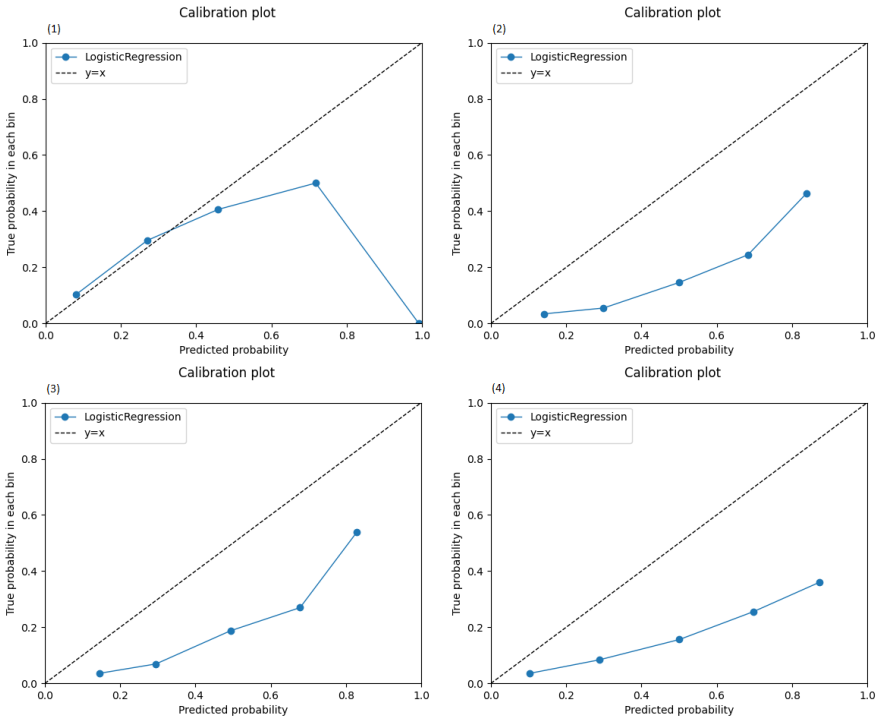

Figure 4.3: Calibration Plots: (1) Without Resampling Strategy; (2) Random Oversampling; (3) Stefanowski; (4) Borderline SMOTE.

Chapter 5

Conclusions

The main objective of this dissertation was to identify data pre-processing strategies to deal with imbalanced data and could lead to improvements and maximise the utility of the available data since in the medical field, information is hard to access, and imbalanced class distributions are widespread. This leads to working with a small number of cases/patients and may affect the results of the studies.

A benchmark and comparison between a batch of resampling strategies were conducted from the main objectives defined at the start of this work. The positive results allowed us to assess better strategies for the learning algorithms used in the study. Overall, the balance is positive since the main objectives proposed were met.

From the study, it was possible to underline some key aspects to be investigated further, such as the need to use more learning algorithms to know which resampling strategies were better for them. Using features from medical imaging would be good research to help in this field. The use of feature engineering in this problem might be valuable since little research was done, and the results look promising.

