

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

An Interpretable Machine Learning Approach for Predicting Mutations in Lung Cancer using Gene Expression

Mafalda Colaço Parente Morais Da Costa



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Tânia Pereira, PhD

Co-Supervisor: Helder Oliveira, PhD

Supervisor: Prof. José Carlos Machado, PhD

August 8, 2023

An Interpretable Machine Learning Approach for Predicting Mutations in Lung Cancer using Gene Expression

Mafalda Colaço Parente Morais Da Costa

Mestrado Integrado em Engenharia Informática e Computação

August 8, 2023

Resumo

O cancro do pulmão é o cancro mais letal devido ao diagnóstico tardio e aos tratamentos limitados, sendo o cancro do pulmão de células não pequenas a principal forma da doença. A terapia direcionada é uma estratégia promissora para combater o cancro do pulmão. Terapias baseadas em inibidores foram desenvolvidas para o KRAS, EGFR, ROS1, ALK, BRAF e RET mas, após algum tempo, estas terapias perdem efeito e é necessário combiná-las com outras terapias mais evasivas. Encontrar padrões nas alterações genéticas dos pacientes com uma determinada mutação permite um tratamento mais eficiente. Para prever mutações, foi provado que a expressão genética é o tipo de dados que resulta melhor.

Neste estudo, *Machine Learning* é usado para encontrar padrões na expressão genética dos pacientes, com o objetivo final de ajudar médicos e cientistas a prever mutações no cancro do pulmão. A abordagem a seguir foi uma exaustiva seleção de *features* baseada em estudos da literatura, de forma a selecionar os genes mais relevantes para a previsão das várias classes. O modelo baseado em *Multi-layer Perceptron* (MLP) ofereceu a melhor performance, com um valor para a Area under the ROC Curve (AUCROC) de 93%, para a melhor previsão que foi EGFR, no Lung Adenocarcinoma (LUAD). O conjunto de *features* ideal foi selecionado com base na importância dos genes determinada pelo *Random Forest Selector* e, posteriormente, reduzido para as 100 melhores features, selecionadas por *Minimum Redundancy — Maximum Relevance* (MRMR). Uma abordagem multi classe foi testada com o objetivo de obter um conjunto ideal de features para a previsão em cada tipo de cancro do pulmão, LUAD e Lung Squamous Cell Carcinoma (LUSC). A performance obtida foi mais baixa, em comparação com os modelos binários, com uma AUCROC de 65% para LUAD e 58% para LUSC. O modelo final é interpretável, oferece os valores da importância de cada feature e ao oferecer explicações do tipo Shapley Additive Explanations (SHAP) para as previsões, de forma a que os médicos possam compreender quais os genes relacionados com as mutações mencionadas acima. O resultado deste trabalho pode contribuir para um diagnóstico mais precoce e para encontrar tratamentos direcionados eficientes, salvando milhões de vidas.

Abstract

Lung cancer is the deadliest form of cancer due to late diagnosis and limited treatments, with Non-small-cell lung cancer being the main form of the disease. Cancer Target Therapy is a promising strategy for fighting lung cancer. Therapies based on Target Inhibitors were developed for Kirsten Rat Sarcoma viral oncogene homolog (KRAS), Epidermal Growth Factor Receptor (EGFR), Receptor Tyrosine Kinase (ROS1), Anaplastic Lymphoma Kinase (ALK), V-Raf Murine Sarcoma Viral Oncogene Homolog B1 (BRAF), and Rearranged During Transfection (RET) but later on, they lose their effect, and it is necessary to combine them with other evasive therapies. Finding patterns along the gene expression of cancer patients allows for more effective treatment. For predicting mutations, it has been proven that RNA-seq gene expression is the type of data that works best.

In this study, Machine Learning (ML) is used to find patterns along RNA-seq information of the patients with the end goal of helping doctors predict mutations in lung cancer. The approach to follow was based on an exhaustive feature selection supported by related studies to choose the most relevant features for the different classes. From the experiments implemented, Multi-layer Perceptron (MLP) provided better performance, with an Area under the ROC Curve (AUCROC) value of 93% for the best prediction, which was EGFR, on Lung Adenocarcinoma (LUAD). The ideal feature set was selected based on the importance of the features determined by the Random Forest Selector followed by a dimensionality reduction of the 100 best selected by Minimum Redundancy — Maximum Relevance (MRMR). A multiclass approach was also experimented with the objective of obtaining an optimal subset for each cancer type, capable of accurately predicting the mutations mentioned before. The performance was lower compared with the binary models, with an AUCROC of 65% for LUAD and 58% for Lung Squamous Cell Carcinoma (LUSC). The final model is interpretable by providing feature importance scores and Shapley Additive Explanations (SHAP) of the predictions so that doctors can understand the genes related to the mutations mentioned before. The results of this work can contribute to a solution for earlier diagnosis and finding efficient, targeted treatments, saving millions of lives.

Acknowledgments

Com esta dissertação, encerro o capítulo da minha vida académica. Um capítulo desafiante, recheado de muito trabalho árduo e foco, mas também de muitas alegrias e desenvolvimento pessoal. Sem a ajuda da minha rede de apoio, não teria chegado tão longe e por isso, quero tirar um momento para agradecer a todos vocês que fizeram parte desta caminhada.

Em primeiro lugar, quero agradecer aos meus orientadores Tânia Pereira e Hélder Oliveira por confiarem em mim este trabalho e por me guiarem ao longo deste percurso. Queria também agradecer ao Francisco Silva e ao Marco por tudo o que me ensinaram com as suas contribuições.

Em segundo lugar quero agradecer à minha família por acreditar sempre em mim. À minha mãe por me dar apoio em todas minhas decisões, por me ouvir e procurar sempre compreender-me. À minha irmã quero agradecer o amor, a companhia e a empatia ao longo de todo o meu percurso académico. À minha avó quero agradecer o apoio incondicional e as palavras de amor. E por último mas não menos importante, ao Teddy pelo apoio emocional.

Quero agradecer aos amigos que conheci ao longo destes anos e que os foram tornando mais leves. À Marisinha, por ter um coração do tamanho do mundo, por cuidar de mim a todas as horas e por me motivar quando eu não encontrava força de vontade. À Xana, por todo o apoio e incontáveis ajudas, pela energia feroz de mulher do norte e pela paciência. Quero também agradecer aos meus roomies e colegas de intercâmbio, que me proporcionaram um semestre cheio de amizade e amor, fazendo de Buenos Aires casa. Quero ainda agradecer aos meus queridos amigos de Braga que nunca falhavam num momento de lazer.

Agradecer à U.DREAM que me ajudou a tornar-me numa melhor versão de mim mesma e que me proporcionou momentos inesquecíveis. Obrigada também a todas as pessoas que lá conheci que tornaram a minha passagem por lá tão especial. Obrigada Natacha, por seres um suporte e um exemplo de responsabilidade para mim. Obrigada Miguel e Sobral, por serem amigos atentos e por todos os sorrisos.

Quero agradecer ao João que acredita sempre em mim e nunca me deixa desistir. Obrigada por fazeres milhares de perguntas e, dessa forma, me ajudar a aprender.

Finalmente, agradecer aos meus amigos de Viana que cresceram comigo e que são parte da pessoa que sou hoje. Agradecer ao Tiago e à Alicia por serem como irmãos e os meus ombros amigos seja para o que for. Agradecer aos "Basilares" pelas conversas filosóficas e por uma amizade que nunca se desvanece e às "Amigas forever" que me aquecem sempre o coração.

Mafalda

“I tell young people: Do not think of yourself, think of others. Think of the future that awaits you, think about what you can do and do not fear anything.”

Rita Levi-Montalcini

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Objectives	2
1.3	Contributions	3
1.4	Document Structure	3
2	An Introduction to Cancer	4
2.1	Gene mutation	5
2.2	RNA-sequencing	6
2.3	Target therapy	8
2.4	Summary	8
3	Literature Review	9
3.1	Predictive Models using gene expression	10
3.2	Cancer-Gene Signature Identification	11
3.3	ML and DL interpretability	11
3.4	Discussion	13
3.5	Summary	14
4	Methodology	15
4.1	Data Description	15
4.2	Data Preparation	16
4.3	Feature Selection	17
4.4	Model Architecture	18
4.5	Summary	22
5	Results	23
5.1	Baseline Models	23
5.2	Experiments on multiple Feature Selection techniques	25
5.2.1	Gene Signature for Feature Selection	25
5.2.2	Feature Selection Results	27
5.2.3	Dimensionality Reduction	30
5.3	Hyperparameter Optimization Results	32
5.4	Multi-Class Approach	37
5.5	Discussion	39
5.6	Summary	41
6	Conclusions	42

A	Multiple occurrence of mutations Tables	49
B	Feature Selection	50
B.0.1	Feature Selector Library	50
B.0.2	KL Divergence	51
B.0.3	Extra Trees Classifier	52
B.0.4	Random Forest Experiments on MLP	53

List of Figures

2.1	Representation of a protein [1].	5
2.2	Example of substitution, deletion and addition mutations [24].	6
2.3	RNA-Seq experiment [50].	7
4.1	Architectural Design	19
5.1	Force plot for the prediction for the 6th patient on the test dataset.	34
5.2	Waterfall plot for the prediction for the 6th patient on the test dataset.	35
5.3	Beeswarm plot for the prediction for the 6th patient on the test dataset.	36
5.4	Stacked forced plot to explain all predictions on the test dataset.	36
5.5	The set of the 20 most relevant features for KRAS, for the final model trained on MLP.	38
5.6	The set of the 20 most relevant features for EGFR, for the final model trained on MLP.	38

List of Tables

4.1	Table of single mutations on LUAD.	16
4.2	Table of single mutations on LUSC.	16
4.3	Parameters to optimize LightGBM and MLP.	21
5.1	Baseline model results trained with cross-validation on GaussianNB.	24
5.2	Baseline model results trained with cross-validation on LightGBM.	24
5.3	Baseline model results trained with cross-validation on MLP.	25
5.4	Results based on the gene signature designed by "Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer" [53] using LightGBM.	26
5.5	Results based on the gene signature designed by "Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer" [53] using MLP.	26
5.6	Results based on the gene signature designed by "Widespread redundancy in -omics profiles of cancer mutation" [21] using LightGBM.	27
5.7	Results based on the gene signature designed by "Widespread redundancy in -omics profiles of cancer mutation" [21] trained on MLP.	27
5.8	Random Feature Selector results trained on LightGBM. All the features with importance >0 were selected.	29
5.9	Random Feature Selector results trained on MLP. All the features with importance >0 were selected.	29
5.10	Results from using ANOVA after KL Divergence, trained on LightGBM.	30
5.11	Results from using MRMR after KL Divergence, trained on LightGBM.	31
5.12	Random Forest Selector with ANOVA trained on MLP.	31
5.13	Results from using MRMR after Random Forest Selector, trained on MLP.	32
5.14	Final model results, trained on LightGBM.	33
5.15	Tuned Parameters for MLP.	33
5.16	Final model results, trained on MLP.	34
5.17	Multiclass problem trained on MLP.	37
5.18	Multiclass problem trained on MLP using Random Forest Selector.	37
5.19	Multiclass problem trained on MLP using Random Forest Selector with MRMR and ADASYN.	37
A.1	Table of multilabel mutations for LUAD.	49
A.2	Table of multilabel mutations for LUSC.	49
B.1	Results from removing the features with zero importance selected by the feature selector library, trained on LightGBM.	50

B.2	Results from removing the features with zero importance selected by the feature selector library, trained on MLP.	51
B.3	Results from using KL Divergence for feature selection, trained on LightGBM. .	51
B.4	Results from using KL Divergence for feature selection, trained on MLP. . . .	52
B.5	Extra Trees Classifier results trained on LightGBM.	52
B.6	Extra Trees Classifier results trained on MLP.	53
B.7	Random Feature Selector results trained on MLP. All the features with importance higher than 0.001 were selected.	53
B.8	Random Feature Selector results trained on MLP. The best 100 features were selected.	54
B.9	ADASYN implementation along with the Random Feature Selector and ANOVA, trained on MLP.	54
B.10	Tuned parameters for LightGBM.	55

Abreviaturas e Símbolos

ALK	Anaplastic Lymphoma Kinase
ANOVA	Analysis of Variance
AUCROC	Area under the ROC Curve
BRAF	V-Raf Murine Sarcoma Viral Oncogene Homolog B1
CNN	Convolutional Neural Network
DL	Depp-Learning
EGFR	Epidermal Growth Factor Receptor
ERT	Extremely Randomized Trees
GDC	Genomic Data Commons
ILIME	Influence-based Local Interpretable Model-agnostic Explanation
KL	Kullback–Leibler
KRAS	Kirsten Rat Sarcoma viral oncogene homolog
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanation
LORE	Local Rule-based Explanations
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MAPLE	Model Agnostic Supervised Local Explanations
ML	Machine-Learning
MLP	Multi-layer Perceptron
MRMR	Minimum Redundancy — Maximum Relevance
NN	Neural Network
NSCLC	Non-Small Cell Lung Cancer
RET	Rearranged During Transfection
RF	Random Forest
ROC	Receiver Operating Characteristic
ROS1	Receptor Tyrosine Kinase
SHAP	Shapley Additive Explanations
SCLC	Small Cell Lung Cancer
SMOTE	Synthetic Minority Oversampling Technique
TCGA	The Cancer Genome Atlas
SVM	Support Vector Machines

Chapter 1

Introduction

Lung cancer can be categorized as Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). NSCLC is the most common form of lung cancer (around 84%) and is sub-categorized into Adenocarcinoma (LUAD) and Squamous Cell Carcinoma (LUSC) [29]. Most patients of NSCLC are diagnosed at an advanced stage because cell degradation and metastasis are slower in NSCLC (than in SCLC), which lowers the chances of finding effective treatment, and so reduces the chances of survival. For a complete understanding of this research goal, it is essential to understand the concept of biomarkers. A biomarker is a change in the DNA of the tumor that causes the tumor to grow. Biomarkers indicate a more detailed scenario of a tumor, an identity, like a fingertip. To target therapy in NSCLC, six biomarkers are used, among others, named EGFR, KRAS, ROS1, BRAF, RET, and ALK.

KRAS mutations are present in 30% of lung Adenocarcinoma cancers and 5% of squamous cell carcinomas. KRAS mutations are usually related to people that have a record of smoking associated, but not necessarily. It can also be driven by pollution and carcinogenic materials. This mutation is usually present in Western European descent people. EGFR mutation is generally unrelated to smoking, being present in lung adenocarcinoma, usually in women, and more current in Asian people (approximately from 30% to 40%). This type of tumor is also more common in young adults. [35] ROS1 mutations, as RET mutations, are present in between 1 and 2% of the patients, and usually, it is present in lung adenocarcinoma. The patients with this mutation activated are, in general, younger and with a low record of smoking. ROS1-derived cancer is usually treated with a ROS1 inhibitor. Still, with time, the patient gains resistance to this inhibitor, and cancer evolves, so the approach to follow would be to treat with a targeted therapy approved to treat a different mutation(biomarker) [12]. As ROS1, ALK-derived cancer patients tend to be young and with a low smoking history. This type of cancer, like BRAF, is present in about 4% of the patients on LUAD [3]. BRAF is, at the same time, the name of a protein and the name of a gene, and the protein helps to control the mutation growth [4].

With the increase of new technologies and cloud storage possibilities, storage databases with open-access data appeared and made it possible to develop Biotechnology models. Since cancer is one of the deadliest and most complex diseases in the world, cancer prediction became a focus,

and access to clinical data and gene expression of the patients made these studies possible. In order to predict cancer, it is essential to analyze the gene mutations and the relationships of the patient's gene expression to understand the origin of the cancerous cells and plan the treatment. The problem with gene expression data is that usually, the datasets are small when related to the samples, the data is imbalanced and the number of features (genes in this case) is massive. This requires meticulous data preparation as the secret of the model's success. The innovation, compared to what has already been done, can be focused on data preparation and feature selection to choose the optimal subset of genes with more relevance for the problem we are looking to address. Machine Learning can work with complex information and find relationships in non-linear datasets.

1.1 Motivation

Cancer is a leading cause of death globally, killing thousands of people every year. Being a complex disease, it is necessary to understand its origin and the mutations that lead to the development of the disease in each patient in order to address appropriate and personalized treatment. The genetic information data that currently exists is highly informative but also quite voluminous. Deep learning combines the advancement of technology with biology, managing to find patterns and predict mutations that may be related to cancer in a record time that human work would never achieve.

Within cancer disease, lung cancer is the form that affects most patients. The problem to be solved focuses on six mutations, specifically, KRAS, EGFR, ROS1, ALK, RET, and BRAF. Therapies designed to target these biomarkers already exist, but, in general, patients develop resistance to these therapies. Therefore, it is necessary to understand better why and how the resistance happens. For example, for the EGFR mutation, there is an inhibitor, and the treatment has a positive balance, whereas, for the KRAS, most of the treatments fail. Since KRAS mutation is present in a significant amount of individuals, it is necessary to understand the genetic information of patients to predict the mutation and find an effective treatment. This investigation aims to facilitate the work of doctors and scientists who study the oncology area, providing them with a practical and fast method of predicting mutations and explaining on what bases the predictions are formulated, so the community can trust Artificial Intelligence and apply the models to similar problems.

1.2 Objectives

This investigation is focused on predicting mutations, specifically KRAS, EGFR, ROS1, RET, ALK, and BRAF mutations, and providing an interpretation for the models so doctors can understand them and work on finding the right therapies for lung cancer. These mutations have been studied, and target inhibitors were built based on previous information. The limitation of these

focused therapies is noticeable when the patient gains resistance, and it becomes necessary the implementation of more traditional therapies. The goal of this work is to develop model experiments and interpretable solutions based on ML, using techniques that resulted in a good performance in related studies. These models should result in new information about the mutations based on the gene expression, being the foundation of more powerful inhibitors. Target therapy based on genetic expression provides personalized treatment and can help save millions of lives.

