FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

An Explainable Approach for Lung Cancer Classification and Integrative Survival Analysis using Omics Data

Bernardo Manuel Faria Ramos

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Tânia Pereira, Hélder Oliveira, José Luis Costa

February 26, 2021

An Explainable Approach for Lung Cancer Classification and Integrative Survival Analysis using Omics Data

Bernardo Manuel Faria Ramos

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Rui Camacho
External Examiner: Nuno Moniz
Supervisor: Tânia Pereira

February 26, 2021

Abstract

Cancer is a major cause of death worldwide, the *Global Burden of Disease* — a global report on the causes of death — reported 9.56 million cancer deaths in 2017, making it the second-highest mortality rate and causing one in six deaths that year. Primary lung cancer accounts for 13% of new cancer cases annually, with the histologic subtypes lung adenocarcinoma and squamous cell carcinoma composing approximately 60% of all types of lung cancer cases.

Cancer prediction relates to the process of differentiating cancerous from normal tissue, whereas histologic subtype classification differentiates groups within the same type of cancer based on specific characteristics of the cancer cells. Lung cancer represents a group of histologically and molecularly heterogeneous diseases even within the same histological subtype. Moreover, accurate histological subtype diagnosis influences the specific subtype's target genes, and thus discovering risk genes for each subtype will help define a personalised treatment plan that can target those genes in therapy.

The *American Joint Committee on Cancer* staging system originated in the 1950s from the need for an accurate, consistent, universal cancer outcome prediction system, and since then, new prognostic factors have been identified and new methods for integrating prognostic factors have been developed. Even though the analysis of clinical parameters such as cancer staging and histopathological assessment can provide useful prognostic insights, they fail to capture the complex associations needed for accurate prognosis. Understanding what underlying genomic changes are affecting survivability will improve patients' risk stratification and has the potential to enhance the prognostic and treatment mechanism.

In this work, we tackle two different problems: 1) cancer prediction and subtype classification using gene expression data; 2) predicting the survival outcome of patients using clinical and multiple genomic data. For problem one, we use two different approaches: a deep learning model, which is the standard strategy in the state-of-art; and gradient boosted trees that we provide explainability on using Shapley additive explanations to retrieve valuable biological insight. For the second problem, we use unsupervised learning techniques to extract representations of each modality, in order to reduce dimensionality and avoid the data redundancy induced by combining heterogeneous data types. Furthermore, we propose a method of late fusion to combine the latent representations of each modality into a deep learning network with a Cox regression layer that predicts the survival of lung cancer patients.

On cancer prediction, the best performing model achieved an area under the receiver operating characteristic curve of 0.984, and 0.971 on subtype classification, leading to an improvement over the previous state of the art's results. Furthermore, due to our interpretable approach for cancer classification, two sets of gene signatures were extracted: a set that differentiates normal from cancerous tissue and another that distinguishes adenocarcinoma from squamous cell samples. These genes were analysed by performing hierarchical clustering to find commonly regulated genes, which resulted in the identification of relevant subtype-specific gene signatures that might be potential targets for personalised subtype therapy. For the survival outcome prediction problem,

using a variational autoencoder on the unsupervised learning stage allowed us to reduce the high-dimensional gene expression profiles without suffering from overfitting and vanishing gradients due to the limited sample size. To validate the late fusion model's effectiveness and analyse the features generated by the unsupervised learning methods, we compared their effectiveness in the risk stratification of cancer patients with interpretable features used by single-modality learners. The results showed a significant performance gain of the late fusion model compared with single modalities, with a concordance index of 0.701 for adenocarcinoma's survival prediction and 0.622 for squamous cell carcinoma. The inclusion of extracted features from multiple modalities led to the selection of prognostic factors fitter for survival prediction, which allows for better risk stratification of lung cancer patients and can lead to an improvement of the treatment and prognostic mechanism.