Appendix A

Hyperparameter Optimisation

Table A.1: Hyperparameters for Random Forest and Resampling Strategies

Resampling Strategies	Hyperparameters	Search Space	Optimal Values
No Resampling Strategy	Max Depth	5, 10, 15	15
	Min Samples Split	2, 5, 10, 15	2
	N estimators	100, 250, 500	100
Random Undersampler	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.9
	Max Depth	5, 10, 15	5
	Min Samples Split	2, 5, 10, 15	5
	N estimators	100, 250, 500	500
Random Oversampler	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	10
	Min Samples Split	2, 5, 10, 15	15
	N estimators	100, 250, 500	500
SMOTE	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.95
	K Neighbors	1, 3, 4, 5	5
	Max Depth	5, 10, 15	5
	Min Samples Split	2, 5, 10, 15	10
	N estimators	100, 250, 500	250
Borderline SMOTE	Proportion	0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	2
	Min Samples Split	2, 5, 10, 15	15
	N estimators	100, 250, 500	250
ADASYN	Beta	0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	5
	Min Samples Split	2, 5, 10, 15	5
	N estimators	100, 250, 500	250
Cluster SMOTE	Proportion	0.9, 0.95, 1.0	0.9
	Max Depth	5, 10, 15	5
	Min Samples Split	2, 5, 10, 15	10
	N estimators	100, 250, 500	100
Stefanowski	Strategy	weak_amp, weak_amp_relabel, strong_amp	strong_amp
	Max Depth	5, 10, 15	5
	Min Samples Split	2, 5, 10, 15	15
	N estimators	100, 250, 500	100
Safe Level Smote	Proportion	0.9, 0.95, 1.0	1.0
	Max Depth	5, 10, 15	10
	Min Samples Split	2, 5, 10, 15	15
	N estimators	100, 250, 500	500
SMOTE Tomeklinks	Proportion	0.9, 0.95, 1.0	1.0
	Max Depth	5, 10, 15	5
	Min Samples Split	2, 5, 10, 15	10
	N estimators	100, 250, 500	250
SMOTE PSO	K	2, 4	4
	N Pop	4, 10	4
	Num It	4, 10	10
	Max Depth	5, 10, 15	10
	Min Samples Split	5, 10, 15	15
	N estimators	100, 250, 500	250

Table A.2: Hyperparameters for Logistic Regression and Resampling Strategies

Resampling Strategies	**Hyperparameters**	**Search Space**	**Optimal Values**
No Resampling Strategy	Penalty	l2, l1, elasticnet, none	none
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	500
Random Undersampler	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.5
	Penalty	l2, l1, elasticnet, none	none
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	250
Random Oversampler	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.9
	Penalty	l2, l1, elasticnet, none	l2
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	700
SMOTE	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.9
	K Neighbors	1, 3, 4, 5	5
	Penalty	l2, l1, elasticnet, none	l1
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	700
Borderline SMOTE	Proportion	0.9, 0.95, 1.0	0.95
	Penalty	l2, l1, elasticnet, none	none
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	500
ADASYN	Beta	0.9, 0.95, 1.0	0.9
	Penalty	l2, l1, elasticnet, none	l1
	Solver	lbfgs, liblinear, saga	liblinear
	Max Iter	250, 500, 700	700
Cluster SMOTE	Proportion	0.9, 0.95, 1.0	1.0
	Penalty	l2, l1, elasticnet, none	l1
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	250
Stefanowski	Strategy	weak_amp, weak_amp_relabel, strong_amp	strong_amp
	Penalty	l2, l1, elasticnet, none	l1
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	250
Safe Level Smote	Proportion	0.9, 0.95, 1.0	0.95
	Penalty	l2, l1, elasticnet, none	l2
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	500
SMOTE Tomeklinks	Proportion	0.9, 0.95, 1.0	0.95
	Penalty	l2, l1, elasticnet, none	l2
	Solver	lbfgs, liblinear, saga	saga
	Max Iter	250, 500, 700	700
SMOTE PSO	K	2, 4	4
	N Pop	4, 10	4
	Num It	4, 10	10
	Penalty	l2, l1, elasticnet, none	l1
	Solver	lbfgs, liblinear, saga	liblinear
	Max Iter	250, 500, 700	250

Table A.3: Hyperparameters for XGBoost and Resampling Strategies

Resampling Strategies	**Hyperparameters**	**Search Space**	**Optimal Values**
No Resampling Strategy	Max Depth	5, 10, 15	5
	N estimators	100, 250, 500	500
Random Undersampler	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.5
	Max Depth	5, 10, 15	15
	N estimators	100, 250, 500	100
Random Oversampler	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	5
	N estimators	100, 250, 500	100
SMOTE	Sampling Strategy	0.5, 0.9, 0.95, 1.0	0.5
	K Neighbors	1, 3, 4, 5	3
	Max Depth	5, 10, 15	10
	N estimators	100, 250, 500	250
Borderline SMOTE	Proportion	0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	10
	N estimators	100, 250, 500	500
ADASYN	Beta	0.9, 0.95, 1.0	0.9
	Max Depth	5, 10, 15	10
	N estimators	100, 250, 500	250
Cluster SMOTE	Proportion	0.9, 0.95, 1.0	1.0
	Max Depth	5, 10, 15	10
	N estimators	100, 250, 500	100
Stefanowski	Strategy	weak_amp, weak_amp_relabel, strong_amp	strong_amp
	Max Depth	5, 10, 15	5
	N estimators	100, 250, 500	100
Safe Level Smote	Proportion	0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	15
	N estimators	100, 250, 500	100
SMOTE Tomeklinks	Proportion	0.9, 0.95, 1.0	0.95
	Max Depth	5, 10, 15	10
	N estimators	100, 250, 500	500
SMOTE PSO	K	2, 4	4
	N Pop	4, 10	4
	Num It	4, 10	10
	Max Depth	5, 10, 15	10
	N estimators	100, 250, 500	500