1.3 Contributions

This investigation will provide the following contributions:

1. Identification of the relevant feature genes related to the problem we aim to solve
2. Discussion about techniques that aim to solve the problem of interpretability
3. Implementation of ML and DL models to predict mutations based on RNA-seq data
4. Analysis between several feature selection techniques and dimensionality reduction techniques
5. Comparison between independent predictions performance to a multiclass-based prediction
6. Gene interpretation of the best performance model

1.4 Document Structure

This document is divided into 4 chapters. The first chapter introduces what it is trying to be solved and why it is a problem, explains the motivation behind this research, and clarifies the objectives. The second chapter explains the biological concepts needed to understand the problem. The third chapter reviews the state of art literature to clarify what has been done concerning the problem and its limitations. The fourth chapter describes the data exploited, the preparation it needs regarding the problem, the feature selection, and the methodology to be followed. The fifth chapter makes an analysis of the results along the multiple approaches, comparing performances in order to break down the best model. The sixth and final chapter presents the best performance, the limitations of this work, and what could be improved in future work.

Chapter 2

An Introduction to Cancer

In this chapter a small context necessary to understand the problem is given. Firstly, understanding how cancer is formed and how gene mutations happen is crucial to understand why predicting mutations can be essential to target therapy. Second, the data used in this investigation will be RNA-seq information, therefore, it is important to acknowledge how it works and why it should be chosen over DNA information. Finally, a small introduction to cancer therapies is key to understanding why this investigation is important and how can be applied by doctors and scientists.

Cancer is a disease that comes from malignant cells that grow in an uncontrollable way and spread around the human body. During the normal function of the human body, cells grow and multiply themselves through cell division, and when they are old or damaged, they disappear to make space for new functional cells. Sometimes this process breaks down, and damaged cells multiply, which can lead to cancerous tumors. A tumor is defined by the American Cancer Institute as a lump or growth. The tumors can be either malignant (cancerous) or benign (not cancerous) [33]. Normal and malignant cells differ in a few obvious ways, the most notable of which is that cancer cells can metastasize, or migrate to new locations inside the body, creating new tumors. [31]. Malignant cells multiply by ignoring signals that should stop the division of cells, they hide from the immune system, which usually eliminates the damaged cells and, in some cases, they trick the immune system into protecting them and helping them grow. Another characteristic is that they may change their chromosomes by duplicating and deleting genome information [33]. Cancer can rely on errors in cell division, can be inherited, or can develop because of damaged DNA that comes from external factors such as harmful substances. The immune system is expected to fight the damaged cells, but as we age, it gets weaker, and it's more probable that a malignant tumor will grow. Different persons have different combinations of genetic changes, and as cancer continues to grow, additional changes continue to happen, meaning that the same tumor in different persons deserves individual attention. There are three main types of genes that can cause or help develop malignant tumors, called the "driver" genes of cancer. DNA repair genes are one of them and mutations in these kinds of genes tend to develop, being responsible for changing the genomic

information of a patient. Scientists found that specific mutations commonly occur in various types of cancer, so studies are being conducted on cancer treatments targeting these gene mutations.

2.1 Gene mutation

Gene mutations are characterized as alterations in the genetic sequence that commonly refer to changes affecting nucleic acids, the building blocks of DNA, that take place during cell division. The substances DNA and RNA store necessary information for reproduction in long-term memory. Genetic mutations can originate different consequences and they can occur at different levels, but most have a negative impact on the human body [38]. Somatic mutations are defined as mutations that happen in an individual organism and that are not passed to descendants. These DNA alterations happen after the individual is born and many cancers are caused by accumulated somatic mutations.

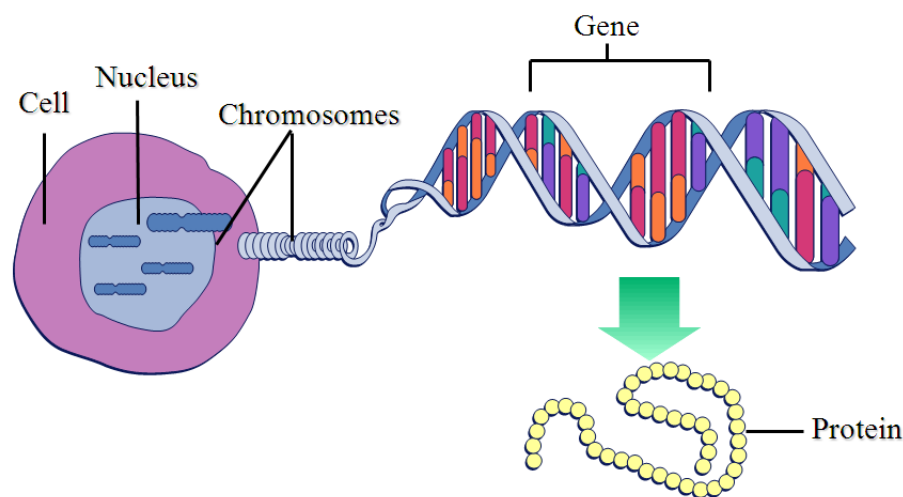


Figure 2.1: Representation of a protein [1].

Mutations can affect a single base pair (point mutation) or change the entire amino acid sequence (nonsynonymous mutation), which can lead to the urge of a protein or to the termination of one. Synonymous mutations do not alter protein sequences, but they only occur in sequences that code for amino acids. Indels are mutations that behave as insertions or deletions that can affect parts of a gene or whole chromosomes. Finally, there exist transposable mutations, elements that move around within the genome, producing new genes or disrupting existing gene functions. Driver mutations can be responsible for tumor initiation, progression, as well as metastasis of cancer.

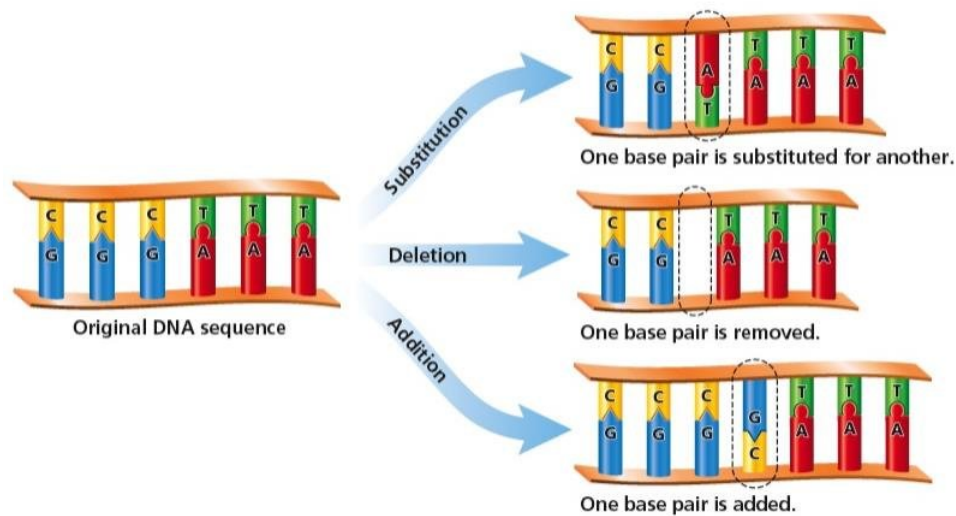


Figure 2.2: Example of substitution, deletion and addition mutations [24].

2.2 RNA-sequencing

The transcriptome is the entire set of transcripts present in a cell at a given developmental stage. To completely understand the disease and how it grows, a deep understanding of the transcriptome is crucial, such as determining what relevant molecular cells and genomes operate[40]. RNA-seq are molecules that have the information of the transcriptome using deep-sequencing methods with less noise and more normalized data than the other techniques used for prediction purposes. RNA-seq maps and quantifies the transcriptomes. The process starts with a population of RNA being converted to a library of cDNA fragments with adaptors attached to one(single-end sequencing) or both ends(pair-end sequencing). To obtain short sequences from one or both ends, each molecule is sequenced. Following the sequencing procedure, the reads are either aligned or reassembled to a reference genome or transcript to build a transcription map that gives information about the structure of the transcriptome and/or the expression level for each gene [50].

The main goal of a transcriptome is determining the differentially expressed genes or detecting variations in genes at different levels, that can be the most relevant genes/features. RNA-seq helps scientists to obtain reliable knowledge on the various genes that influence the development of a disease, such as if the genes are enabled, if they are activated or not, and in what circumstances. This kind of technology works well in investigating both known and unknown transcripts. There are some important advantages of this technology compared to previous ones. RNA-seq is useful for studying complex transcriptomes by revealing a precise location of transcription boundaries and giving information about the connectivity between multiple exons. Relative to DNA microarrays, RNA-seq has the advantage of having very low background signal, so, not having an upper limit for quantification, it depends on the number of sequences obtained. Opposite from DNA microarrays, RNA correct maps to unique regions of the genome. Since RNA does not have an

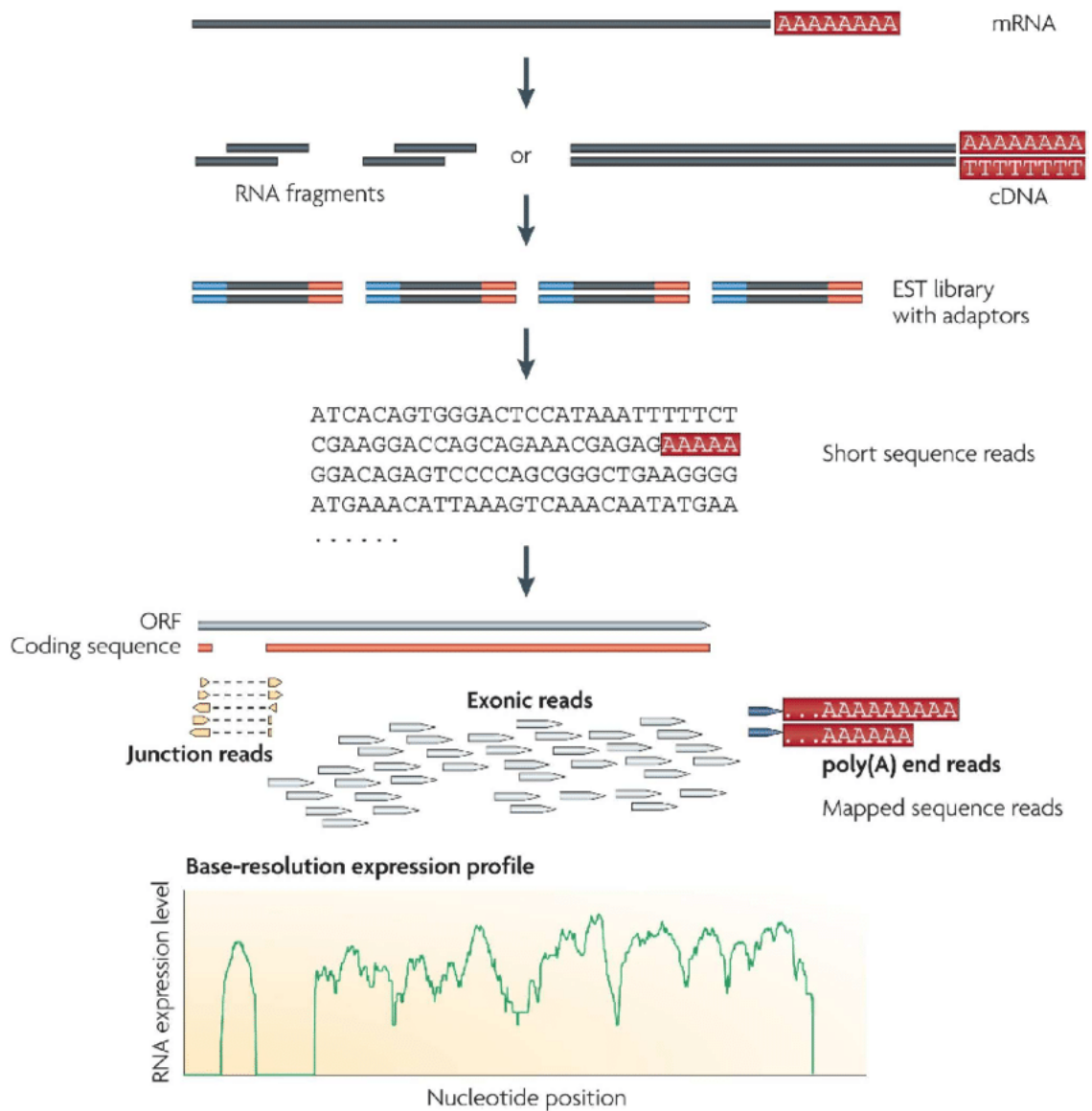


Figure 2.3: RNA-Seq experiment [50].

amplification step, requires less sample. RNA-sequencing is shown to be highly accurate for quantifying expression levels and showed high levels of reproducibility [50]. Rare alterations in transcriptome sequences can also be found in RNA-sequencing studies of cell populations adjusting to various environmental situations, helping in understanding evolutionary dynamics. [40]. Another advantage is that RNA-seq combined with computational analysis can be used to detect gene fusion events in disease tissues, which can be highly helpful in cancer research. Translocations or genomic rearrangements can be found using single-read or pair-end read applied to genomic DNA [43].

2.3 Target therapy

Target therapy acts in the cancer cells that enable the tumor to grow and develop, using drugs to control target proteins. National Cancer Institute describes it as "the foundation of precision medicine" [15]. This type of therapy can be implemented by itself, but, in general, it is more effective when combined with other traditional treatments. The creation of these therapies is based on identifying the pathways that contribute to genetic alterations. In order to improve these therapies, the investigation into the DNA changes and proteins that contribute to the progression of the tumor should be continued [14]. Target therapy can be an advantage compared to traditional therapies because, opposite to, for example, chemotherapy, it just acts on the cancer cells, not damaging the healthy ones. In addition, it acts on a level that triggers the immune systems and interrupts chemical signals, blocking molecular functions that cause cell division and cancer spreading. Advances have been made in controlling lung cancer, and target inhibitors for KRAS, EGFR, ROS1, ALK, BRAF and RET have been developed. The drawback is that after some time, cancer cells become resistant to the inhibitors, and, as it has mentioned before, these treatments become more effective when combined with the traditional ones, losing the advantages that arise from using target drugs when it comes to the side effects of traditional therapies. However, target therapy is not free from side effects, and while some can be mild, others can truly damage the patient. It is necessary to evaluate the advantages of the various techniques focusing on the patient background and resistance.

2.4 Summary

Cancer is a disease caused by an accumulation of gene mutations that mislead the immune system, spreading through the human body. These mutations happen at the DNA level. Cancer affects individuals differently since their genetic information is different. RNA-seq is a novel technology that helps scientists to obtain reliable knowledge on transcriptomes. Target therapy is focused on the protein driver of the tumor and can be a promising solution to save patients from cancer.

Chapter 3

Literature Review

To better approach this problem three main areas of the literature were considered. The first one, consists in ML and DL applied to gene expression data. Includes studies whose objective is to predict mutations on lung cancer based on gene expression, studies that apply machine and deep learning to predict mutations or to predict other variables based on RNA-seq data and other types of -omics data, model architectures based on deep learning to predict prognosis and to identify essential genes that affect the development of cancer disease. A lot of studies related to predicting mutations are conducted on data based on images. These studies were discarded since the data type is not the same and so, the architecture to be applied is also not the same. For the second group of searches, the search has made mostly on Google Scholar and the search has "deep learning" AND "lung cancer" AND "RNA" AND "predict" -"images". The second one is considering biological insights about cancer-driven genes where we dive into some of the techniques. The third subsection talks about the interpretability of ML and DL models. A lot of dedication has been put into the research of this subject, especially when it comes to analyzing medical data. There are already several libraries that focus on understanding, through data visualization and performance metrics, how black box algorithms work. In this chapter, we will discuss these libraries and their advantages and disadvantages compared to others. For both areas, the cast was limited to 5 years of studies since the recent ones covered the limitations of the oldest researches.

Machine Learning and Deep Learning have been broadly used for studies related to cancer in order to predict mutations. In the case of lung cancer, the patients have different clinical responses and results, even when the tumor type is the same. Finding patterns among cancer patients can be complex and time-consuming, but ML and DL, by helping to obtain abstract data representations, makes it easier and possible to identify useful subsets of genes that can anticipate the survival outcome of liver cancer. Machine Learning methods such as Principal Component Analysis, Clustering, and Autoencoding have been used to distinguish cancer types. ML has been widely used for the prediction of cancer therapy responses using generally Semi-supervised algorithms. The most promisor algorithms are Neural Networks and they can be either supervised, unsupervised or semi-supervised. The advantage of unsupervised is that it finds patterns in the input data based

on their distributions and does not require pre-existing labels. RNA-seq gene expression datasets usually do not have pre-existing prediction labels and because of that, the models are based on distribution[46].

As in NSCLC the effectiveness of the therapy widely depends on the individuals, predictive models are essential to plan therapy based on the specific characteristics of an individual instead of the general population.

3.1 Predictive Models using gene expression

Studies suggest that connections between genomic changes and patterns of gene expression data can serve as a foundation for building precise molecular relationships related to tumorigenesis.

To demonstrate that genomic alterations can be predicted using gene expression, for every tumor type, Mercatelli et al. build a model [42] for predicting single somatic mutations and copy number variations based on TCGA gene expression data of different cancer types. The focus mutations were TP53, which is one of the most common driver genes of cancer, and KRAS. The investigation concluded that genomic alterations could indeed be modeled through machine learning by gene expression across multiple cancer types. The AUROC of the models was between 0.6 and 0.9, depending on the cancer type. The resampling used was bootstrap with a recursive feature elimination on the 10000 highest variance gene expression values. Multiple machine learning models were tested, but the conclusion was that GBM was optimal for the prediction task.

In 2020, a study [34] was conducted on predicting the PIK3CA mutation in breast cancer, using gene expression data extracted from the TCGA pan-cancer dataset. The methodology implements a model that was built in the context of an investigation to detect the Pan-Cancer RAS Pathway Activation with Machine Learning [51], by Way et al. Although the goal of the two studies is distinct, the same data is used, and Way et al. suggest that their approach can be used to detect other genomic alterations, such as mutations. First, the genes with a large median absolute deviation are selected and after, centering and scaling are implemented. The model is based on a supervised elastic net penalized logistic regression model, the performance was measured by AUCROC with 10-fold cross-validation. The model performed with 84% on the test data. This approach is only applicable to genetic changes that have a significant impact on gene expression, like the Ras Pathway and the PIK3CA mutation.