CCS Concepts: • **Applied computing** → Life and medical sciences → Genomics → Computational genomics; • **Computing methodologies** → Machine learning → Machine learning approaches → Neural networks;

Additional keywords and phrases: Machine learning, Deep Learning, Cancer Prediction, Subtype Classification, Survival Prediction

Resumo

O cancro é uma das principais causas de morte no mundo, e de acordo com o *Global Burden of Disease* — um relatório mundial sobre causas de morte — em 2017 foram reportadas 9,56 milhões de mortes derivadas de cancro, sendo assim a segunda maior causa de mortalidade e resultando em uma em cada seis mortes nesse ano. A neoplasia proveniente do pulmão é responsável por 13% de todos os tipos de cancro registados anualmente e reparte-se em vários subtipos histológicos, entre os quais o adenocarcinoma pulmonar e o carcinoma de células escamosas compondo aproximadamente 60% de todos os casos de cancro do pulmão.

A previsão ou classificação de cancro está relacionada com o processo de distinguir tecido canceroso de tecido normal; por sua vez, a classificação de subtipo histológico diferencia grupos dentro do mesmo tipo de cancro, com base no estudo histopatológico do tumor. O cancro do pulmão é uma doença histologicamente e molecularmente heterogénea, apresentando diferenças mesmo dentro do mesmo subtipo histológico. Para além disso, o subtipo histológico influencia os genes de risco que estão ativos e, portanto, a descoberta dos mesmos possibilita o desenvolvimento de novas terapias personalizadas para subtipos histológicos.

O sistema de estadiamento de cancro do *American Joint Committee on Cancer* foi iniciado em meados de 1950 com o intuito de criar um sistema universal e consistente para o prognóstico de pacientes com cancro, e desde então surgiram novas metodologias que integram vários fatores de prognóstico. Ainda que a análise de parâmetros clínicos, como o estágio clínico ou o relatório histopatológico ajudem no prognóstico de pacientes de cancro, não conseguem, porém, captar as complexas correlações necessárias para oferecer um bom prognóstico. A compreensão dos processos genómicos que estão a acontecer é fundamental para entender como estas mudanças afetam a sobrevivência dos pacientes, o que pode aumentar o nível de estratificação dos pacientes e tem o potencial para melhorar o mecanismo de prognóstico e de tratamento de doentes.

Neste trabalho abordamos duas temáticas: 1) a classificação de cancro e de subtipo histológico, usando dados de expressão genética; 2) a previsão da sobrevivência de pacientes com cancro do pulmão, usando dados clínicos e vários tipos de dados genómicos. Para o primeiro problema usamos duas estratégias: recorremos um modelo de *deep learning*, já que este é o padrão no atual estado da arte; e usamos métodos de *gradient boosted trees* de uma forma interpretável com o intuito de obter resultados relevantes do ponto de vista biológico. Para o segundo problema, utilizamos métodos de aprendizagem não supervisionada, de formar a extrair representações significativas de cada modalidade, com o intuito reduzir a alta dimensionalidade dos dados genómicos e evitar qualquer tipo de redundância causada pela heterogeneidade dos dados. Para além disso, propomos um método de *late fusion* de forma a combinar as representações latentes de cada modalidade que são fundidas numa rede neuronal com uma *Cox regression layer* para efetuar a previsão da sobrevivência dos pacientes com cancro.

O melhor modelo para previsão de cancro obteve uma performance *area under the receiver operating characteristic curve* de 0.984, e de 0.971 para a classificação do subtipo, o que representa uma melhoria em relação ao estado da arte. Além disso, como resultado da nossa abordagem

interpretável na classificação de cancro e subtipo, dois conjuntos de assinaturas genéticas foram extraídas: um grupo que distingue tecido normal de cancerígeno e outro que distingue cancro adenocarcinoma de escamoso. Estes genes foram analisados, usando *clustering* hierárquico, de forma a identificar genes co regulados, o que permitiu a identificação de assinaturas sob e sobre expressas em cada subtipo histológico que podem ser potenciais alvos de terapia personalizada. Para o problema da previsão da sobrevivência, o uso de um *variational autoencoder* possibilitou a redução da alta dimensionalidade dos dados de expressão genética sem haver perdas de eficiência devido à quantidade reduzida de dados disponíveis. Para validar a eficácia do modelo de *late fusion* e as representações geradas pelos modelos de aprendizagem não supervisionada, comparamos a eficácia destes, na estratificação de pacientes com cancro com os parâmetros clínicos, que são interpretáveis. Os resultados mostraram melhorias significativas no desempenho preditivo do modelo de *late fusion* com um *concordance index* de 0.701 na previsão de sobrevivência da coorte adenocarcinoma e de 0.622 para o escamoso. As *features* geradas pelos modelos de aprendizagem não supervisionada e usadas pelo modelo de *late fusion*, mostraram melhor separabilidade em relação aos melhores parâmetros prognósticos clínicos, o que contribui para uma melhor estratificação de pacientes com cancro do pulmão e por sua vez da previsão da sobrevivência.