References

[1] National lung screening trial (nlst). Available at https://www.cancer.gov/types/lung/research/nlst, Accessed last time in August 2022.

[2] Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[3] Quitting smoking after diagnosis of lung cancer improves survival and reduces the risk of disease progression. *International Agency for Research on Cancer*, Jul 2021.

[4] Maha Alriyami and Constantin Polychronakos. Somatic mutations and autoimmunity. *Cells*, 10(8), 2021.

[5] S. Aruna and L. Nandakishore. *Empirical Analysis of the Effect of Resampling on Supervised Learning Algorithms in Predicting the Types of Lung Cancer on Multiclass Imbalanced Microarray Gene Expression Data*, pages 15–27. 02 2022.

[6] Gustavo Batista, Ronaldo Prati, and Maria-Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6:20–29, 06 2004.

[7] Paula Branco, Luis Torgo, and Rita P. Ribeiro. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343:76–99, 5 2019.

[8] Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains, 8 2016.

[9] Mateusz Buda, Atsuto Maki, and Maciej Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 10 2017.

[10] Sheng Chen, Yaqi Han, Jinqiu Lin, Xiangyu Zhao, and Ping Kong. Pulmonary nodule detection on chest radiographs using balanced convolutional neural network and classic candidate detection. *Artificial Intelligence in Medicine*, 107:101881, 05 2020.

[11] David Cieslak, Nitesh Chawla, and Aaron Striegel. Combating imbalance in network intrusion datasets. pages 732–737, 01 2006.

[12] Federica Corso, Giulia Tini, Giuliana Presti, Noemi Garau, Simone Angelis, Federica Bellerba, Lisa Rinaldi, Francesca Botta, Stefania Rizzo, Daniela Origgi, Chiara Paganelli, Marta Cremonesi, Cristiano Rampinelli, Massimo Bellomi, Luca Mazzarella, Pier Pelicci, Sara Gandini, and Sara Raimondi. The challenge of choosing the best classification method in radiomic analyses: Recommendations and applications to lung cancer ct images. *Cancers*, 13:3088, 06 2021.

[13] Google Developers. Imbalanced data. Available at https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data, Accessed last time in February 2022, 2021.

[14] Joana Diz, Goreti Marreiros, and Alberto Freitas. Applying Data Mining Techniques to Improve Breast Cancer Diagnosis. *JOURNAL OF MEDICAL SYSTEMS*, 40(9), SEP 2016.

[15] Shu Fang and Zhehai Wang. Egfr mutations as a prognostic and predictive marker in non-small-cell lung cancer. *Drug design, development and therapy*, 8:1595–1611, 09 2014.

[16] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. 01 2018.

[17] José Ferreira Junior, Marcel Koenigkam Santos, Federico Garcia-Cipriano, Alexandre Fabro, and Paulo Azevedo-Marques. Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Computer Methods and Programs in Biomedicine*, 159, 02 2018.

[18] Devon M. Fitzgerald and Susan M. Rosenberg. What is mutation? a chapter in the series: How microbes "jeopardize" the modern synthesis. *PLOS Genetics*, 15(4):1–14, 04 2019.

[19] Olivier Gevaert, Sebastian Echegaray, Amanda Khuong, Chuong Hoang, Joseph Shrager, Kirstin Jensen, Gerald Berry, Haiwei Guo, Charles Lau, Sylvia Plevritis, Daniel Rubin, Sandy Napel, and Ann Leung. Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific Reports*, 7:41674, 01 2017.