Later in 2022, it has questioned if multi-omics data, instead of using a single type of data such as RNA-seq data, would not be an advantage when predicting mutations [21]. Logistic regression was the algorithm of choice to predict mutation samples from a given target instance. Since multi-omics data was used, the number of features compared to the number of samples was highly imbalanced, and an elastic net penalty was implemented in order to prevent overfitting, a strategy based on feature weights. The model was trained using 4-fold cross-validation and hyperparameter tuning with grid search was implemented to get the best AUROC performance. A few theories were experimented on in this study, and I would like to highlight three conclusions.

First, selecting the cancer-related genes with the help of domain knowledge improved the prediction performance. Second, gene expression data gives a better performance in predicting mutation than DNA. Third and final conclusion, multi-omics data does not have a significant impact on performance compared to using gene expression alone for prediction.

Focusing on deep learning approaches, three methods [20] [36] [28] designed for survival prediction based on Neural Networks and RNA-seq data were compared between them based on a concordance index, an algorithm that describes how well models are capable of distinguishing groups, and the p-value of the log-rank test. The evaluation was made using datasets from 12 different cancers extracted from TCGA. The conclusion was that Cox-nnet performed better, supporting that when it comes to biological data, simple architectures work better.

3.2 Cancer-Gene Signature Identification

Cancer has complex molecular expressions, it is dependent on the molecular expressions of their patients, and the therapies and consequences can be very different from individual to individual. A study was conducted [19] that analyzed if an individual gene was of significant relevance when predicting cancer. The goal of this study was to develop a universal classifier for various cancer types. The base idea of this investigation was that "most cancers from different tissues tend to have a common nature that is universally different from that of normal tissues". An obstacle that TaeJin et al. faced was that deep learning has difficulty predicting the importance of a specific feature, and this, in biology and in cancer studies is crucial since the identification of the cancer driver gene is essential both for prognosis and therapy. The conclusion was that there are two group types of cancer, the first group, where the expression profile of multiple genes can be enough to identify a cancer type, and the second group, where oncogene driver genes belong, meaning that a single gene can contribute a lot to a particular tumor. This study suggests that Deep Learning is not only useful for the prediction of cancer but also for learning essential characteristics of tumors in order to develop new strategies for therapy. Ye et al. [53] propose an unsupervised feature selection method to identify gene signatures for Non-Small-Cell Lung Cancer. The method was tested on gene expression data from TCGA on the LUAD and LUSC subtypes. The outcome was a gene signature based on 17 genes. Crawford et al. [21] proposed a set of 268 cancer-related genes selected based on prior knowledge from previous studies and on the COSMIC Cancer Gene Census.

3.3 ML and DL interpretability

When talking about ML and DL applied to medicine, doctors and scientists are usually skeptical of the solutions because they do not understand how the algorithm works, questioning the fairness and biases of the outcome. Even though, in general, black box models outperform interpretable models, such as decision trees, the last ones are the favorites because of the transparency

they provide. In the medical field is definitely necessary to understand how decisions and predictions are made, considering these assumptions affect human lives directly. Much research has been and is still carried out on interpreting decisions made by black box models so doctors and patients can trust Artificial Intelligence, using efficient technology to accelerate medicine's evolution. Currently, many techniques and libraries are available and are simple to use. Interpretable models can be global or local. Global techniques make conclusions about the model behavior and decision-making process, understanding patterns and relationships inside the data and feature importance. On the other side, Local techniques focus on explaining at the prediction level for a specific input. These techniques aim to answer questions such as, what features influence a specific prediction and how changes in those features can affect the prediction. In this section, we will discuss techniques of both types and how they can be helpful.

One concern around model interpretability is how to evaluate the techniques and what metrics to use in order to choose the best one. ElShawi et al. [23] conducted a study in 2020 on interpretability in healthcare, more specifically, on the most popular local techniques, namely LIME, SHAP, LORE, ILIME, and MAPLE. The conclusions show that no clear winner exists because there is no technique that achieves great performance for all the metrics. LIME is a technique where the prediction is explained by training an interpretable model on a disturbed sample centered around the input instance of interest. This model is limited regarding clear coverage because it is uncertain that an explanation for a particular sample is applicable in the region of interest around. ILIME addresses the problem by weighting the perturbed samples according to their impact on the feature that is being analyzed. Therefore, another weakness is that it heavily depends on the weights supplied to the perturbed samples. MAPLE combines Random Forests to select a relevant neighborhood by assigning weights to the features, followed by a feature selection technique to score the instances based on the reduction of impurity related to the label. MAPLE is a highly accurate model that provides clear explanations and, despite being primarily a local technique, has the ability to detect global patterns within the data. In contrast with LIME, ANCHORS guarantee that instances present in the same neighborhood are similar by finding the features enough to make a prediction. LORE provides a rule-based explanation working by perturbing the feature of interest and observing the variations in the model's predictions. SHAP is a technique based on game theory applying the logic of a fair distribution of gains among players on features. When it comes to individual explanations, SHAP is the most robust technique. On the downside, it is computationally expensive. To assess the evaluation, the following metrics were used: identity, separability, similarity, execution time, bias detection, and trust. MAPLE proves to be the best technique to detect *identity*¹, opposite of LIME, that has the highest performance on *separability*², which goes by the assumption that all features are relevant for the prediction. SHAP and MAPLE are the best technique son detecting *bias*, and ANCHORS achieve the highest score on the *trust*³ metric.

¹Identity stands for if two instances are identical, the explanations must be identical too.

²Separability stands for if two instances are different, the explanations are different. This metric is based on the assumption that every model feature is relevant for the prediction.

³Trust measures the level of agreement between the explanations generated by interpretability frameworks.

Later, in 2022[26], a study was conducted to compare a local technique(a global surrogate model based on decision trees) and a global technique(LIME). Using medical datasets, the experiments were made using SVM and MLP. The main objective of the study is to determine the overall accuracy of each model, to assess the global and local interpretability of each model, and to compare if one technique is better than the other using *fidelity*⁴. The outcome is that Global Surrogate outperforms LIME. However, the experiment was based only on MLP and SVM as representatives of all the black box models, as cautioned by the authors.

A plausible metric to evaluate an interpretability model is based on the features contributions. For this purpose, it is important to take into account the task's final objective, the dataset that serves as the input, and the model response. In this case, the limitation is that assessing this metric, it is required domain knowledge available to support the comparison. Meng et al. conducted an experiment [41] to evaluate feature importance on mortality prediction based on interpretable models and fairness metrics. Several interpretable models were assessed such as gradient-based models, perturbation-based models, a glass box model interpretation, and a model based on a random shuffle of features. ArchDetect, a perturbation-based model, gives the best-performing feature importance overall. It is a curious outcome because, at the same time, that offers the worst AUC performance, feature importances are the most reasonable for the task when compared with domain knowledge.

3.4 Discussion

In section 3.1, it is demonstrated the potential of using RNA-seq data for prediction problems, when compared to other types of omics data. Besides that, the studies prove the efficiency of ML and DL when it comes to predicting genomic alterations and survival outcomes along cancer types. The discussion presented, highlights the importance of feature selection, using domain knowledge to build the models, and choosing appropriate models for the task, in order to achieve better performance. In the section 3.2, the methods presented prove that understanding the characteristics of tumors or developing a gene signature can be beneficial in achieving higher performance on ML and DL. It is possible, along with domain knowledge and DL, to discover general cancer gene signatures that will work along different cancer types or focus on a subtype.

In the section 3.3 it is presented the importance of model interpretability in healthcare and how difficult can be to evaluate the techniques. The need for transparency in medical decision-making is urgent, considering the direct impact on human lives. This discussion is needed because black box models outperform interpretable models when it comes to genetic data. Various techniques for model interpretability are discussed, including global and local approaches. Global techniques provide insights into the overall behavior and decision-making process of the model, while local techniques focus on explaining individual predictions and the influence of features on those predictions, and depending on the task goal, both can be useful and interesting. There are exposed

⁴Fidelity refers to how the explanations generated by a model mimic its behavior and accuracy.

different techniques, and the choice of the method to follow is going to depend on the dataset, the requirements of the task, and the desired level of interpretability.

3.5 Summary

With a view on this chapter, ML and DL applied to healthcare, Gene signature, and Interpretability were discussed. The first subsection had an overview of studies using ML and DL for predictions based on gene expression data. The conclusions are relevant to the research being developed, and the results give us insights into what technique to use and what could work on our problem. Feature Selection is an important part when using gene expression because of the number of features when compared to the available samples. It is important to select features with high importance for the problem we are trying to solve and, for that reason, studies on discovering a cancer-gene signature have been performed, either in general or either focused on Lung cancer. This discussion is presented in the second chapter. Finally, interpretability is an issue for doctors and researchers regarding ML and DL. In the third chapter, multiple techniques are exposed, and a comparison is presented between them based on relevant metrics. This chapter was necessary to define how to approach our problem.

Chapter 4

Methodology

In this section, there is presented an analysis of the lung cancer datasets used for this project, the data cleaning steps, the used feature selection mechanisms are explained, and finally, the model architecture is described. The goal of this project is to understand better KRAS, EGFR, ROS1, ALK, RET, and BRAF mutations tumors by analyzing the gene expression of the patients. Machine Learning and Deep Learning using RNA-seq information have proven to be efficient in cancer and gene prediction. RNA transcriptomes data is complex because of the high dimensions (genes) and the existence of duplicates. Therefore, cautious data preparation and feature selection are needed in order to avoid over-fitting.

4.1 Data Description

In this work, four datasets were used containing information from LUAD and LUSC patients. The Cancer Genome Atlas (TCGA) [32] offers open-source data from American cancer patients. The data used is available in the Genomic Data Commons (GDC) [30] portal. For the purpose of this work, it is necessary to gather information related to gene expression, RNA-seq information, and information about the gene mutations present in the patient. This second information is presented as MAF files that can be obtained by filtering the information on the data category "*Simple Nucleotide Variation*". Regarding the RNA information, there are 539 cases of **LUAD** and 502 cases of **LUSC**. The mutation data has 618 cases for LUAD and 549 for LUSC. Combining the two datasets of information for each cancer type, 532 and 488 samples of LUAD and LUSC, respectively, remain. Initially, there were 19962 gene expression features for both LUAD and LUSC.

Analyzing the final datasets, the number of features is much higher than the number of samples. A considerable imbalance can also be observed when comparing the minority class with the majority class, classes defined for the presence or not of gene mutations. By examining the tables 4.1 and 4.2, it is possible to conclude that the genes are differentially expressed both in LUAD and LUSC.

	KRAS	EGFR	ROS1	RET	ALK	BRAF
LUAD	145	78	24	21	38	45
LUSC	9	15	43	23	20	14

Table 4.1: Table of single mutations on LUAD.

Between a multiclass or multilabel problem, we chose to explore a multiclass model because the occurrences for the multilabel model were not enough, and the conclusion would be that these occurrences are not that common. The occurrences of multi mutations can be analyzed at [A.1](#) and [A.2](#). The performance also was not expected to be better, and the objective of this work is to predict the mutations based on gene expression and not based on other genomic alterations. Taking into account single mutations occurrence for multiclass classification, 496 samples were used for LUAD, 220 of the samples without mutation, and 478 samples for LUSC, 374 without mutation.

4.2 Data Preparation

The TCGA provides a JSON file with all the metadata from a patient, being possible to relate different data to a patient through the TCGA Barcode and the RNA-seq information was organized accordingly for each patient. Since a patient can have multiple genes mutated, a label for each gene mutation was created, and the mutation information was added to the gene expression data based on the patient's *Barcode*.

The first crucial step of our methodology is to remove redundant features and nonsignificant features to reduce the size of the dataset and increase model performance. Some duplicated columns existed and were removed, and columns with only 0 values were also removed because that information was not relevant to the prediction. After removing these columns, we are left with 19509 features for LUAD and 19507 features for LUSC. Our study has a specific focus, and relevant features will be the ones that are directly associated either with the gene mutations we are looking to predict or lung cancer. The features with low variance or a small range of values were also removed in the majority of the models because these features do not provide vital information for the model to learn the patterns. In the case of removing the features with low variance, we are left with 17368 features for LUAD and 17566 features for LUSC. After cleaning our data, the next step will be the Feature Selection stage.

	KRAS	EGFR	ROS1	RET	ALK	BRAF
LUAD	129	70	12	12	20	33
LUSC	9	12	37	20	15	11

Table 4.2: Table of single mutations on LUSC.

4.3 Feature Selection

A concern that arises when studying cancer prediction is the ability to generalize the methods to future patients, based on new data and new information that can come to light. Genetic data have high dimensionality (genes features), which leads to model overfitting and decreased prediction accuracy. One solution to reduce over-fitting is to apply feature selection in order to choose the most relevant features for the prediction. The features used in this investigation will be the gene counts of the patients. Studies and methods on how to select the most relevant feature selections are being explored when it comes to cancer prediction, multiple techniques were tested to find an optimal method or combination. As mentioned in 3.2, gene signatures based on other studies were also tested as a form of dimensionality reduction in a relevant way for this project. In total, 12 combinations were tested for feature selection for the best training and testing performance, granted by MLP.

Liu et al. [37] propose a gene selection method based on **Kullback–Leibler divergence** (KL divergence) in order to avoid ML models from overfitting when using genetic data. KL divergence is defined as

$$KL(P||Q) = \sum x P(x) * \log(P(x)/Q(x))$$

and can be used for gene selection to build a deep-learning network model. In this study [37], KL divergence is used to select the lung-cancer related genes as the input of a neural network. This method uses deep learning instead of machine learning because gene expression datasets are unbalanced small datasets with plenty of features, leading to insufficient learning parameters for machine learning methods. KL is a method that measures the difference in two probabilistic distributions, in the case being the difference of the gene expression features with the target column over the same variable. If the result of KL Divergence is equal to zero, the feature and the target variable are independent, otherwise, the larger the value, the higher the relationship between the feature and the target.

Extremely Randomized Trees (ETR) [25] is an ensemble technique of decision trees. Each decision tree is based on the training dataset, and at each node, a random subset of features is considered. The best feature to split the data is selected based on Gini Impurity in this case. The randomness in feature selection creates diverse decision trees. This method can be used for feature selection by ordering the features based on their Gini Importance and selecting the top k features. For the implementation of this feature selector *ExtraTreesClassifier* from *scikit-learn* was used.

Random Forest Trees (RF Trees) is also an ensemble technique, and the logic is similar to Extremely Randomized Trees. The big difference is while ERT samples the entire training dataset for each tree, RF samples subsets the dataset through bootstrapping [6]. So, in order to avoid overfitting, not every tree sees all the features or all the samples. Another difference is that while the nodes of the trees in ETR split randomly, in RF the nodes are split analyzing the best split. In the same way that ERT can serve as a feature selector, RF can be as well used for the same purpose. RF

Trees for feature selection were implemented using the class *RandomForestClassifier* from *scikit-learn*. Some parameters were adjusted such as the *n_estimators* was set to 1000, *random_state* to 42, and *class_weight* to 'balanced'.

Minimum Redundancy — Maximum Relevance (MRMR) works by selecting the minimal-optimal subset of features, by maximizing the relevance and minimizing the redundancy of the selected features [22]. It works by, first, computing the relevance between each feature and the target variable, followed by measuring redundancy between features based on mutual information. A number of *n*-features wished can be selected. Interestingly, this method was developed and focused on applying to gene expression data. This technique was implemented with the help of a library available on *Git Hub* [13].

Analysis of Variance (ANOVA) [49] calculates the variation between a target label and a feature. Starting by defining a hypothesis of either, there are no significant differences between the means of the groups, either that at least one group mean is different. The overall mean is calculated based on all samples and then between-group variability is compared to within-group variability. This comparison is made by calculating the *F-statistic* and the final features by determining a *p-value*.

Feature Selector is a library available on *GitHub* [18] with five methods implemented for dimensionality reduction. In our study this library was used to find the zero-importance features to be removed, calculated using a gradient boosting machine. This library also incorporates some interpretability, such as plotting the the feature importances and plotting the cumulative feature importance.

This step was the most laborious since it is challenging to choose the most relevant features for the task in a high-dimensional and imbalanced dataset. Also, in this work, feature selection requires special attention since one of the goals of this study is to understand what are the most important features for predicting a given mutation.

4.4 Model Architecture

As it is possible to analyze based on the image below 4.1, the next step after cleaning the data is to normalize and scale the data. Normalization is done according to L2, scaling the samples by dividing them by their Euclidean norm. Normalization is used to adjust the values to a similar scale so the scale of the samples does not impact the performance of the models. Followed by standardization using *StandardScaler* from *scikit-learn*. Data was split into 70% of training data, and 30% for test data, as commonly defined around classification problems. *StandardScaler* is fitted into the training data and used to transform the test data. **StandardScaler** centers the data around 0 and scales it to have a standard deviation of 1. This step is essential because in this work gradient descent-based algorithms that require the data to be scaled are implemented [5].