Conceitos ACM CCS: • **Computação aplicada** → Ciências médicas e sociais → Genética → Computação genética; • **Metodologias de computação** → Aprendizagem automática → Estratégias de aprendizagem automática → Redes Neuronais;

Palavras chave: Aprendizagem automática, Redes Neuronais, Previsão de cancro, Classificação de subtipo, Previsão de sobrevivência

Acknowledgements

First of all, I wish to express my deepest gratitude to my supervisors Tânia Pereira and Hélder Oliveira for their constant feedback and promptness to help at any time, and whose perceptive reasoning helped me evolve my critical thinking, and José Luis Costa which guided me to understand the genomics domain.

I acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for the free publicly available TCGA database used in this work and INESC-TEC for providing the computational requirements needed for this work.

Lastly but not least, I am very grateful to my family, particularly to my parents who are my inspiration, and my friends who supported me throughout this challenging year.

Bernardo Ramos

‘ I am one of those who think like Nobel, that humanity will draw more good than evil from new discoveries.”

Marie Curie

Contents

List of Figures

List of Tables

Abbreviations

ACC	Adenoid Cystic Carcinoma
ADAM	Adaptive Moment Estimation
ADASYN	Adaptive Synthetic Sampling Technique
AIC	Akaike Information Criterion
AJCC	The American Joint Committee on Cancer
ANN	Artificial Neural Network
AUC	Area under the ROC curve
BCR	Breakpoint Cluster Region Gene
BLCA	Urothelial Bladder Carcinoma
BN	Bayesian Network
BNorm	Batch Normalisation
BO	Bayesian Optimisation
BRCA	Breast Cancer
BT	Boosted Tree
CESC	Cervical Squamous Cell Carcinoma
CGF	Clustered Gene Filtering
CGN	Graph Convolutional Network
CGS	Clustered Gene Selection
CNN	Convolutional Neural Network
COAD	Colon Adenocarcinoma
Cox-PH	Cox Proportional-Hazards
C-Index	Concordance Index
DA	Data Augmentation
DBN	Deep Belief Network
DFF	Deep Feedforward
DL	Deep Learning
DLN	Deep Learning Network
DNA	Deoxyribonucleic acid
DNN	Deep Neural Network
DT	Decision Tree
EF	Early Fusion
EFB	Exclusive Feature Bundling
EI	Expected Improvement
EMBC	Engineering in Medicine and Biology Society
ESMO	European Society for Medical Oncology
FE	Feature Extraction
FS	Feature Selection
GAN	Generative Adversarial Networks
GBDT	Gradient Boosted Decision Tree
GBM	Gradient Boosted Machine
GCN	Graph Convolutional Network

GDC	Genomic Data Commons
GO	Gene Ontology
GOSS	Gradient-based One-Side Sampling
GP	Gaussian Process
GP-UCB	GP Upper Confidence Bounds
GRCh38	Human Reference Genome
HCC	Hepatocellular Carcinoma
HGVS	Human Genome Variation Society
HNSC	Head-Neck Squamous Cell Carcinoma
HOG	Histogram of Oriented Gradient
HR	Hazard Ratio
HUGO	Human Genome Organisation
ICD-10	International Classification of Diseases 10
IGSR	The International Genome Sample Resource
ISR	Index-Sparsity Reduction
KL	Kullback-Leibler
KM	Kaplan-Meier
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
kPCA	kernel PCA
k-NN	k-Nearest Neighbours
LBP	Local Binary Patterns
LF	Late Fusion
LCC	Large-Cell carcinoma
LGG	Low Grade Glioma
LGBM/LightGBM	Light Gradient Boosting Machine
LR	Likelihood Ratio
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinomas
miRNA	microRNA
ML	Machine Learning
mRNA	messenger RNA
mRMR	Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy
MSE	Mean Squared Error
NAdam	Nesterov-accelerated Adaptive moment estimation
NB	Naïve Bayes
NCCN	National Comprehensive Cancer Network
NCI	National Cancer Institute
NGS	Next-Generation Sequencing
NN	Neural Network
NPV	Negative Predictive Value
NSCLC	Non-Small Cell Lung Cancer
NT	Normal Tissue
OV	Ovarian Cancer
PAAD	Pancreatic Adenocarcinoma
PCA	Principal Component Analysis