[20] Srwa Hasan, Ali Sagheer, and Hadi Veisi. Improving breast cancer classification using (smote) technique and pectoral muscle removal in mammographic images. *Mendel*, 27:8, 12 2021.

[21] Zhongyi Hu, Raymond Chiong, Ilung Pranata, Yukun Bao, and Yuan Lin. Malicious web domain identification using online credibility and performance data by considering the class imbalance issue, 10 2018.

[22] National Cancer Institute. Anatomy of the lung. Available at https://training.seer.cancer.gov/lung/anatomy, Accessed last time in January 2022.

[23] National Cancer Institute. Risk factors for cancer. Available at https://www.cancer.gov/about-cancer/causes-prevention/risk, Accessed last time in January 2022, 2015.

[24] National Cancer Institute. Targeted therapy to treat cancer. Available at https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies, Accessed last time in February 2022, 2020.

[25] Wenxiao Jiang, Guiqing Cai, Peter Hu, and Yue Wang. Personalized medicine in non-small cell lung cancer: A review from a pharmacogenomics perspective. *Acta Pharmaceutica Sinica B*, 8, 04 2018.

[26] Justin Johnson and Taghi Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27, 03 2019.

[27] Qi Kang, Xiaoshuang Chen, Sisi Li, and Mengchu Zhou. A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Transactions on Cybernetics*, 47:4263–4274, 11 2017.

[28] Harsurinder Kaur, Husanbir Pannu, and Avleen Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, 52:1–36, 08 2019.

[29] Bartosz Krawczyk, Mikel Galar, Lukasz Jelen, and Francisco Herrera. Evolutionary under-sampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, 38, 10 2015.

[30] Philippe Lambin, Emmanuel Rios Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud Stiphout, Patrick Granton, Karen Zegers, Robert Gillies, Ronald Boellaard, André Dekker, and Hugo Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990)*, 48:441–6, 03 2012.

[31] Choong Ho Lee and Hyung-Jin Yoon. Medical big data: promise and challenges. *KIDNEY RESEARCH AND CLINICAL PRACTICE*, 36(1):3–11, MAR 2017.

[32] Geewon Lee, Ho Yun Lee, Hyunjin Park, Mark Schiebler, Edwin Beek, Yoshiharu Ohno, Joon Beom Seo, and Ann Leung. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. *European Journal of Radiology*, 86, 09 2016.

[33] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based under-sampling in class-imbalanced data. *Information Sciences*, 409, 05 2017.

[34] Yong-Qiang Liu, Xiao-Lu Wang, Dan-Hua He, and Yong-Xian Cheng. Protection against chemotherapy- and radiotherapy-induced side effects: A review based on the mechanisms and therapeutic opportunities of phytochemicals. *Phytomedicine*, 80:153402, 2021.

[35] Nuno Moniz and Vitor Cerqueira. Automated imbalanced classification via meta-learning. *Expert Systems with Applications*, 178, 9 2021.

[36] Mehdi Naseriparsa and Mohammad Riahi Kashani. Combination of pca with smote resampling to boost the prediction rate in lung cancer dataset. *International Journal of Computer Applications*, 77, 03 2014.

[37] Radiological Society of North America and American College of Radiology. Lung cancer screening. Available at https://www.radiologyinfo.org/en/info/screening-lung, Accessed last time in February 2022, 2021.

[38] World Health Organization. Promoting cancer early diagnosis. Available at https://www.who.int/activities/promoting-cancer-early-diagnosis, Accessed last time in January 2022.

[39] Lizhi Peng, Bo Yang, Yuehui Chen, and Xiaoqing Zhou. An under-sampling imbalanced learning of data gravitation based classification. pages 419–425, 08 2016.

[40] Lizhi Peng, Haibo Zhang, Bo Yang, Yuehui Chen, and Xiaoqing Zhou. Smote-dgc: An imbalanced learning approach of data gravitation based classification. volume 9772, pages 133–144, 08 2016.

[41] 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. *Scientific Reports*, 10, 12 2020.

[42] Endang Purba, Ei-ichiro Saita, and Ichiro Maruyama. Activation of the egf receptor by ligand binding and oncogenic mutations: The "rotation model". 05 2017.