In the next step, an exhaustive feature selection and a posterior dimensionality reduction were performed. This two-step approach has been conducted to combine the advantages of feature

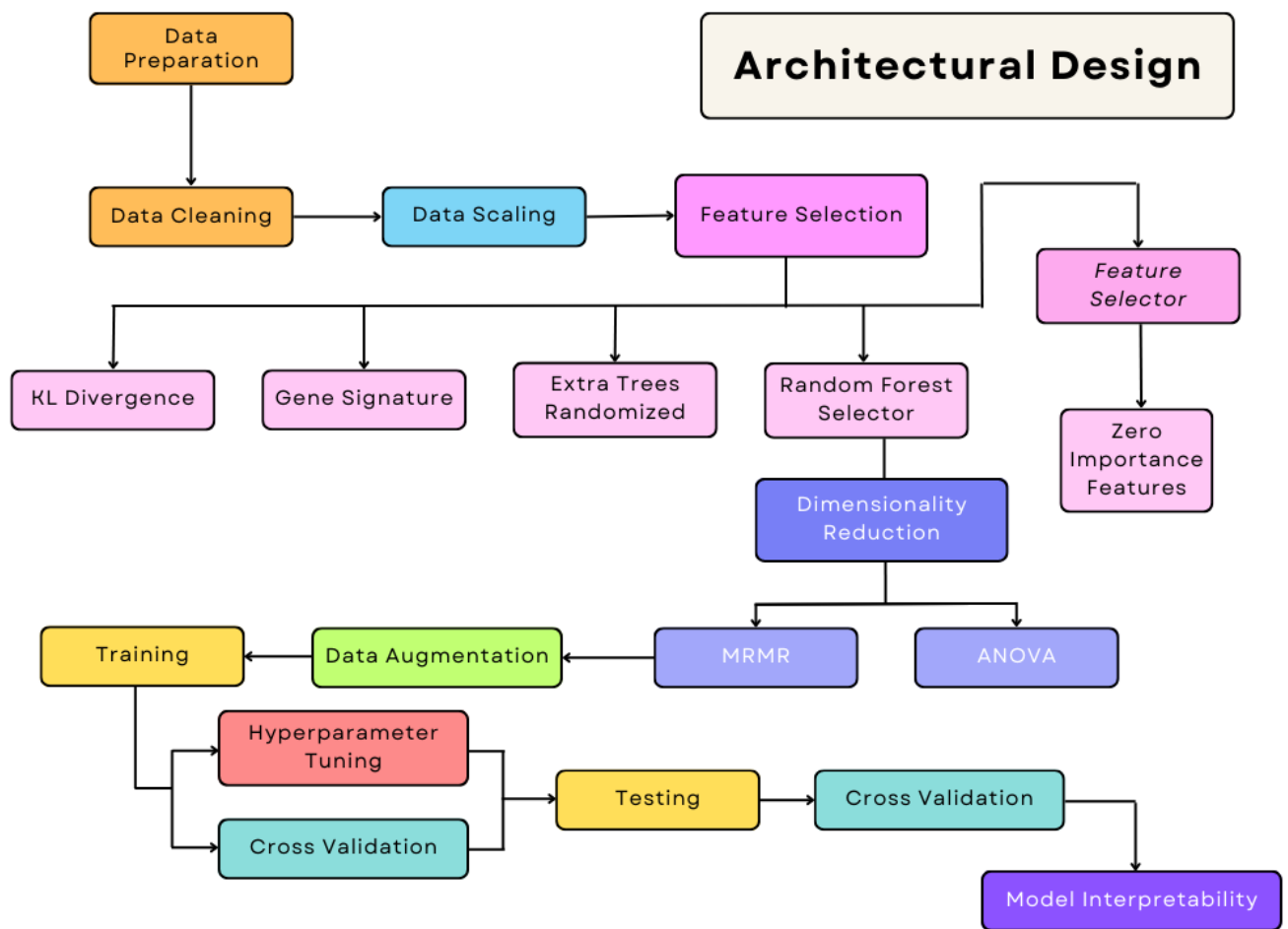


Figure 4.1: Architectural Design

selection and dimensionality reduction to improve performance. By combining these two techniques, a more relevant feature selection is expected, making sure that the irrelevant and redundant features are not taken into account. These techniques are useful to avoid overfitting since, with the combination of the two methods is unlikely that the model memorizes noise, resulting in a better generalization of unseen data. Additionally, computational efficiency is improved, making the models faster and resulting in better model performance, simultaneously, since the features selected are highly informative for the task. Each of the models detailed in the feature selection 4.3 was tested individually, and based on the best performance, either MRMR or ANOVA were applied to select the best 100 features.

For the binary approach, 12 predictions for the model were performed since 6 gene mutations are being studied both on LUAD and LUSC. For the multiclass approach, as mentioned before, the prediction was based on unique occurrences of the mutation, and a *OnevsRest* approach was followed. The multiclass model methodology will be based on the binary model offering better performance.

Based on the literature, based on the data, and based on the complexity of the task, two algorithms regarding Machine Learning and a simple neural network were selected. For the ML models, Gaussian Naive Bayes served as a base to obtain a comprehensive understanding of the data and to see how a generical and simple model would work for the purpose of the task. **Gaussian Naive Bayes (GaussianNB)** is a probabilistic classification algorithm based on Gaussian distribution. GaussianNB is based on the premise that the features are independent of one another and follow a Gaussian distribution. After defining the class probabilities, the Naive Bayes Theorem is employed to compute the posterior probability of each class, considering the feature values of a new instance [7].

LightGBM was selected based on the literature on predictive models using gene expression data. **LightGBM** is a gradient-bosting algorithm based on decision trees. It is known for its computational efficiency and high performance. Gradient boosting models use gradient-based optimization to train the set of decision trees by updating, in each iteration, the weights of the model to reduce loss. LightGBM differs from other gradient-boosting algorithms on using leaf-wise growth instead of depth-wise growth [8], meaning it grows the tree by splitting the leaves. Besides that, it uses Gradient-based One Side Sampling by keeping the instances with larger gradients, that will contribute more to information gain and by randomly dropping the instances with smaller ingredients. This technique reduces the computational cost without affecting accuracy. After that, LightGBM uses Exclusive Feature Bundling(EFB) to improve efficiency. EFB groups the features that have similar gradients and optimizes them as a group, resulting in less memory usage during training [9].

Deep Learning is an area of interest being explored in healthcare since it handles better big data. Based on the size of the dataset and the complexity of the task, and as we have seen above, simple networks should work better with the data being handled in this research, and that will be the approach to follow. MLP was the chosen model for its simplicity, and fast performance. **MLP** stands for Multilayer Perceptron. MLP is an artificial neural network, structured by an input layer, one or more hidden layers, and an output layer. It is a feedforward neural network, which means that data flows from the input layer all the way to the output layer in a single direction. Backpropagation allows MLP to adjust the weights on the input layer in each interaction by computing the mean squared error across the input and output pairs. This mechanism is what allows the network to learn the patterns across the data [10].

GaussianNB results were discarded in the first experiments since LightGBM and MLP showed better performance on precision and recall.

The problem of imbalanced data is a big constraint based on the limited number of examples of the minority class, limiting the model on learning the decision limit. In order to overcome the problem, three techniques of data augmentation were experimented, namely *SMOTE*, *Borderline-SMOTE*, and *Adaptive Synthetic Sampling (ADASYN)*. *SMOTE* performs data augmentation by generating samples along the lines generated between the minority class points, by using KNN to select the neighborhoods. *Borderline SMOTE* generates synthetic samples along the decision boundary between the two classes. *ADASYN* uses the density distribution to set the number of

samples per neighborhood. The sampling strategy used through all the techniques tested is only to generate samples from the minority class.

The final phase was to optimize the algorithm’s parameters. This step had a major impact on model performance and on balancing precision and recall values. Tuning Hyperparametrization was implemented using BayesSearch, on LightGBM, and grid search, on MLP. The list of values for each parameter that is tuned to optimize the algorithms is detailed in table A.1. Along with those parameters, in order to control the randomness of the splitting between the train and test set, the *random_state* was defined as 42 for both algorithms. Additionally, for LightGBM, *boosting_type* was settled as *'dart'* because according to LightGBM documentation [11] *'dart'* is suggested to get better accuracy, *is_unbalanced* was set to *True* because the dataset is imbalanced, and *objective* as set to *'binary'* since the classification followed a binary approach.

Algorithm	Parameters	Values
LightGBM	learning_rate	from 0.001 to 0.99
	n_estimators	from 50 to 300
	max_depth	from 1 to 30
	num_leaves	from 2 to 1000
	min_data_in_leaf	from 1 to 100
	feature_fraction	from 0.1 to 1.0
	subsample	from 0.1 to 1.0
	bagging_freq	from 0 to 20
	bagging_fraction	from 0.1 to 1.0
MLP	hidden_layer_sizes	(50,), (100,), (50,50,50), (50,100,50)
	activation	'identity', 'logistic', 'tanh', 'relu'
	solver	'lbfgs', 'sgd'
	alpha	0.0001, 0.01, 0.05
	learning_rate	'constant', 'invscaling', 'adaptive'
	batch_size	'128', '512'
	early_stopping	True, False

Table 4.3: Parameters to optimize LightGBM and MLP.

To evaluate the performance of the model, *Balanced-Accuracy*, *AUROC*, *Precision*, and *Recall* were measured. When splitting the data, some randomness is introduced. To reduce the generalization error and to prevent over-fitting, cross-validation was used. Accuracy is defined as [2]

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

where TP^1 , TN^2 , FP^3 , FN^4 are referring to the predicted values. Accuracy is one of the most common metrics to evaluate performance, but it is not the most appropriate to use when the dataset is imbalanced. The reason behind this is that if the model only predicts for the majority class, accuracy will be very optimistic, not considering the minority class that, in imbalanced binary classification, is usually the class of interest. On the other hand, *balanced_accuracy_score* was formulated to adjust better on imbalanced data. *Balanced_accuracy_score* is defined as

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

where *Sensitivity* evaluates how a model accurately predicts the actual positive labels and *Specificity* evaluates how a model predicts the actual negative labels. In this work, the labels of interest are the ones belonging to the minority class, and for that reason, *Precision* and *Recall* are used to evaluate the models. *Precision* is defined as [?]

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

and measures the ability of the model to not label negative samples as positive. *Recall* is formulated as [?]

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

and evaluates the ability of the model to detect positive labels. *AUROC* it is an important metric to measure the capability of a model to distinguish labels. ROC is a curve that plots the true positive rate against the false positive rate. AUC computes the degree of separability [16]. The set of metrics complement each other and guarantee that both the performance of the model to predict the majority and minority class are evaluated.

4.5 Summary

For the final models, two datasets were used. The datasets are highly imbalanced, so the task of selecting the most important features for each prediction is particularly challenging. To ease that task, data preparation is implemented based on removing redundant, non-significant features and low variance features. Feature Selection was an exhaustive implementation of methods based on a literature review and a crucial step to improve performance, choosing the most relevant genes for predicting a mutation. Three distinct algorithms, between binary and multiclass, were optimized to train and test the data.

¹TP = True Positive

²TN = True Negative

³FP = False Positive

⁴FN = False Negative

Chapter 5

Results

In this chapter, the results of the experiments design on the methodology for both binary classification and multiclass will be presented and discussed.

The models are trained with cross-validation, and the results are the mean average of the results for all folds. This is a way to mitigate the randomness along the models, present when the data is split and when applying data augmentation.

5.1 Baseline Models

To get insights into how the data would behave in terms of performance, 3 baseline models were implemented using 3 different algorithms to train the data. This is important to establish a benchmark for performance evaluation, to identify overfitting, and, based on this information, to evaluate if more complex models can lead to better predictions. Baseline models are more simple and easy to interpret and can be helpful to understand what features can be important for the prediction by comparing the models performance.

The baseline modes were trained with normalized and standardized data, with cross-validation. The results for the baseline models can be analyzed in table 5.1 when training with GaussianNB, in table 5.2 when training with LightGBM and in table 5.3 for MLP. In the models can be observed low performance on LUSC in general. The reason is that all the classes in LUSC are highly imbalanced, the metrics that we are focusing on for this dataset will be *Precision* and *Recall* to verify if the model can actually predict mutations. Another pattern observed is that in LUAD, KRAS and EGFR are expected to have better performance, meaning that the imbalance is not so significant in these mutations.

GaussianNB					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.638+/-0.062	0.631+/-0.067	0.449+/-0.080	0.511+/-0.111
	EGFR	0.675+/-0.064	0.669+/-0.052	0.335+/-0.055	0.529+/-0.092
	ROS1	0.532+/-0.067	0.532+/-0.067	0.200+/-0.400	0.067+/-0.133
	ALK	0.556+/-0.046	0.552+/-0.034	0.450+/-0.332	0.124+/-0.063
	BRAF	0.530+/-0.076	0.539+/-0.073	0.172+/-0.137	0.176+/-0.139
	RET	0.553+/-0.075	0.523+/-0.058	0.040+/-0.080	0.067+/-0.133
LUSC	KRAS	0.500+/-0.000	0.500+/-0.000	0.000	0.000
	EGFR	0.500+/-0.000	0.500+/-0.000	0.000	0.000
	ROS1	0.493+/-0.059	0.465+/-0.027	0.000	0.000
	ALK	0.500+/-0.000	0.500+/-0.000	0.000	0.000
	BRAF	0.500+/-0.000	0.500+/-0.000	0.000	0.000
	RET	0.494+/-0.006	0.497+/-0.004	0.000	0.000

Table 5.1: Baseline model results trained with cross-validation on GaussianNB.

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.807+/-0.058	0.694+/-0.033	0.717+/-0.071	0.463+/-0.068
	EGFR	0.855+/-0.049	0.697+/-0.028	0.768+/-0.131	0.420+/-0.053
	ROS1	0.826+/-0.048	0.512+/-0.025	0.200+/-0.400	0.025+/-0.050
	ALK	0.673+/-0.063	0.545+/-0.040	0.250+/-0.224	0.100+/-0.082
	BRAF	0.560+/-0.064	0.516+/-0.023	0.300+/-0.400	0.036+/-0.045
	RET	0.582+/-0.115	0.500+/-0.000	0.000	0.000
LUSC	KRAS	0.675+/-0.189	0.499+/-0.001	0.000	0.000
	EGFR	0.624+/-0.082	0.500+/-0.000	0.000	0.000
	ROS1	0.527+/-0.040	0.498+/-0.002	0.000	0.000
	ALK	0.458+/-0.051	0.500+/-0.000	0.000	0.000
	BRAF	0.459+/-0.065	0.499+/-0.001	0.000	0.000
	RET	0.579+/-0.095	0.499+/-0.001	0.000	0.000

Table 5.2: Baseline model results trained with cross-validation on LightGBM.

Based on the model's performances in general and having a focus on AUC score, GaussianNB has the lowest achievement with *AUCROC* scores between 49,5% and 68%, followed by MLP with values between 45% and 86%, being the best range obtained with LightGBM, between 35% and 84%. The values *Precision* and *Recall* are very different along the different models. GaussianNB cannot predict mutations on LUSC neither LightGBM. While MLP does not provide outstanding performance, it can predict mutations in 8 genes better than the other algorithms. Having in mind these observations, GaussianNB is discarded for model improvements. A conclusion based on this information is also that a more complex algorithm will not necessarily perform better since Neural Networks are considered more complex than Machine Learning algorithms and the performance of LightGBM and MLP is similar on this problem.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.811+/-0.045	0.712+/-0.057	0.601+/-0.086	0.586+/-0.126
	EGFR	0.839+/-0.048	0.788+/-0.029	0.631+/-0.090	0.650+/-0.047
	ROS1	0.604+/-0.110	0.518+/-0.058	0.033+/-0.067	0.067+/-0.133
	ALK	0.597+/-0.079	0.521+/-0.052	0.115+/-0.108	0.119+/-0.109
	BRAF	0.530+/-0.077	0.508+/-0.048	0.135+/-0.127	0.090+/-0.074
	RET	0.530+/-0.041	0.480+/-0.015	0.000	0.000
LUSC	KRAS	0.612+/-0.187	0.487+/-0.009	0.000	0.000
	EGFR	0.496+/-0.192	0.530+/-0.107	0.100+/-0.200	0.100+/-0.200
	ROS1	0.411+/-0.129	0.448+/-0.008	0.000	0.000
	ALK	0.548+/-0.185	0.538+/-0.087	0.107+/-0.137	0.133+/-0.163
	BRAF	0.349+/-0.322	0.470+/-0.018	0.000	0.000
	RET	0.514+/-0.182	0.505+/-0.050	0.050+/-0.100	0.050+/-0.100

Table 5.3: Baseline model results trained with cross-validation on MLP.

5.2 Experiments on multiple Feature Selection techniques

5.2.1 Gene Signature for Feature Selection

Gene Signatures selected from the literature were tested as feature selection and dimensionality reduction. Using the gene Signatures as Feature Selection, the low variance features were not removed because the objective was to verify if the gene signatures were adequate and, if selected the most informative features. In the set of features selected from the study of *Widespread redundancy in -omics profiles of cancer mutation* [21], GPR98, SSPO, XIST, BAGE2, GPR112, C1orf173, BAI3, MALAT1, LPHN3 were removed because these genes were not present on our gene expression. The initial set had 268 features and the final set was left with 259 features.

For the gene signature from the *Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer* [53], a selection of 17 features, the P13 gene was excluded from the group since it was not present in our datasets. The final set has 16 genes selected for non-small lung cancer.

Table 5.4 and Table 5.5 exhibit the results of using the gene signature from the study [53] as features for predicting mutations, trained with LightGBM and MLP, respectively. For LightGBM, the *AUCROC* range is in between 39% and 78%, and the accuracy in the range of 50%. For MLP, the *AUCROC* range is in between 28% and 87%, and the accuracy is in the range of 50%, but almost no mutation was predicted on the test set. LightGBM has a better performance overall because the precision and recall values show that, for some genes, the model can predict the mutation.

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.667+/-0.104	0.596+/-0.085	0.415+/-0.101	0.426+/-0.159
	EGFR	0.768+/-0.085	0.570+/-0.058	0.340+/-0.176	0.245+/-0.085
	ROS1	0.642+/-0.131	0.490+/-0.003	0.000	0.000
	ALK	0.446+/-0.074	0.496+/-0.038	0.067+/-0.133	0.033+/-0.067
	BRAF	0.506+/-0.086	0.509+/-0.031	0.079+/-0.102	0.062+/-0.076
	RET	0.541+/-0.141	0.492+/-0.007	0.000	0.000
LUSC	KRAS	0.612+/-0.366	0.493+/-0.012	0.000	0.000
	EGFR	0.515+/-0.207	0.494+/-0.007	0.000	0.000
	ROS1	0.384+/-0.116	0.487+/-0.008	0.000	0.000
	ALK	0.453+/-0.112	0.492+/-0.015	0.000	0.000
	BRAF	0.487+/-0.282	0.496+/-0.004	0.000	0.000
	RET	0.672+/-0.063	0.513+/-0.045	0.050+/-0.100	0.050+/-0.100

Table 5.4: Results based on the gene signature designed by "Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer" [53] using LightGBM.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.557+/-0.022	0.541+/-0.021	0.395+/-0.078	0.164+/-0.060
	EGFR	0.684+/-0.018	0.495+/-0.003	0.000	0.000
	ROS1	0.746+/-0.023	0.500+/-0.000	0.000	0.000
	ALK	0.431+/-0.039	0.489+/-0.005	0.000	0.000
	BRAF	0.629+/-0.074	0.500+/-0.000	0.000	0.000
	RET	0.508+/-0.049	0.500+/-0.000	0.000	0.000
LUSC	KRAS	0.862+/-0.025	0.499+/-0.002	0.000	0.000
	EGFR	0.409+/-0.027	0.500+/-0.000	0.000	0.000
	ROS1	0.417+/-0.040	0.500+/-0.000	0.000	0.000
	ALK	0.436+/-0.065	0.500+/-0.000	0.000	0.000
	BRAF	0.555+/-0.045	0.500+/-0.000	0.000	0.000
	RET	0.276+/-0.017	0.499+/-0.001	0.000	0.000

Table 5.5: Results based on the gene signature designed by "Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer" [53] using MLP.