pGBRT	Parallel Boosted Regression Trees
PI	Probability of Improvement
PRAD	Prostate Adenocarcinoma
RAS	Rat Sarcoma Virus Protein
RBM	Restricted Boltzman Machines
ReLU	Rectified Layer Unit
RF	Random Forest
RNA-seq	Ribonucleic Acid Sequencing
ROC	Receiver operating characteristic
ROI	Regions Of Interest
SA	Stacked Autoencoders
SCLC	Small Cell Lung Cancer
SDAE	Stacked Denoising Autoencoder
SEER	The Surveillance, Epidemiology, and End Results
SHAP	SHapley Additive exPlanations
SKCM	Skin Cutaneous Melanoma
SMOTE	Synthetic Minority Over-Sampling Technique
SNF	Similar Network Fusion
SNP	Single Nucleotide Polymorphism
SPC	Surfactant Protein C
SSA	Stacked Sparse Autoencoder
STAD	Stomach Adenocarcinoma
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas Program
THCA	Thyroid Cancer
TMN	Tumour, Nodes and Metastasis
TP	Primary Tumour Tissue
t-SNE	t-distributed Stochastic Neighbour Embedding
UCS	Uterine Carcinosarcoma
VAE	Variational Autoencoder
VIF	Variance Inflation Factor
WES/WXS	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
XGboost	Extreme Gradient Boosting Machine
WHO	World Health Organization
WSI	Whole-Slide Images
2D	Two-Dimensional
3D	Three-Dimensional

Chapter 1

Introduction

Cancer is a leading cause of death; in 2017, 9.56 million cancer deaths were reported worldwide, making it the second-highest mortality rate and causing one in six deaths that year [134]. Demographically, cancer mortality is more significant in developed countries, mainly due to the lower incidence of other diseases. Furthermore, the economical impact of cancer is significant and is increasing, in 2010, it was estimated at approximately US$ 1.16 trillion [134], so there is an increasing motivation to research it. Figure 1.1 presents the death toll of cancer distributed by type, with lung cancer totalling 1,88 million deaths.