[43] Enislay Ramentol, Yaile Caballero, Rafael Bello, and Francisco Herrera. Smote-rsb *: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowledge and Information Systems*, 33, 11 2011.

[44] Stefania Rizzo, Francesco Petrella, Valentina Buscarino, Federica Maria, Sara Raimondi, Massimo Barberis, Caterina Fumagalli, Gianluca Spitaleri, Cristiano Rampinelli, Filippo de Marinis, Lorenzo Spaggiari, and Massimo Bellomi. Ct radiogenomic characterization of egfr, k-ras, and alk mutations in non-small cell lung cancer. *European radiology*, 26, 05 2015.

[45] Madeleine Scrivener, Evelyn E. C. de Jong, Janna E. van Timmeren, Thierry Pieters, Benoit Ghaye, and Xavier Geets. Radiomics applied to lung cancer: a review. *TRANSLATIONAL CANCER RESEARCH*, 5(4):398–409, AUG 2016.

[46] National Health Service. Biopsy. Available at https://www.nhs.uk/conditions/biopsy, Accessed last time in February 2022, 2021.

[47] National Health Service. What do cancer stages and grades mean? Available at https://www.nhs.uk/common-health-questions/operations-tests-and-procedures/what-do-cancer-stages-and-grades-mean, Accessed last time in January 2022, 2021.

[48] Takehito Shukuya and Kazuhisa Takahashi. Germline mutations in lung cancer. *Respiratory Investigation*, 57, 05 2019.

[49] Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71:7–33, 1 2021.

[50] American Cancer Society. What is lung cancer? Available at https://www.cancer.org/cancer/lung-cancer/about/what-is, Accessed last time in January 2022, 2019.

[51] American Cancer Society. What is cancer? Available at https://www.cancer.org/treatment/understanding-your-diagnosis/what-is-cancer, Accessed last time in January 2022, 2020.

[52] Bofan Song, Shaobai Li, Sumsum Sunny, Keerthi Gurushanth, Pramila Mendonca, Nirza Mukhia, Sanjana Patrick, Shubha Gurudath, Subhashini Raghavan, Tsusennaro Imchen, Shirley Leivon, Trupti Kolur, Vivek Shetty, Vidya Bushan, Rohan Ramesh, Tyler Peterson, Vijay Pillai, Petra Wilder-Smith, Alben Sigamani, and Rongguang Liang. Classification of imbalanced oral cancer image data from high-risk population. *Journal of Biomedical Optics*, 26, 10 2021.

[53] Jerzy Stefanowski and Szymon Wilk. Selective pre-processing of imbalanced data for improving classification performance. pages 283–292, 09 2008.

[54] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71:209–249, 5 2021.

[55] Seba Susan and Amitesh . Sso maj -smote-sso min : Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets. *Applied Soft Computing*, 78, 02 2019.

[56] Seba Susan and Amitesh . *Hybrid of Intelligent Minority Oversampling and PSO-Based Intelligent Majority Undersampling for Learning from Imbalanced Datasets*, pages 760–769. 01 2020.

[57] Seba Susan and Amitesh . The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art. *Engineering Reports*, 3, 04 2021.

[58] Rajat Thawani, Michael McLane, Niha Beig, Soumya Ghose, Prateek Prasanna, Vamsidhar Velcheti, and Anant Madabhushi. Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer*, 115, 11 2017.

[59] Pascale Tomasini, Preet Walia, Catherine Labbé, Kevin Jao, and Natasha Leighl. Targeting the kras pathway in non-small cell lung cancer. *The Oncologist*, 21, 11 2016.

[60] Ivan Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6, 11 1976.

[61] Cancer Research UK. How cancer starts. Available at https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts, Accessed last time in January 2022, 2020.

[62] Hualong Yu, Jun Ni, and Jing Zhao. Acosampling: An ant colony optimization-based undersampling method for classifying imbalanced dna microarray data. *Neurocomputing*, 101:309–318, 02 2013.

[63] Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. Oversampling method for imbalanced classification. *Computing and Informatics*, 34:1017–1037, 01 2015.