As for the gene signature selected from [21], table 5.6 and exhibit the results. Observing the results from LightGBM, AUCROC is in the range between 47% and 84% and the accuracy between 49% and 75%. As for MLP, AUCROC is in the range between 41% and 76% and the accuracy between 49% and 68%. Although the AUROC and Accuracy values are higher using LightGBM, MLP provide is a better predictive model in this situation.

Overall, the results are not impressive, taking into account that these gene signatures were selected for the type of cancer and the same data that we are using.

On the following methods of Feature Selection, low variance features were removed previously, and normalization on data was implemented.

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.836+/-0.014	0.743+/-0.020	0.620+/-0.049	0.610+/-0.030
	EGFR	0.748+/-0.028	0.619+/-0.052	0.523+/-0.147	0.276+/-0.097
	ROS1	0.641+/-0.066	0.498+/-0.002	0.000	0.000
	ALK	0.679+/-0.084	0.546+/-0.040	0.333+/-0.365	0.100+/-0.082
	BRAF	0.710+/-0.056	0.535+/-0.050	0.329+/-0.382	0.091+/-0.100
	RET	0.540+/-0.084	0.499+/-0.001	0.000	0.000
LUSC	KRAS	0.727+/-0.071	0.498+/-0.002	0.000	0.000
	EGFR	0.666+/-0.075	0.518+/-0.041	0.200+/-0.400	0.040+/-0.080
	ROS1	0.589+/-0.042	0.495+/-0.005	0.000	0.000
	ALK	0.518+/-0.104	0.499+/-0.003	0.000	0.000
	BRAF	0.468+/-0.074	0.514+/-0.029	0.200+/-0.400	0.029+/-0.057
	RET	0.641+/-0.120	0.497+/-0.002	0.000	0.000

Table 5.6: Results based on the gene signature designed by "Widespread redundancy in -omics profiles of cancer mutation" [21] using LightGBM.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.740+/-0.023	0.678+/-0.020	0.501+/-0.040	0.528+/-0.026
	EGFR	0.757+/-0.021	0.658+/-0.044	0.404+/-0.081	0.410+/-0.077
	ROS1	0.482+/-0.070	0.495+/-0.004	0.000	0.000
	ALK	0.661+/-0.043	0.574+/-0.003	0.263+/-0.060	0.167+/-0.000
	BRAF	0.731+/-0.074	0.565+/-0.029	0.208+/-0.057	0.182+/-0.057
	RET	0.452+/-0.103	0.494+/-0.004	0.000	0.000
LUSC	KRAS	0.410+/-0.048	0.497+/-0.003	0.000	0.000
	EGFR	0.633+/-0.074	0.498+/-0.003	0.000	0.000
	ROS1	0.514+/-0.048	0.497+/-0.021	0.079+/-0.112	0.033+/-0.044
	ALK	0.624+/-0.036	0.586+/-0.006	0.229+/-0.086	0.200+/-0.000
	BRAF	0.621+/-0.043	0.527+/-0.036	0.400+/-0.490	0.057+/-0.070
	RET	0.645+/-0.067	0.658+/-0.004	0.323+/-0.102	0.333+/-0.000

Table 5.7: Results based on the gene signature designed by "Widespread redundancy in -omics profiles of cancer mutation" [21] trained on MLP.

5.2.2 Feature Selection Results

Results from Feature selection by removing the zero importance features using the *Feature Selector library* are presented in the tables B.1 and B.2, for LightGBM and MLP. Overall, the results from this Feature Selection are better when training on LightGBM. On this algorithm, the *AUCROC* average is 68% and the *Accuracy* values range from 49% to 73%. Training the data on MLP, the *AUCROC* average is 59% and the *Accuracy* values are between 48% to 75%. On predicting mutations, both models fail on half of the models. Although this feature selector provides good performance when compared to the others, something not desirable is that the feature importances are non-deterministic, which means that they are a bit random. The purpose of this work is directly related to helping doctors to make decisions based on these predictions, so we are

looking for a more stable model.

Results of running Extra Trees Classifier can be seen in table B.5 and B.6. For LightGBM, the *AUCROC* values range between 48% and 75%, the *Accuracy* values from 49% to 63% and the model can not predict mutations for 7 genes. While with MLP, the *AUCROC* values range from 47% to 78%, the *Accuracy* values from 49% to 69% and the model can not predict mutations for 5 classes. The performance is slightly better on MLP.

KL Divergence results can be observed in the tables B.3 and B.4, for LightGBM and MLP, respectively. **KullbackLeiblerSelector** from the *kydavra* library was used to implement the KL Divergence. With **KL Divergence**, we aim to get the features with less divergence from the target class, so the minimum divergence is established at 1, and the maximum is in the range of 2 and 3.5. Looking at the results with LightGBM, the range in *AUCROC* is between 48% and 83%, with a mean average of 66% with promising values for *Precision* and *Recall*. Although when compared with the results from the **Feature Selector** usage B.1, all the metrics performances are lower, using KL Divergence, we can predict 2 more labels. Looking at MLP results, the range in *AUCROC* is between 27% and 81%, with a mean average of 49%, the lowest of all the feature selection experiments. Although the *Accuracy*, *Precision*, *Recall* values are not the worst along the experiments, the results with Feature Selector are more promising for this algorithm, and the *AUCROC* values are more linear.

Random Forest Selector returns the feature set ordered by scores. Higher the score, the more informative the prediction of the feature will be. Multiple tests were conducted with MLP to see what range of features would give the best result, respectively, features with a score higher than 0, features with a score higher than 0.001, and the best 100 features. The best results were obtained by using all the features with a score above 0. When comparing the Tables B.7, B.8 and , it is obvious how choosing all the features with importance higher than 0 gives a better performance around all the metrics. These experiments are trained only with MLP. The *AUCROC* range taking into account all the features with a score higher than 0, is 62%, and the accuracy is around 55%, and as for *Precision* and *Recall*, the values are higher than for the other feature selection methods tested. Only for KRAS in LUSC, the model is not capable of predicting any mutated sample. When choosing features with a score higher than 0.001, the *AUCROC* mean is 66%, the *Accuracy* values are around 50%, and the precision and recall values are lower than the first experiment. Even if the *AUCROC* is higher compared to the values on , the model is losing important features for prediction with the lower *Precision* and *Recall* values. When selecting the 100 features with higher scores, the *AUCROC* range is 60%, and the *Accuracy* values are also around 55%, but as for the precision and recall values, only for 4 genes can the model predict mutations. Measuring the values around the 4 metrics, the decision was to use all the features with scores higher than zero.

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.833+/-0.007	0.723+/-0.018	0.645+/-0.038	0.544+/-0.044
	EGFR	0.867+/-0.024	0.725+/-0.042	0.696+/-0.085	0.486+/-0.092
	ROS1	0.805+/-0.058	0.499+/-0.002	0.000	0.000
	ALK	0.656+/-0.080	0.561+/-0.031	0.273+/-0.167	0.133+/-0.067
	BRAF	0.656+/-0.054	0.506+/-0.020	0.200+/-0.400	0.018+/-0.036
	RET	0.616+/-0.119	0.500+/-0.000	0.000	0.000
LUSC	KRAS	0.677+/-0.110	0.499+/-0.001	0.000	0.000
	EGFR	0.651+/-0.114	0.499+/-0.002	0.000	0.000
	ROS1	0.616+/-0.043	0.498+/-0.002	0.000	0.000
	ALK	0.479+/-0.144	0.500+/-0.000	0.000	0.000
	BRAF	0.531+/-0.096	0.498+/-0.003	0.000	0.000
	RET	0.694+/-0.106	0.493+/-0.006	0.000	0.000

Table 5.8: Random Feature Selector results trained on LightGBM. All the features with importance >0 were selected.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.832+/-0.022	0.712+/-0.022	0.512+/-0.038	0.615+/-0.028
	EGFR	0.740+/-0.023	0.671+/-0.050	0.360+/-0.027	0.467+/-0.118
	ROS1	0.611+/-0.058	0.547+/-0.034	0.159+/-0.086	0.125+/-0.079
	ALK	0.638+/-0.059	0.620+/-0.012	0.128+/-0.030	0.333+/-0.000
	BRAF	0.544+/-0.087	0.576+/-0.061	0.255+/-0.137	0.200+/-0.106
	RET	0.649+/-0.026	0.491+/-0.043	0.025+/-0.050	0.040+/-0.080
LUSC	KRAS	0.632+/-0.063	0.486+/-0.003	0.000	0.000
	EGFR	0.712+/-0.025	0.630+/-0.050	0.440+/-0.301	0.280+/-0.098
	ROS1	0.473+/-0.019	0.490+/-0.014	0.077+/-0.067	0.033+/-0.027
	ALK	0.661+/-0.033	0.541+/-0.051	0.102+/-0.085	0.120+/-0.098
	BRAF	0.577+/-0.043	0.500+/-0.027	0.033+/-0.067	0.029+/-0.057
	RET	0.461+/-0.149	0.510+/-0.063	0.020+/-0.040	0.067+/-0.133

Table 5.9: Random Feature Selector results trained on MLP. All the features with importance >0 were selected.

Table 5.8 show the results of applying the Random Forest as a feature Selector, trained with LightGBM. The *AUCROC* range is between 53% and 87% and the *Accuracy* in the range of 49% and 73%. For LUSC, *Precision* and *Recall* values are 0, which means no mutation was predicted for this cancer type. Looking at Table 5.9, the results of training a model on MLP using a **Random Forest Selector** can be analyzed. The *AUCROC* ranges from between 46% and 83% and the accuracy from between 49% and 72%. The model, for each class, can at least predict, for the test data, 1 occurrence of the mutation. The performance for the labels with less imbalance is promising, but for both models, the *Accuracy* is still low. MLP has a better performance on predicting mutations, as we can observe on the Precision and Recall values, although the AUROC is not so high as with LightGBM. Analyzing independently, it could be more promising to design

a model from this point for LUAD with LightGBM and LUSC with MLP, but the goal is to find a unique model that can predict the gene mutations on both LUAD and LUSC.

Focusing on the importance of predicting mutations, and then, *Precision* and *Recall* values being higher than 0, from all the Feature Selection methods, **RF** showed better performance on MLP, and **KL Divergence** better performance on LightGBM, For that reason, dimensionality reduction will only be tested on this model.

5.2.3 Dimensionality Reduction

Dimensionality Reduction will not necessarily mean an improvement in performance on all models, but it will mean an improvement in computing efficiency and interpretability. The next step will be optimizing the parameters, and fewer features will provide faster training. Although some important features can be lost along the way, the trade-off is worth it. In this section, instead of comparing LightGBM with MLP, we will compare the results of the dimensionality reduction instead.

5.2.3.1 LightGBM Model

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.737+/-0.016	0.643+/-0.025	0.476+/-0.060	0.451+/-0.026
	EGFR	0.795+/-0.027	0.695+/-0.022	0.544+/-0.067	0.448+/-0.038
	ROS1	0.746+/-0.011	0.493+/-0.002	0.000	0.000
	ALK	0.660+/-0.107	0.494+/-0.032	0.022+/-0.044	0.033+/-0.067
	BRAF	0.553+/-0.047	0.523+/-0.031	0.140+/-0.116	0.073+/-0.068
	RET	0.795+/-0.046	0.635+/-0.102	0.423+/-0.283	0.280+/-0.204
LUSC	KRAS	0.657+/-0.210	0.532+/-0.067	0.200+/-0.400	0.067+/-0.133
	EGFR	0.556+/-0.097	0.499+/-0.001	0.000	0.000
	ROS1	0.609+/-0.032	0.510+/-0.026	0.233+/-0.244	0.044+/-0.042
	ALK	0.535+/-0.051	0.508+/-0.036	0.029+/-0.057	0.040+/-0.080
	BRAF	0.676+/-0.020	0.497+/-0.001	0.000	0.000
	RET	0.636+/-0.022	0.654+/-0.006	0.262+/-0.127	0.333+/-0.000

Table 5.10: Results from using ANOVA after KL Divergence, trained on LightGBM.

Table 5.10 shows the results of applying ANOVA after selecting the most important features using KL Divergence, trained on LightGBM. Table 5.11 shows the results of applying MRMR to select the best features. The differences are not significant but analyzing Precision with ANOVA, the range of values is between 2,2% and 54%, with a mean average of 19%, and for Recall, the range is between 4% and 45%, with a mean average of 15%. As for MRMR, the Precision values are between 3,3% and 59%, with a mean average of 20%, and for recall, 3,3% and 57%, with a

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.802+/-0.013	0.720+/-0.024	0.587+/-0.039	0.569+/-0.038
	EGFR	0.822+/-0.015	0.695+/-0.021	0.516+/-0.068	0.457+/-0.049
	ROS1	0.697+/-0.066	0.495+/-0.005	0.000	0.000
	ALK	0.676+/-0.044	0.495+/-0.036	0.033+/-0.067	0.033+/-0.067
	BRAF	0.563+/-0.062	0.503+/-0.025	0.090+/-0.111	0.036+/-0.045
	RET	0.794+/-0.045	0.635+/-0.102	0.423+/-0.283	0.280+/-0.204
LUSC	KRAS	0.597+/-0.223	0.499+/-0.002	0.000	0.000
	EGFR	0.667+/-0.072	0.499+/-0.002	0.000	0.000
	ROS1	0.572+/-0.031	0.505+/-0.017	0.207+/-0.194	0.033+/-0.027
	ALK	0.560+/-0.095	0.512+/-0.043	0.100+/-0.200	0.040+/-0.080
	BRAF	0.679+/-0.052	0.511+/-0.030	0.200+/-0.400	0.029+/-0.057
	RET	0.668+/-0.028	0.654+/-0.107	0.223+/-0.123	0.333+/-0.211

Table 5.11: Results from using MRMR after KL Divergence, trained on LightGBM.

mean average of 20%. Besides that, AUCROC gets higher values with a mean average of 67% and, for those reasons, the model using MRMR to select the best features will be the one to optimize.

5.2.3.2 MLP Model

Table 5.12 and 5.13 expose the results of applying dimensionality reduction to our model trained on MLP. When looking at the AUCROC ranges, when using ANOVA, the ranges are between 41% and 84%, and when using MRMR, between 36% and 83%. When comparing accuracy, the mean average with ANOVA is 55%, and when using MRMR is 56%. When looking at Precision and Recall, the average precision and recall with ANOVA is 18% and 14%, respectively, and for MRMR, is 21% and 16%. Based on the average of all metrics, although the difference is not significant, MRMR was the chosen technique.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.837+/-0.012	0.718+/-0.020	0.570+/-0.031	0.579+/-0.026
	EGFR	0.763+/-0.022	0.660+/-0.028	0.438+/-0.038	0.400+/-0.071
	ROS1	0.784+/-0.042	0.506+/-0.023	0.050+/-1.000	0.025+/-0.050
	ALK	0.519+/-0.041	0.602+/-0.037	0.247+/-0.072	0.233+/-0.082
	BRAF	0.615+/-0.040	0.593+/-0.020	0.332+/-0.028	0.218+/-0.045
	RET	0.461+/-0.093	0.495+/-0.003	0.000	0.000
LUSC	KRAS	0.414+/-0.073	0.499+/-0.002	0.000	0.000
	EGFR	0.590+/-0.050	0.556+/-0.047	0.340+/-0.377	0.120+/-0.098
	ROS1	0.425+/-0.011	0.488+/-0.004	0.000	0.000
	ALK	0.628+/-0.096	0.535+/-0.050	0.200+/-0.245	0.080+/-0.098
	BRAF	0.669+/-0.018	0.500+/-0.000	0.000	0.000
	RET	0.441+/-0.134	0.522+/-0.059	0.022+/-0.044	0.067+/-0.133

Table 5.12: Random Forest Selector with ANOVA trained on MLP.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.835+/-0.011	0.716+/-0.026	0.535+/-0.053	0.605+/-0.026
	EGFR	0.826+/-0.014	0.711+/-0.024	0.513+/-0.032	0.495+/-0.057
	ROS1	0.715+/-0.058	0.509+/-0.024	0.067+/-0.133	0.025+/-0.050
	ALK	0.789+/-0.062	0.621+/-0.041	0.304+/-0.111	0.267+/-0.082
	BRAF	0.683+/-0.043	0.621+/-0.050	0.385+/-0.123	0.273+/-0.100
	RET	0.603+/-0.070	0.516 +/- 0.040	0.100+/-0.200	0.040+/-0.080
LUSC	KRAS	0.562+/-0.101	0.492+/-0.004	0.000	0.000
	EGFR	0.731+/-0.091	0.580+/-0.040	0.600+/-0.490	0.120+/-0.098
	ROS1	0.360+/-0.041	0.484+/-0.005	0.000	0.000
	ALK	0.602+/-0.108	0.551+/-0.034	0.082+/-0.046	0.160+/-0.080
	BRAF	0.588+/-0.072	0.500+/-0.000	0.000	0.000
	RET	0.399+/-0.126	0.494+/-0.002	0.000	0.000

Table 5.13: Results from using MRMR after Random Forest Selector, trained on MLP.

For the next step, Optimization, Data Augmentation was applied in an attempt to balance the class weights. The result of applying Data Augmentation using the ADASYN smote technique can be analyzed in the Table B.9, and already an improvement in Accuracy and Recall is noticeable, although the Precision values get worse for some classes, meaning that the model is labeling negative samples as positive.