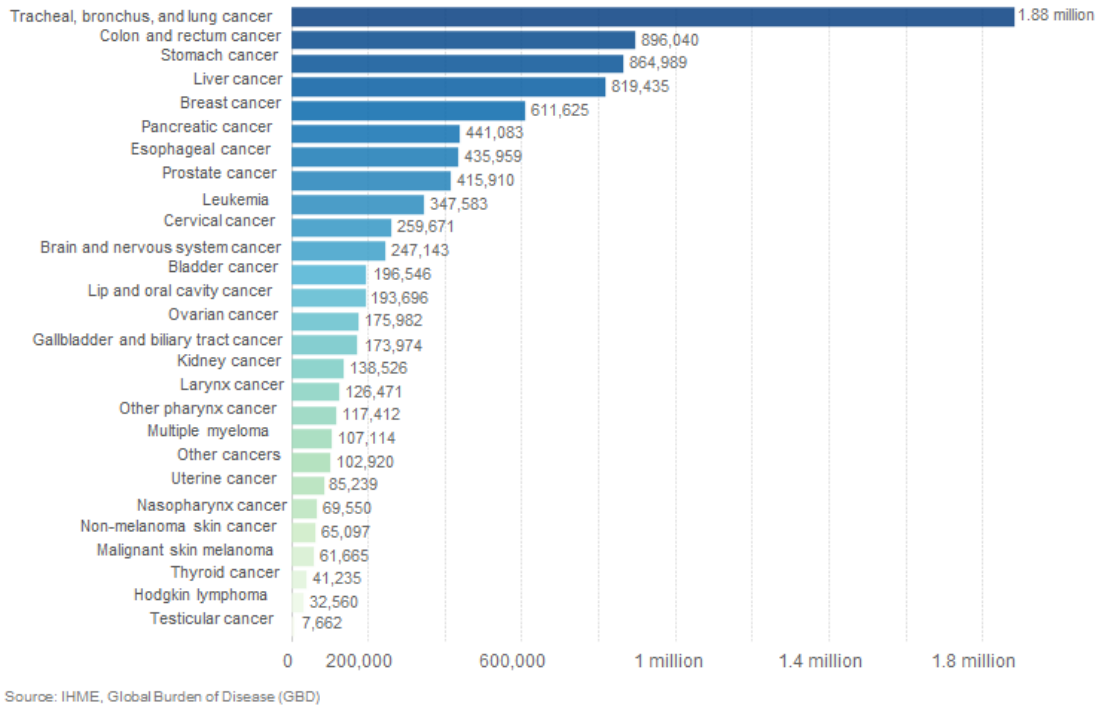


Figure 1.1: Number of deaths across all ages and both sexes, broken down by cancer type in 2017. Adapted from [80].

Around one-third of deaths from cancer are due to the five leading behavioural and dietary risks: high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco usage, and alcohol abuse. Tobacco use is the most critical risk factor for cancer and is responsible for approximately 22% of cancer deaths [100].

Lung cancer constitutes 13% of new cancer cases annually, Figure 1.2 shows an increase in the mortality rates for lung cancer in the last 30 years, that is directly correlated with an increase in incidence rate; however, the survival rate remains at 20% [134]. Studies show that between 30 to 50% of cancers can currently be prevented by avoiding risk factors and implementing existing evidence-based prevention strategies [100], yet early diagnosis is essential and a critical factor for chances of survival.

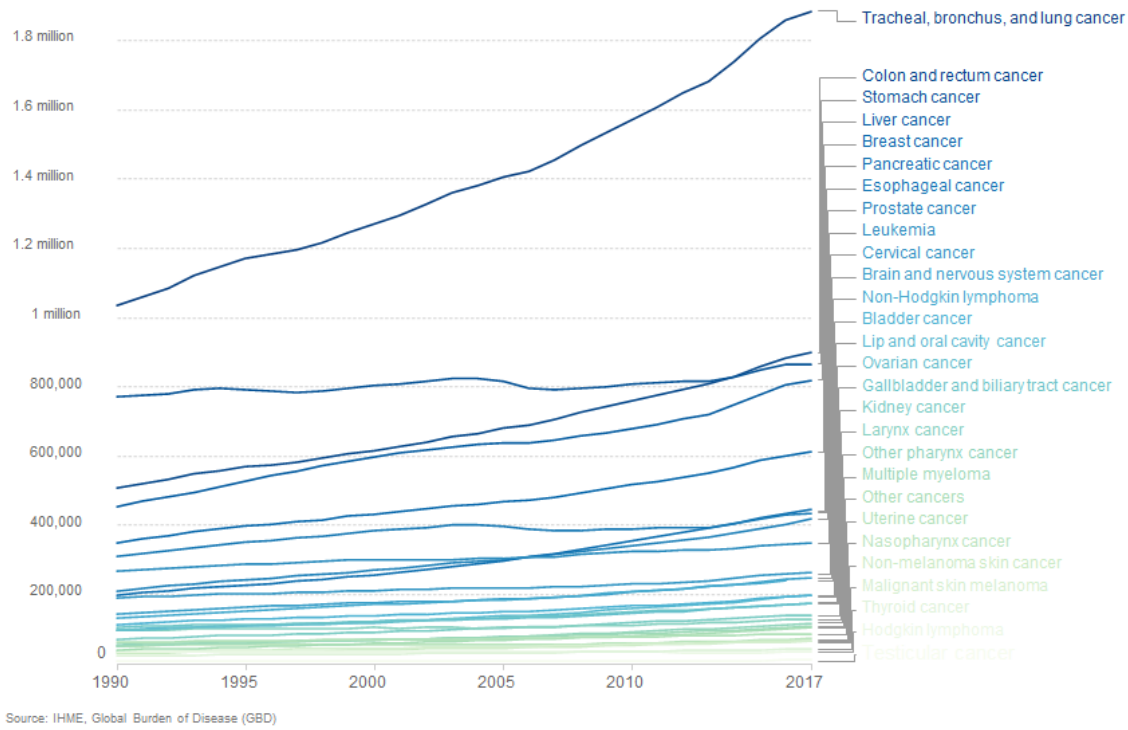


Figure 1.2: Progression of deaths broken down by type from 1990-2017. Adapted from [80].