5.3 Hyperparameter Optimization Results

The list of the optimized parameters trained with LightGBM is available at B.10. The optimized parameters are a result of BayesSearch with 5 cross-validation. BayesSearch is computationally efficient, by selecting the parameters based on previous evaluations instead of trying all the possible values. The results of the final model can be analyzed in 5.14. The AUCROC values are between 56% and 91%, with a mean average of 67%, and the accuracy values are between 48% and 72%. Comparing the results with optimization with the previous results in 5.11, previously, the model could not predict mutations on 3 classes and currently can not predict only for 1 class. The Precision values are lower than in the previous model, but on the Recall, a significant difference is noticeable. Overall, a higher improvement was expected with the optimization, the results are slightly better if we focus the attention on Recall.

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.829+/-0.015	0.717+/-0.021	0.586+/-0.040	0.564+/-0.032
	EGFR	0.785+/-0.019	0.713+/-0.013	0.364+/-0.012	0.581+/-0.036
	ROS1	0.478+/-0.036	0.485+/-0.027	0.047+/-0.005	0.800+/-0.245
	ALK	0.563+/-0.043	0.484+/-0.048	0.031+/-0.027	0.100+/-0.082
	BRAF	0.588+/-0.051	0.496+/-0.019	0.040+/-0.080	0.018+/-0.036
	RET	0.718+/-0.080	0.775+/-0.054	0.281+/-0.112	0.480+/-0.160
LUSC	KRAS	0.914+/-0.041	0.500+/-0.000	0.000	0.000
	EGFR	0.750+/-0.068	0.539+/-0.049	0.300+/-0.400	0.080+/-0.098
	ROS1	0.556+/-0.030	0.556+/-0.030	0.148+/-0.014	0.589+/-0.044
	ALK	0.606+/-0.089	0.518+/-0.045	0.043+/-0.035	0.120+/-0.098
	BRAF	0.640+/-0.065	0.601+/-0.040	0.104+/-0.027	0.371+/-0.070
	RET	0.687+/-0.061	0.662+/-0.042	0.042+/-0.011	0.667+/-0.000

Table 5.14: Final model results, trained on LightGBM.

Tuned Parameters for MLP								
		hidden_layer_sizes	activation	solver	alpha	learning_rate	batch_size	early_stopping
LUAD	KRAS	(100,)	relu	lbfgs	0.0001	constant	auto	True
	EGFR	(50,)	logistic	sgd	0.0001	constant	auto	False
	ROS1	(50, 100, 50,)	relu	sgd	0.0001	adaptive	auto	True
	ALK	(50,)	tanh	sgd	0.05	constant	128	False
	BRAF	(50,)	logistic	sgd	0.0001	constant	128	False
	RET	(50, 50, 50,)	relu	sgd	0.0001	adaptive	512	True
LUSC	KRAS	(50, 100, 50,)	relu	sgd	0.0001	constant	128	True
	EGFR	(50, 50, 50,)	relu	sgd	0.0001	adaptive	512	True
	ROS1	(50, 100, 50,)	identity	sgd	0.0001	invscaling	512	False
	ALK	(50,)	identity	sgd	0.05	constant	auto	False
	BRAF	(50,)	identity	lbfgs	0.05	identity	128	True
	RET	(50,)	logistic	sgd	0.0001	constant	128	False

Table 5.15: Tuned Parameters for MLP.

For MLP, the optimized parameters are displayed in Table 5.15, organized by mutation. The parameter optimization is a result of a Grid Search with 5 interactions, and the final parameters are the ones that offered the best AUCROC score. GridSearch exhaustively explores all the combinations between the hyperparameters by order. Comparing the final model 5.16 with the model before optimization 5.13 the improvements are significant, especially when looking at the Precision and Recall values. The model can predict mutations for all the genes, which is a significant improvement from all the models we have seen, and behaves relatively well for the classes with less imbalance. Overall, the AUCROC improved in all classes, the values ranging between 52% and 93%, and the average mean 73%. The accuracy improved slightly but, for the unbalanced classes, is still in the range between 53% and 61%. By analyzing the tuned parameters in 5.15, we can conclude that Stochastic Gradient Descent (SGD) is the best solver for the task, the alpha is set at 0.0001.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.905+/-0.038	0.825+/-0.033	0.633+/-0.052	0.849+/-0.054
	EGFR	0.931+/-0.030	0.847+/-0.062	0.531+/-0.064	0.827+/-0.120
	ROS1	0.695+/-0.153	0.537+/-0.090	0.103+/-0.122	0.267+/-0.249
	ALK	0.761+/-0.109	0.650+/-0.075	0.279+/-0.142	0.467+/-0.236
	BRAF	0.703+/-0.150	0.693+/-0.108	0.253+/-0.090	0.614+/-0.154
	RET	0.460+/-0.142	0.560+/-0.123	0.091+/-0.090	0.317+/-0.213
LUSC	KRAS	0.524+/-0.415	0.634+/-0.206	0.167+/-0.211	0.300+/-0.400
	EGFR	0.648+/-0.115	0.555+/-0.083	0.030+/-0.037	0.200+/-0.245
	ROS1	0.772+/-0.099	0.583+/-0.070	0.104+/-0.030	0.840+/-0.196
	ALK	0.805+/-0.105	0.604+/-0.118	0.149+/-0.140	0.267+/-0.249
	BRAF	0.850+/-0.178	0.59+/-0.202	0.133+/-0.267	0.200+/-0.400
	RET	0.804+/-0.040	0.597+/-0.103	0.171+/-0.114	0.300+/-0.187

Table 5.16: Final model results, trained on MLP.

By using SHAP [17], it is possible to analyze single predictions or all the predictions on the test set. SHAP values break down a prediction by showing the impact of each feature, by measuring the difference between a baseline model prediction and the actual prediction. To understand the values and their significance, we will analyze some plots for single predictions and the force plot that explains all the projections. For the purpose of the example, the plots will only display the 10 more relevant features, in the appendix, the images with all the feature scores can be visualized.

Starting by explaining single predictions, in figure 5.1, it is possible to analyze an explanation for a single prediction. First, the base value is close to 0, which means that the average prediction of the model on the training data is 0, meaning that the patient does not have the gene mutated. The original SHAP paper [39] explains the base value as being "the value that would be predicted if we did not know any features for the current output." In this specific example, the model predicted that the probability of having a mutation on EGFR is 70%. The red and blue bars represent the influence of the RNAseq genes on this prediction. Genes represented by the red bar, push the prediction to a higher score, genes represented in blue have a negative influence on the prediction, pushing it to a lower prediction. In this case, IBSP is the gene with a higher impact on the prediction, and CYSLTR1 has a more meaningful effect on decreasing the prediction.

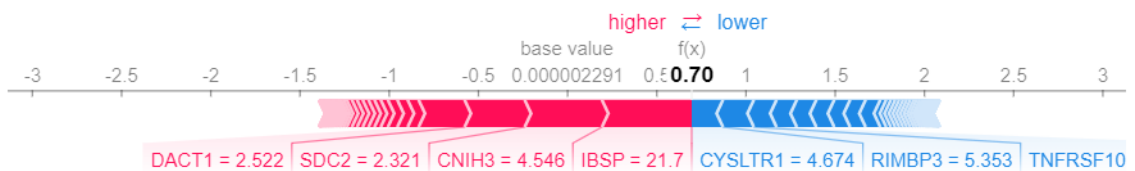


Figure 5.1: Force plot for the prediction for the 6th patient on the test dataset.

The waterfall plot 5.2 represents the feature impacts on the final model prediction, as the force plot, but in a more effortless and cleaner way to visualize. $E[f(x)]$ represents the base value, and $f(x)$ represents the actual prediction. The values on the x_axis are the raw counts of the gene presence in this patient. The values in between are the SHAP values that measure the impact on the prediction, for instance, the gene IBSP increased the prediction value by 0.5.

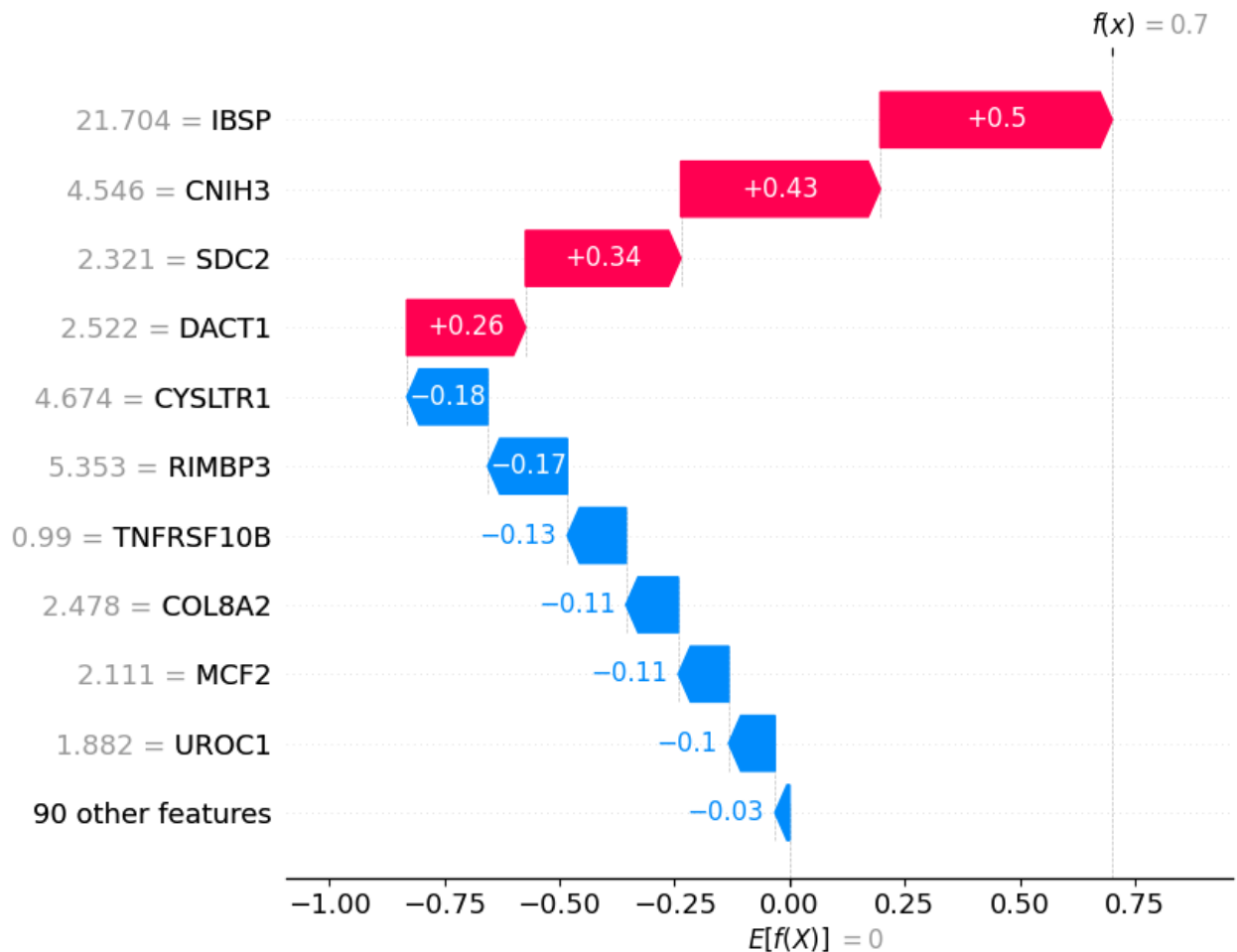


Figure 5.2: Waterfall plot for the prediction for the 6th patient on the test dataset.

By looking at the beeswarm plot in figure 5.3, it is possible to see the relative impact of all the features in the entire dataset. The average SHAP value of the feature organizes the x_axis, the higher the value, the more up in the top the feature will be. In this case, higher values of the 10 features presented will result in higher SHAP values, therefore, a higher prediction score.

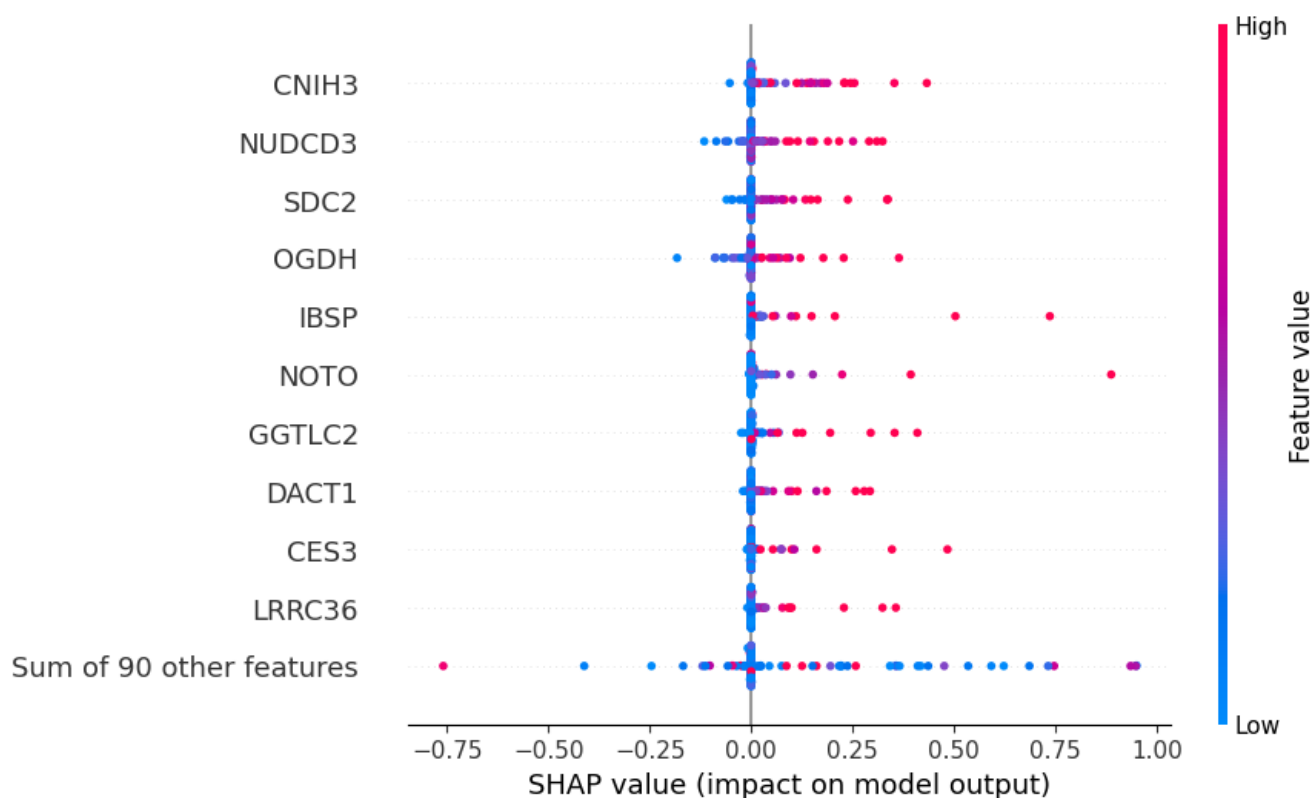


Figure 5.3: Beeswarm plot for the prediction for the 6th patient on the test dataset.

All the previous plots explain the feature importances of one single prediction. To show how the model makes the predictions in general, the SHAP values are aggregated into a stacked force plot, which consists on a combination of all force plots. The stacked forced plot is interactive, and it is possible to visualize the feature combinations between the genes. In the image 5.4 it is possible to visualize the feature importances for the prediction of the patient number 136. The value predicted is 0.1447, meaning that for this patient will be predicted no mutation. The red and blue features can be interpreted was in a single force plot mentioned before.

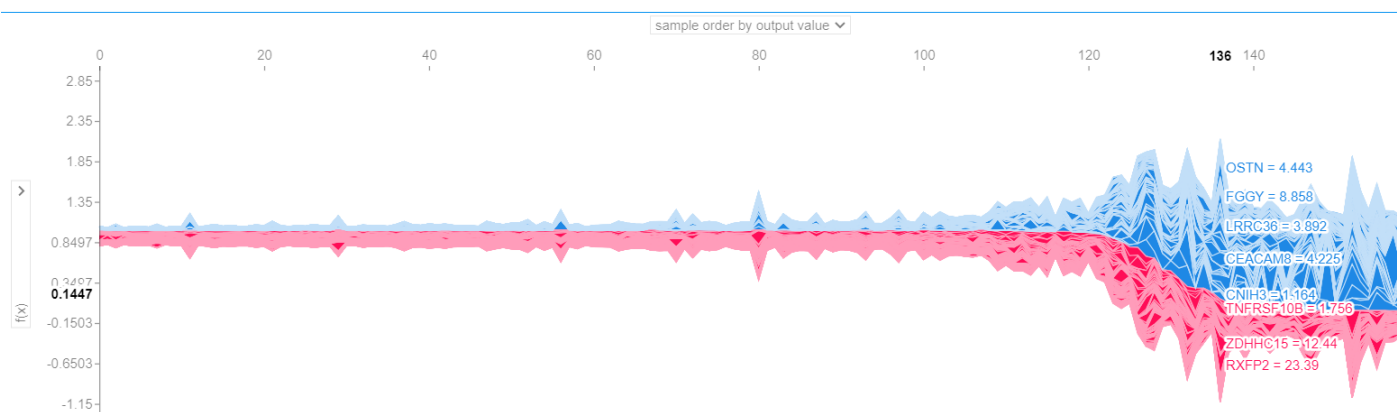


Figure 5.4: Stacked forced plot to explain all predictions on the test dataset.

5.4 Multi-Class Approach

For the Multi-Class model, an approach of One vs Rest was chosen. This technique turns the Multi-Class problem into a binary class model for each class. In our case, the classification is divided into cancer types. For each cancer type, we aim to predict 6 gene mutations and for that reason, the multiclass approach can be divided into 6 binary classification problems. For example, to predict KRAS, a binary classification model KRAS vs [EGFR, ROS1, ALK, BRAF, RET] will be created. The first experiment was based on the methodology used on the binary classification that offered the best performance. For multiclass, the same metrics were used but with an additional parameter *average*, suitable for multiclass classification problems. The *average* parameter was set to *macro*, meaning that the precision and recall values are calculated for each class independently, according to their occurrences. The final value is a mean average of all the precision and recall values of the different labels.

MLP Baseline				
	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	0.550+/-0.020	0.156+/-0.021	0.167+/-0.070	0.156+/-0.021
LUSC	0.520+/-0.027	0.133+/-0.006	0.122+/-0.015	0.133+/-0.006

Table 5.17: Multiclass problem trained on MLP.

Random Forest Selector				
	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	0.592+/-0.019	0.289+/-0.028	0.290+/-0.027	0.289+/-0.028
LUSC	0.550+/-0.020	0.156+/-0.021	0.167+/-0.070	0.156+/-0.021

Table 5.18: Multiclass problem trained on MLP using Random Forest Selector.

Random Forest Selector + MRMR				
	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	0.649+/-0.014	0.291+/-0.027	0.304+/-0.052	0.291+/-0.027
LUSC	0.578+/-0.025	0.145+/-0.010	0.137+/-0.009	0.145+/-0.010

Table 5.19: Multiclass problem trained on MLP using Random Forest Selector with MRMR and ADASYN.

By analyzing the previous tables, an improvement in AUCROC can be observed along the way, with 52% and 55% by removing only the low variance features for LUAD and LUSC, respectively, with 55% and 59% using only the Random Forest to select features with importance higher than 0 and 58% and 65% by applying a dimensionality reduction to the best 100 features to the set selected by Random Forest. As for the overall Accuracy, Precision and Recall values,

an improvement is noticed on both datasets when using the Random Forest selector, but when the subset is reduced, the performance is lower on LUSC, supporting that the relevant features are not the same for different cancer driver genes and by reducing the subset, we are eliminating relevant information.

In an attempt to improve performance, a Multi-Class problem was approached, but the results were not satisfactory. One possible explanation is that the relevant features are not the same for all classes, therefore, it is hard to select a subset of features and the Multi-Class is not appropriate for the problem. To support this assumption, figure 5.5 and figure 5.6 show the 20 most relevant features for predicting KRAS and EGFR on LUAD, respectively, with the final model trained on MLP.

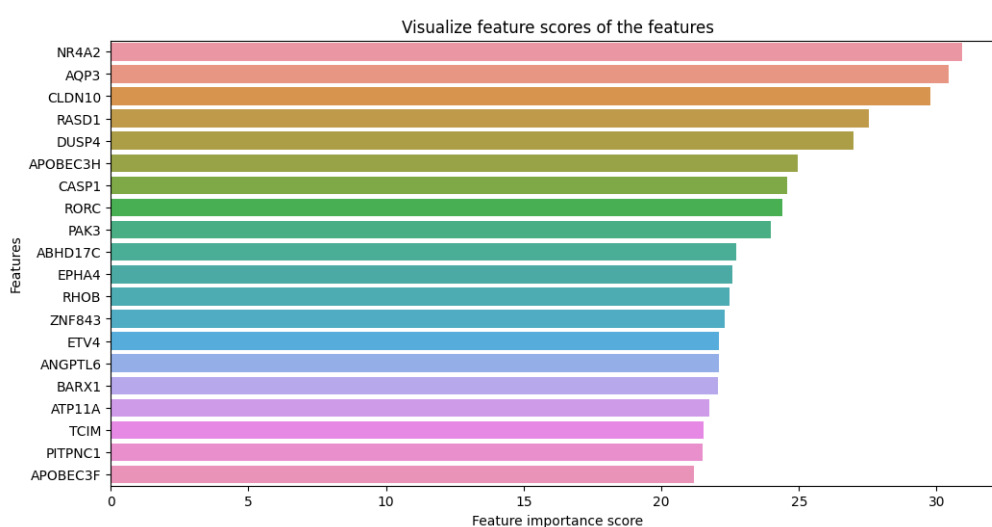


Figure 5.5: The set of the 20 most relevant features for KRAS, for the final model trained on MLP.

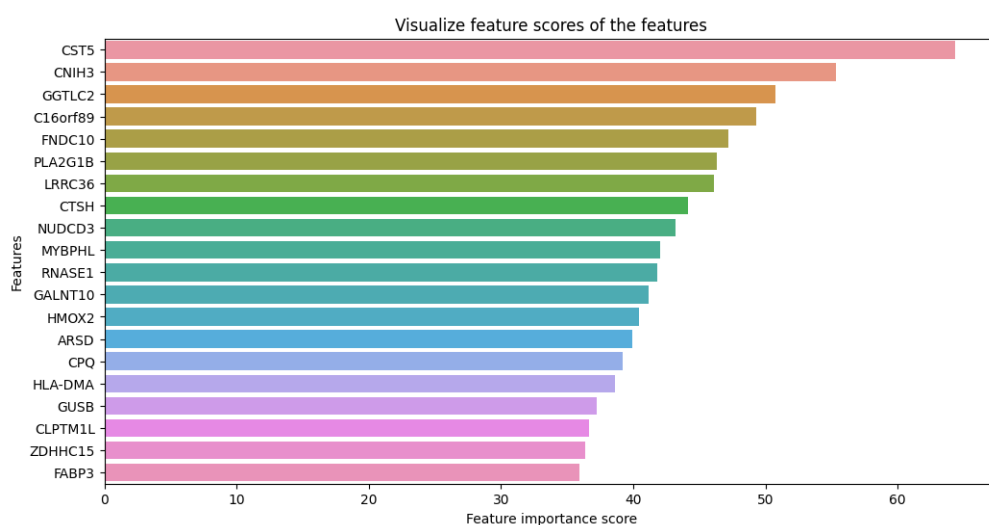


Figure 5.6: The set of the 20 most relevant features for EGFR, for the final model trained on MLP.

As it is possible to visualize, the gene set for the two predictions is entirely different, leading to the conclusion that a multiclass problem is not suitable for this task having in account the imbalance of the majority of the classes. For the less imbalanced classes, KRAS and EGFR on LUAD, the prediction is more promising, leading to another assumption that by collecting more data to balance the classes, the performance would be better.

5.5 Discussion

- **The best DL model performance was acquired using Random Forest + MRMR feature selection.** Training on MLP, the prediction with higher performance is on the gene EGFR, acquiring an AUCROC of 93,1%, an Accuracy score of 84,7%, a Precision of 53,1%, and a Recall of 82,7%.
- **The best ML model performance was acquired using KL Divergence + MRMR feature selection.** Trained on LightGBM, KRAS was the class with higher performance prediction. AUCROC score is 82,9%, followed by an Accuracy score of 71,7%, a Precision of 58,6%, and a Recall of 56,4%.

The following discussion is on the performance based on the selected features of the different techniques that were experimented, analyzing each class independently. This discussion is helpful for future work.

By focusing on LUAD:

- **For KRAS, the features selected on "*Widespread redundancy in -omics profiles of cancer mutation*" [21] offer the best performance among the Feature Selection techniques experimented.** Training on LightGBM, AUCROC is 83,6% with an Accuracy of 74,3%, a Precision of 62% and a Recall of 61%.
- **For EGFR, the feature set selected by removing Feature Importance features based on the *Feature Selector* is optimal when trained with MLP.** AUCROC performance is 81,4%, with the Accuracy being 75%, a Precision of 50%, and a Recall of 49%.
- **For ROS1, the feature set selected by Random Forest Feature Selector + ANOVA is optimal when trained with MLP.** AUCROC performance is 75,5%, with the Accuracy being 76,1%, a Precision of 11,6%, and a Recall of 87,5%.
- **For ALK, the features set selected by removing Feature Importance features based on the *Feature Selector* is optimal when trained with LightGBM.** AUCROC performance is 65,1%, with the Accuracy being 57,7%, a Precision of 46,7%, and a Recall of 16,7%.
- **For BRAF, the optimal feature set is selected by using Random Forest Selector + MRMR, trained on MLP.** The AUCROC score is 68,3%, with the Accuracy being 62,1%, a Precision of 38,5%, and a Recall of 27,3%.

- **For RET, the optimal feature set is selected by using KL Divergence + ANOVA, trained with LightGBM.** The AUCROC score is 79,5%, with the Accuracy being 63,5%, a Precision of 42,3%, and a Recall of 28%.

By examining LUSC:

- **The best subset of features selected for KRAS is given by Random Forest Feature Selector + ANOVA, trained on MLP.** AUCROC is 62,8% with an Accuracy of 63,2%, a Precision of 60% and a Recall of 26,7%.
- **For EGFR, the optimal set is obtained by removing Feature Importance features based on the *Feature Selector* when trained on MLP.** AUCROC is 73,7% with an Accuracy of 63,2%, a Precision of 47% and a Recall of 28%.
- **For ROS1, the optimal set is selected by KL Divergence + ANOVA when trained on LightGBM.** AUCROC is 60,9% with an Accuracy of 51%, a Precision of 23,3% and a Recall of 4,4%.
- **For ALK, the optimal set is selected by Random Forest + ANOVA, when trained on MLP.** AUCROC is 63,8% with an Accuracy of 61,5%, a Precision of 7,7% and a Recall of 40%.
- **For BRAF, the optimal set is the gene signature selected in "*Widespread redundancy in -omics pro les of cancer mutation*" [21], when trained on MLP.** AUCROC is 62,2% with an Accuracy of 52,7%, a Precision of 40% and a Recall of 33,3%.
- **For RET, the set features offering the best performance along all the metrics is selected by KL Divergence + MRMR, trained on LightGBM.** AUCROC is 66,8% with an Accuracy of 65,4%, a Precision of 22,3% and a Recall of 33,3%.

The multiclass approach results prove how distinct the subset of genes influences different mutations. The multiclass problem was built on the discoveries made along the binary classification model, and the model behaves poorly in performance when compared to the final binary model. The best performance for LUSC is when only using Random Forest Selector, with an AUCROC of 55% with an average Accuracy of 15,6%, an average Precision of 16,7% and an average Recall of 15,6%, for each class. For LUAD, the best performance is given by using Random Forest + MRMR to select the best features, with an AUCROC of 64,9% with an average Accuracy of 21,9%, an average Precision of 30,4% and an average Recall of 29,1%, for each class. The results suggest that it is challenging to find a subset of ideal features for all the mutations. The conclusion is that the problem of predicting mutation should be addressed independently based on the mutation we are looking to predict.

5.6 Summary

The first and fundamental step was to train the dataset with simple baseline models based on different algorithms to set a benchmark for assessing performance. The next step is to conduct an exhaustive approach to the binary classification problem, testing methods mentioned in literature and commonly used for designing Machine and Deep Learning models. After selecting the Feature Selection method that provides the best performance, dimensionality reduction is implemented to improve computing efficiency. On the best final models for both ML and DL, optimization is conducted in order to select the best hyperparameters for the task. Based on the results of the binary problem, a multiclass approach is carried on. The best performance is for a binary classification model that selects the best features based on the importance provided by Random Forest followed by a dimensionality reduction to the best 100 features using MRMR, trained on MLP with the optimized parameters found by a Grid Search approach.

Chapter 6

Conclusions

Lung cancer is the cancer type that affects most cancer patients, and it has a mortality rate of over 35%. NSCLC is generally diagnosed at an advanced stage, lowering the survival chances of the patient. Biomarkers can detail the information of a type of cancer, helping address a targeted therapy. Not all lung tumors have the same identity or are originated in the same proteins. KRAS, EGFR, ROS1, BRAF, RET, and ALK are mutation cancer drivers. Cancer inhibitors have been designed for each of these genes, but after some time, the patient gains resistance. In order to create more efficient biomarkers and, therefore, more potent targeted inhibitors, it is essential to track down the gene expression of the patients and the genes that are influencing the tumors to grow. This work attempts to help doctors to address this issue by designing interpretable ML and DL models for predicting mutations based on gene expression data. Using data from TCGA, during this investigation, multiple models using different feature selection techniques were experimented with by incorporating domain knowledge or not. MLP lead to the best results with the highest AUCROC of 93% and Accuracy of 85% on predicting EGFR on LUAD, using Random Forest + MRMR as feature selection, using ADASYN to balance the data and Grid Search to tune the hyperparameters. When analyzing the state-of-art for predicting mutations based on gene expression data, neural network-based solutions are not being explored. By using SHAP to explain the outcomes of the MLP-based model, this work mitigates the assumption that the DL results are challenging to interpret. By approaching the multiclass problem, this work proves how the gene expression of patients with different cancer-driven genes is so distinct and should be addressed separately. Continuous work in this field, along with DL interpretable models, can be promising, and this experiment provides strong insights into how to explore this field. The model's performance along all genes was not the desired. Some limitations, listed below, can be fulfilled in order to improve the models performance.

- **Collect more data.** One limitation of training the models used in this work is the data imbalance along the labels. Future work should focus on gathering more mutated samples to provide knowledge to the model. Another way to mitigate the imbalance would be to use more sophisticated Data Augmentation techniques such as Variational AutoEncoders.

- **Address the multilabel problem.** In this study, occurrences of samples with multiple mutations were observed. The data were insufficient to address this problem. Still, the multilabel model can be beneficial to detect patterns in patients with different mutated genes and discover a general group of genes influenced.
- **Combine domain knowledge with Feature Selection techniques.** In this study, domain knowledge and Feature Selection techniques were used separately to select the optimal subset of features. Combining both can lead to better predictions.
- **Explore more complex NN.** Previous study induces the belief that simpler architectures work better when it comes to biological data, although what happened in this experiment is that when the complexity was increased, the performance was higher. For that reason, we suggest future researchers to introduce more complexity to the models and base their research on more complex Machine Learning algorithms, such as Deep Learning models.

Bibliography

- [1] #127 chromosomes, dna, genes and alleles | biology notes for igcse 2014 & 2024. <http://igbiologyy.blogspot.com/2014/03/chromosomes-dna-genes-and-alleles.html>. (Accessed on 06/28/2023).
- [2] Accuracy vs balanced accuracy, which is the best metric? <https://stephenallwright.com/accuracy-vs-balanced-accuracy/>. (Accessed on 06/29/2023).
- [3] Alk and lung cancer | american lung association. <https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/symptoms-diagnosis/biomarker-testing/alk-lung-cancer>. (Accessed on 06/19/2023).
- [4] Braf and lung cancer | american lung association. <https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/symptoms-diagnosis/biomarker-testing/braf>. (Accessed on 06/19/2023).
- [5] Feature engineering: Scaling, normalization and standardization. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/#What_Is_Normalization? (Accessed on 06/28/2023).
- [6] Feature selection using random forest | by akash dubey | towards data science. <https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f>. (Accessed on 06/27/2023).
- [7] Gaussian naive bayes: What you need to know? | upgrad blog. <https://www.upgrad.com/blog/gaussian-naive-bayes/>. (Accessed on 06/28/2023).
- [8] Gradient boosted decision trees [guide]: a conceptual explanation. <https://neptune.ai/blog/gradient-boosted-decision-trees-guide>. (Accessed on 06/28/2023).

- [9] Lightgbm (light gradient boosting machine) - geeksforgeeks. <https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/>. (Accessed on 06/28/2023).
- [10] Multilayer perceptron explained with a real-life example and python code: Sentiment analysis | by carolina bento | towards data science. <https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-se> (Accessed on 06/28/2023).
- [11] Parameters tuning — lightgbm 3.3.5.99 documentation. <https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html>. (Accessed on 06/28/2023).
- [12] Ros1 and lung cancer | american lung association. <https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/symptoms-diagnosis/biomarker-testing/ros1>. (Accessed on 06/19/2023).
- [13] smazzanti/mrmr: mrmr (minimum-redundancy-maximum-relevance) for automatic feature selection at scale. <https://github.com/smazzanti/mrmr>. (Accessed on 06/27/2023).
- [14] Targeted cancer therapy | targeted drug therapy for cancer. <https://www.cancer.org/cancer/managing-cancer/treatment-types/targeted-therapy/what-is.html>. (Accessed on 06/22/2023).
- [15] Targeted therapy for cancer - nci. <https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies>. (Accessed on 06/22/2023).
- [16] Understanding auc - roc curve | by sarang narkhede | towards data science. <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>. (Accessed on 06/29/2023).
- [17] Welcome to the shap documentation — shap latest documentation. <https://shap.readthedocs.io/en/latest/index.html>. (Accessed on 07/04/2023).
- [18] Willkoehrsen/feature-selector: Feature selector is a tool for dimensionality reduction of machine learning datasets. <https://github.com/WillKoehrsen/feature-selector>. (Accessed on 06/27/2023).
- [19] TaeJin Ahn, Taewan Goo, Chan-hee Lee, SungMin Kim, Kyullhee Han, Sangick Park, and Taesung Park. Deep learning-based identification of cancer or normal tissue using gene expression data. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1748–1752. IEEE, 2018.

- [20] Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.
- [21] Jake Crawford, Brock C Christensen, Maria Chikina, and Casey S Greene. Widespread redundancy in-omics profiles of cancer mutation states. *Genome Biology*, 23(1):1–24, 2022.
- [22] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.
- [23] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 37(4):1633–1650, 2021.
- [24] Grant Hayes Frazier. Defusing a ticking time bomb: The complicated considerations underlying compulsory human genetic editing. *Hastings Sci. & Tech. LJ*, 10:39, 2019.
- [25] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.
- [26] Hajar Hakkoum, Ibtissam Abnane, and Ali Idri. Evaluating interpretability of multilayer perceptron and support vector machines for breast cancer classification. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE, 2022.
- [27] Jason C Hsu, Phung-Anh Nguyen, Phan Thanh Phuc, Tsai-Chih Lo, Min-Huei Hsu, Min-Shu Hsieh, Nguyen Quoc Khanh Le, Chi-Tsun Cheng, Tzu-Hao Chang, and Cheng-Yu Chen. Development and validation of novel deep-learning models using multiple data types for lung cancer survival. *Cancers*, 14(22):5562, 2022.
- [28] Zhi Huang, Travis S Johnson, Zhi Han, Bryan Helm, Sha Cao, Chi Zhang, Paul Salama, Maher Rizkalla, Christina Y Yu, Jun Cheng, et al. Deep learning-based cancer survival prognosis from rna-seq data: approaches and evaluations. *BMC medical genomics*, 13(5):1–12, 2020.
- [29] Kentaro Inamura. Lung cancer: understanding its molecular pathology and the 2015 who classification. *Frontiers in oncology*, 7:193, 2017.
- [30] National Cancer Institute. Genomic Data Commons. Available at: <https://gdc.cancer.gov/>. [Last accessed: 2023/12/10].
- [31] National Cancer Institute. Metastasis. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis>. [Last accessed: 2023/01/18].