1.1 Motivation

The use of computational genomics can impact cancer research in many ways, but maybe the most significant is understanding the underlying causes of cancer.

Cancer is a genetic disease caused by specific changes to genes (mutations) that control the way cells function, especially how they grow and divide [68]. It represents a group of histologically and molecularly heterogeneous diseases even within the same histological subtype [116]. Moreover, accurate histological subtype diagnosis influences the specific subtype's target genes, which will help define the treatment plan to target those genes in therapy. Interpretable subtype classification approaches allow us to identify gene expression signatures that better differentiate lung adenocarcinoma (LUAD) from lung squamous cell carcinoma (LUSC) subtypes, potentially discovering new targetable genes for personalised therapy.

Medical prognosis of cancer is done mainly through the analysis of clinical parameters to date. Although cancer staging, histopathological assessment, and clinical variables can provide useful prognostic insights [135], they fail to capture the complex associations needed for accurate prognosis. Combining multimodal data can lead us towards a deeper understanding of what underlying genomic changes affect survivability, which will improve the risk stratification of patients and has the potential to enhance the prognostic and treatment mechanism.

1.2 Objective

In this work, we aim to implement Machine Learning (ML) solutions for cancer prediction and subtype classification using gene expression data. Furthermore, we aim to develop Deep Learning (DL) solutions to perform a risk analysis assessment using multi-omics and clinical data to determine the outcome of patients diagnosed with cancer.

For the cancer prediction and subtype classification problems, we aim to develop two different ML models types: tree-based learning models that are interpretable and therefore provide biological insight; we also develop a DL model to benchmark the results, which is the standard strategy in the state-of-art. Our interpretable approach allows us to extract two sets of genes with biological relevance: a set that differentiates normal tissue from cancerous tissue, and a set of genes that distinguishes LUAD from LUSC samples.

For the outcome prediction problem, we use different types of genomic data and clinical variables to predict the survival of LUAD and LUSC patients. Our contribution focuses on the inclusion of heterogeneous data types needed for survival prediction that have shown contradictory results in predictive performance on state of the art. We experiment with early and late fusion techniques to effectively combine the different modalities and use a DL model to assess LUAD and LUSC cohorts' survivability.

1.3 Contributions

- **Distinguish cancerous from non-cancerous tissue** - We implemented a deep feedforward network (DFF) to benchmark the results, and a gradient boosted tree model (GBDT) to differentiate between cancerous and non-cancerous gene expression samples. Our best performing classifier, the DFF, obtained an Area Under the ROC Curve (AUC) of 0.984.

- **Distinguish LUAD from LUSC tissue** - A DFF and GBDT were also implemented to differentiate between LUAD and LUSC histologic subtypes' gene expression samples. The best performing classifier, the GBDT model obtained an AUC of 0.971 conferring better results than state of the art.

- **Identify relevant cancer and subtype-specific gene signatures** - We provided interpretability on the tree-based learner in order to identify relevant gene signatures that differentiate between cancerous and normal tissue, and LUAD and LUSC genomic profiles. This resulted in a paper in the process of publication at the IEEE EMBC 2021 conference[1].

- **Predict LUAD and LUSC patients' survival outcome** - We implemented a Late Fusion (LF) model to combine gene expression, mutation and clinical modalities for the survival prediction of the LUAD and LUSC cohorts. Unsupervised learning is used to extract latent representations of each modality, therefore reducing the high-dimensionality of the genomic data. Our LF approach achieved a concordance index (C-Index) of 0.701 for LUAD and 0.622 for the LUSC cohort, presenting better results than single-modality approaches while solving the problem of combining heterogeneous genomic data.