- [32] National Cancer Institute. The Cancer Genome Atlas. Available at: <https://portal.gdc.cancer.gov/>. [Last accessed: 2023/12/10].
- [33] National Cancer Institute. What Is Cancer? <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>, 2021. [Last accessed: 2023/01/18].
- [34] Jun Kang, Ahwon Lee, and Youn Soo Lee. Prediction of pik3ca mutations from cancer gene expression data. *Plos one*, 15(11):e0241514, 2020.
- [35] Niki Karachaliou, Clara Mayo, Carlota Costa, Ignacio Magrí, Ana Gimenez-Capitan, Miguel Angel Molina-Vila, and Rafael Rosell. Kras mutations in lung cancer. *Clinical lung cancer*, 14(3):205–214, 2013.
- [36] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.
- [37] Suli Liu and Wu Yao. Prediction of lung cancer using gene expression and deep learning with kl divergence gene selection. *BMC bioinformatics*, 23(1):175, 2022.
- [38] Laurence Loewe. Genetic Mutation. Available at: <https://www.nature.com/scitable/topicpage/genetic-mutation-1127/>. [Last accessed: 2023/01/18].
- [39] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [40] Samuel Marguerat and Jürg Bähler. Rna-seq: from technology to biology. *Cellular and molecular life sciences*, 67(4):569–579, 2010.
- [41] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.
- [42] Daniele Mercatelli, Forest Ray, and Federico M Giorgi. Pan-cancer and single-cell modeling of genomic alterations through gene expression. *Frontiers in genetics*, 10:671, 2019.
- [43] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011.
- [44] Medha Pandey and M Michael Gromiha. Predicting potential residues associated with lung cancer using deep neural network. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 822:111737, 2021.

- [45] Bernardo Ramos, Tania Pereira, João Moranguinho, Joana Morgado, José Luis Costa, and Hélder P Oliveira. An interpretable approach for lung cancer prediction and subtype classification using gene expression. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1707–1710. IEEE, 2021.
- [46] Laiqa Rukhsar, Waqas Haider Bangyal, Muhammad Sadiq Ali Khan, Ag Asri Ag Ibrahim, Kashif Nisar, and Danda B Rawat. Analyzing rna-seq gene expression data using deep learning approaches for cancer classification. *Applied Sciences*, 12(4):1850, 2022.
- [47] Chiara Scoccianti, Aurélien Vesin, Ghislaine Martel, Magali Olivier, Elisabeth Brambilla, Jean-François Timsit, Luca Tavecchio, Christian Brambilla, John K Field, Pierre Hainaut, et al. Prognostic value of tp53, kras and egfr mutations in nonsmall cell lung cancer: the euelc cohort. *European Respiratory Journal*, 40(1):177–184, 2012.
- [48] Kien Wei Siah, Sean Khozin, Chi Heem Wong, and Andrew W Lo. Machine-learning and stochastic tumor growth models for predicting outcomes in patients with advanced non-small-cell lung cancer. *JCO clinical cancer informatics*, 1:1–11, 2019.
- [49] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.
- [50] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.
- [51] Gregory P Way, Francisco Sanchez-Vega, Konnor La, Joshua Armenia, Walid K Chatila, Augustin Luna, Chris Sander, Andrew D Cherniack, Marco Mina, Giovanni Ciriello, et al. Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas. *Cell reports*, 23(1):172–180, 2018.
- [52] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, 153:1–9, 2018.
- [53] Xiucan Ye, Weihang Zhang, and Tetsuya Sakurai. Adaptive unsupervised feature learning for gene signature identification in non-small-cell lung cancer. *IEEE Access*, 8:154354–154362, 2020.
- [54] Zhong-Si Zhang, Fei Xu, Han-Jing Jiang, and Zhan-Heng Chen. Prognostic prediction for non-small-cell lung cancer based on deep neural network and multimodal data. volume pt.III, pages 549 – 60, Berlin, Germany, 2021//.

Appendix A

Multiple occurrence of mutations Tables

Mutations	Occurences
KRAS, ALK	6
KRAS, RET	5
ROS1, ALK	5
EGFR, BRAF	3
ROS1, ALK, BRAF	2
KRAS, ROS1	2
ALK, BRAF	2
EGFR, ALK	2
ROS1, RET	1
KRAS, BRAF	1
ROS1, BRAF	1
RET, ALK	1
RET, BRAF	1
KRAS, ROS1, BRAF	1
EGFR, RET, BRAF	1

Table A.1: Table of multilabel mutations for LUAD.

Mutations	Occurences
ROS1, ALK	2
EGFR, ROS1	2
ALK, BRAF	1
RET, BRAF	1
RET, ALK	1
ROS1, BRAF	1
ROS1, RET	1
EGFR, ALK	1

Table A.2: Table of multilabel mutations for LUSC.

Appendix B

Feature Selection

B.0.1 Feature Selector Library

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.838+/-0.015	0.710+/-0.014	0.626+/-0.041	0.523+/-0.042
	EGFR	0.863+/-0.016	0.728+/-0.028	0.722+/-0.090	0.486+/-0.047
	ROS1	0.778+/-0.052	0.497+/-0.002	0.000	0.000
	ALK	0.651+/-0.057	0.577+/-0.004	0.467+/-0.282	0.167+/-0.000
	BRAF	0.607+/-0.067	0.520+/-0.040	0.250+/-0.387	0.055+/-0.073
	RET	0.669+/-0.079	0.498+/-0.003	0.000	0.000
LUSC	KRAS	0.788+/-0.129	0.497+/-0.003	0.000	0.000
	EGFR	0.719+/-0.105	0.534+/-0.048	0.133+/-0.163	0.080+/-0.098
	ROS1	0.583+/-0.052	0.508+/-0.012	0.167+/-0.211	0.022+/-0.027
	ALK	0.558+/-0.058	0.499+/-0.003	0.000	0.000
	BRAF	0.454+/-0.061	0.500+/-0.000	0.000	0.000
	RET	0.664+/-0.045	0.493+/-0.005	0.000	0.000

Table B.1: Results from removing the features with zero importance selected by the feature selector library, trained on LightGBM.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.819+/-0.013	0.715+/-0.016	0.500+/-0.014	0.636+/-0.041
	EGFR	0.814+/-0.021	0.750+/-0.043	0.500+/-0.054	0.590+/-0.083
	ROS1	0.554+/-0.042	0.553+/-0.004	0.297+/-0.113	0.125+/-0.000
	ALK	0.425+/-0.017	0.574+/-0.035	0.132+/-0.046	0.200+/-0.067
	BRAF	0.517+/-0.063	0.538+/-0.020	0.220+/-0.093	0.109+/-0.036
	RET	0.591+/-0.033	0.484+/-0.004	0.000	0.000
LUSC	KRAS	0.605+/-0.071	0.500+/-0.000	0.000	0.000
	EGFR	0.737+/-0.041	0.632+/-0.045	0.470+/-0.267	0.280+/-0.098
	ROS1	0.571+/-0.045	0.491+/-0.005	0.000	0.000
	ALK	0.571+/-0.045	0.491+/-0.005	0.000	0.000
	BRAF	0.484+/-0.054	0.500+/-0.000	0.000	0.000
	RET	0.435+/-0.080	0.49+/-0.004	0.000	0.000

Table B.2: Results from removing the features with zero importance selected by the feature selector library, trained on MLP.

B.0.2 KL Divergence

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.818+/-0.030	0.689+/-0.021	0.600+/-0.030	0.482+/-0.050
	EGFR	0.830+/-0.021	0.687+/-0.050	0.702+/-0.112	0.400+/-0.098
	ROS1	0.819+/-0.037	0.509+/-0.024	0.067+/-0.133	0.025+/-0.050
	ALK	0.676+/-0.044	0.495+/-0.036	0.033+/-0.067	0.033+/-0.067
	BRAF	0.484+/-0.067	0.503+/-0.016	0.050+/-0.100	0.018+/-0.036
	RET	0.792+/-0.050	0.617+/-0.116	0.383+/-0.323	0.240+/-0.233
LUSC	KRAS	0.700+/-0.240	0.533+/-0.067	0.200+/-0.400	0.067+/-0.133
	EGFR	0.579+/-0.158	0.499+/-0.001	0.000	0.000
	ROS1	0.568+/-0.030	0.499+/-0.002	0.000	0.000
	ALK	0.611+/-0.115	0.497+/-0.003	0.000	0.000
	BRAF	0.529+/-0.097	0.513+/-0.028	0.100+/-0.200	0.029+/-0.057
	RET	0.530+/-0.092	0.499+/-0.001	0.000	0.000

Table B.3: Results from using KL Divergence for feature selection, trained on LightGBM.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.815+/-0.022	0.724+/-0.022	0.559+/-0.048	0.610+/-0.077
	EGFR	0.803+/-0.028	0.733+/-0.030	0.418+/-0.021	0.590+/-0.065
	ROS1	0.376+/-0.041	0.493+/-0.002	0.000	0.000
	ALK	0.635+/-0.041	0.555+/-0.032	0.173+/-0.092	0.133+/-0.067
	BRAF	0.544+/-0.045	0.502+/-0.018	0.054+/-0.066	0.036+/-0.045
	RET	0.270+/-0.111	0.499+/-0.001	0.000	0.000
LUSC	KRAS	0.068+/-0.029	0.500+/-0.000	0.000	0.000
	EGFR	0.330+/-0.072	0.511+/-0.038	0.040+/-0.080	0.040+/-0.080
	ROS1	0.495+/-0.017	0.497+/-0.026	0.119+/-0.112	0.044+/-0.042
	ALK	0.665+/-0.043	0.487+/-0.006	0.000	0.000
	BRAF	0.335+/-0.052	0.498+/-0.003	0.000	0.000
	RET	0.551+/-0.093	0.495+/-0.005	0.000	0.000

Table B.4: Results from using KL Divergence for feature selection, trained on MLP.

B.0.3 Extra Trees Classifier

LightGBM					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.664+/-0.033	0.627+/-0.027	0.445+/-0.05	0.431+/-0.034
	EGFR	0.751+/-0.034	0.590+/-0.039	0.420+/-0.109	0.229+/-0.082
	ROS1	0.744+/-0.058	0.495+/-0.005	0.000	0.000
	ALK	0.647+/-0.047	0.537+/-0.039	0.16+/-0.185	0.100+/-0.082
	BRAF	0.562+/-0.041	0.518+/-0.039	0.183+/-0.26	0.055+/-0.073
	RET	0.584+/-0.083	0.498+/-0.003	0.000	0.000
LUSC	KRAS	0.535+/-0.124	0.500+/-0.000	0.000	0.000
	EGFR	0.480+/-0.02	0.497+/-0.003	0.000	0.000
	ROS1	0.587+/-0.034	0.506+/-0.015	0.2+/-0.245	0.022+/-0.027
	ALK	0.706+/-0.058	0.495+/-0.003	0.000	0.000
	BRAF	0.583+/-0.024	0.500+/-0.000	0.000	0.000
	RET	0.519+/-0.085	0.499+/-0.001	0.000	0.000

Table B.5: Extra Trees Classifier results trained on LightGBM.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.729+/-0.014	0.649+/-0.023	0.482+/-0.019	0.456+/-0.055
	EGFR	0.783+/-0.012	0.690+/-0.028	0.509+/-0.055	0.448+/-0.057
	ROS1	0.634+/-0.022	0.495+/-0.003	0.000	0.000
	ALK	0.471+/-0.015	0.554+/-0.032	0.163+/-0.086	0.133+/-0.067
	BRAF	0.559+/-0.101	0.583+/-0.030	0.553+/-0.270	0.182+/-0.057
	RET	0.380+/-0.052	0.492+/-0.004	0.000	0.000
LUSC	KRAS	0.348+/-0.051	0.500+/-0.000	0.000	0.000
	EGFR	0.590+/-0.051	0.558+/-0.049	0.400+/-0.374	0.120+/-0.098
	ROS1	0.545+/-0.038	0.510+/-0.012	0.207+/-0.194	0.033+/-0.027
	ALK	0.498+/-0.140	0.500+/-0.000	0.000	0.000
	BRAF	0.492+/-0.032	0.499+/-0.002	0.000	0.000
	RET	0.589+/-0.090	0.530+/-0.065	0.067+/-0.133	0.067+/-0.133

Table B.6: Extra Trees Classifier results trained on MLP.

B.0.4 Random Forest Experiments on MLP

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.771+/-0.009	0.666+/-0.019	0.463+/-0.026	0.533+/-0.038
	EGFR	0.782+/-0.024	0.699+/-0.049	0.439+/-0.063	0.495+/-0.098
	ROS1	0.698+/-0.053	0.506+/-0.023	0.050+/-0.100	0.025+/-0.050
	ALK	0.600+/-0.059	0.573+/-0.003	0.253+/-0.065	0.167+/-0.000
	BRAF	0.536+/-0.032	0.569+/-0.049	0.427+/-0.360	0.164+/-0.089
	RET	0.762+/-0.026	0.495+/-0.002	0.000	0.000
LUSC	KRAS	0.679+/-0.058	0.497+/-0.001	0.000	0.000
	EGFR	0.7900+/-0.056	0.500+/-0.000	0.000	0.000
	ROS1	0.506+/-0.017	0.502+/-0.011	0.130+/-0.069	0.044+/-0.022
	ALK	0.669+/-0.096	0.506+/-0.038	0.033+/-0.067	0.040+/-0.080
	BRAF	0.548+/-0.049	0.494+/-0.004	0.000	0.000
	RET	0.543+/-0.100	0.486+/-0.007	0.000	0.000

Table B.7: Random Feature Selector results trained on MLP. All the features with importance higher than 0.001 were selected.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.812+/-0.011	0.701+/-0.019	0.516+/-0.048	0.585+/-0.019
	EGFR	0.795+/-0.029	0.733+/-0.021	0.469+/-0.015	0.552+/-0.049
	ROS1	0.693+/-0.063	0.495+/-0.003	0.000	0.000
	ALK	0.563+/-0.018	0.575+/-0.006	0.345+/-0.140	0.167+/-0.000
	BRAF	0.506+/-0.030	0.563+/-0.035	0.375+/-0.223	0.145+/-0.073
	RET	0.418+/-0.052	0.495+/-0.003	0.000	0.000
LUSC	KRAS	0.655+/-0.026	0.500+/-0.000	0.000	0.000
	EGFR	0.568+/-0.090	0.500+/-0.000	0.000	0.000
	ROS1	0.493+/-0.031	0.488+/-0.005	0.000	0.000
	ALK	0.600+/-0.095	0.499+/-0.002	0.000	0.000
	BRAF	0.526+/-0.015	0.500+/-0.000	0.000	0.000
	RET	0.620+/-0.080	0.488+/-0.004	0.000	0.000

Table B.8: Random Feature Selector results trained on MLP. The best 100 features were selected.

MLP					
Cancer Type	Mutation	Test AUC	Test Accuracy	Test Precision	Test Recall
LUAD	KRAS	0.838+/-0.012	0.710+/-0.015	0.538+/-0.034	0.585+/-0.041
	EGFR	0.868+/-0.004	0.774+/-0.010	0.352+/-0.021	0.762+/-0.000
	ROS1	0.755+/-0.030	0.761+/-0.046	0.116+/-0.009	0.875+/-0.112
	ALK	0.780+/-0.018	0.741+/-0.031	0.142+/-0.016	0.633+/-0.067
	BRAF	0.681+/-0.020	0.650+/-0.039	0.154+/-0.023	0.154+/-0.023
	RET	0.521+/-0.029	0.507+/-0.028	0.032+/-0.003	0.480+/-0.160
LUSC	KRAS	0.628+/-0.012	0.632+/-0.066	0.600+/-0.374	0.267+/-0.133
	EGFR	0.698+/-0.016	0.580+/-0.022	0.043+/-0.002	0.720+/-0.098
	ROS1	0.535+/-0.011	0.493+/-0.024	0.118+/-0.013	0.400+/-0.054
	ALK	0.638+/-0.041	0.615+/-0.006	0.077+/-0.006	0.400+/-0.000
	BRAF	0.619+/-0.051	0.497+/-0.003	0.000	0.000
	RET	0.622+/-0.021	0.638+/-0.078	0.048+/-0.015	0.467+/-0.163

Table B.9: ADASYN implementation along with the Random Feature Selector and ANOVA, trained on MLP.

Tuned Parameters LightGBM											Feature
		learning_rate	n_estimators	max_depth	num_leaves	min_data_in_leaf	feature_fraction	subsample	bagging_freq	bagging_fraction	
LUAD	KRAS	0.41136577925764123	299	27	971	28	0.18324023937297257	0.7902432761655531	0	0.9075983999825247	
	EGFR	0.005145853438087209	62	5	741	20	0.7732012409128814	1.0	20	0.9153768678509719	
	ROS1	0.99	130	29	111	1	0.9296769145670454	0.1	17	0.35023646796008867	
	ALK	0.99	177	1	2	46	0.1	1.0	0	0.6967057232537828	
	BRAF	0.0010180888324023967	275	16	35	5	0.12979471415551108	0.41498728184381306	17	0.7537960980521498	
	RET	0.3636701559754866	58	23	1000	100	1.0	0.5440962808441718	1	0.8776890622645767	
LUSC	KRAS	0.7071634607739256	85	26	355	7	0.37306909889575746	0.6721561916912533	18	0.8536495199979559	
	EGFR	0.2539746856232988	239	16	873	10	0.6382423697609152	0.9207359753561953	3	0.8311563895216271	
	ROS1	0.006710857837784945	75	1	566	84	0.1	0.12277904239606001	18	0.1	
	ALK	0.017605496796288214	50	12	372	22	0.8844867356168935	0.634941022232247	14	0.9599301876789355	
	BRAF	0.001	50	30	1000	75	1.0	0.1	3	0.7938724807234536	
	RET	0.001	300	1	2	28	0.1	0.5447746030357573	20	0.1	

Table B.10: Tuned parameters for LightGBM.