1.4 Document Structure

Our work is divided into six chapters, as follows. On chapter 2, we contextualise the problem, giving a brief definition on cancer and how it develops, and a more detailed explanation on the genomic aspect of cancer, more specifically lung cancer. On chapter 3, we perform a literature review, where we approach the machine learning models we will be implementing as well as an assessment of the literature using the same datasets. On chapter 4, we perform an exploratory analysis on the datasets we use for data understanding. On chapter chapter 5 we detail our methodology for the for cancer prediction and subtype classification problems, and on chapter 6 for the survival outcome prediction problem.

[1] Conference of the IEEE Engineering in Medicine and Biology Society (https://embc.embs.org/2021)

Chapter 2

Problem Contextualization

In this chapter, we contextualize some of the main biology concepts required for an understanding of the cancer genomics research. We start by defining cancer and give an overview of how cancer develops, and then focus on lung cancer histologic groups and define its subtypes. Finally, we give an overview of the underlying genetics in cancer and briefly explain the data types we use.

2.1 Cancer

2.1.1 Understanding the Cancer Mechanism

On Hanahan and Weinberg [57], the authors build a framework to understand the biological capabilities acquired during the multistep development of human tumours. The model in Figure 2.1, can be seen as a game, where cells need to acquire a set of capabilities without which they are not fit as a cancer cell. In the first work, the authors describe six capabilities, which we explain succinctly:

Resisting cell death: When cells stop behaving properly, they go through apoptosis (launch their own "suicide mission"), so it is necessary to avoid apoptosis.

Sustaining proliferative signalling: Cells have to discover ways to proliferate, by generating growth signals.

Inducing angiogenesis: They need the energy to multiply, so they have to induce angiogenesis, which consists of sending signals to the body to transfer in more blood.

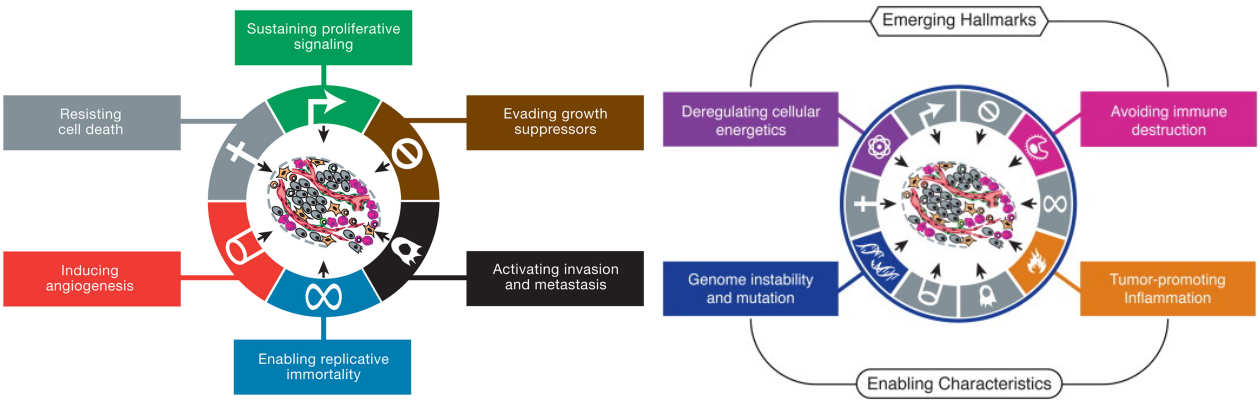
Evading growth suppressors: Antigrowth signals can block proliferation. It needs to stop listening from signals from the rest of the body.

Activating invasion and metastasis: It has to figure out ways to invade tissues and to metastasize so that it can escape the tumour mass and colonize new tissue.

Enabling replicative mortality: It needs to replicate infinitely.

Deregulating cellular energetics: The cell needs the energy to survive, so it has to destabilize the energy-producing machinery.

Avoid immune destruction: The tumour is actively reprogramming immune cells in order not to recognize and attack it.



(a) Six acquired capabilities of cancer (b) Emerging hallmarks and traits

Figure 2.1: Images from Hallmarks of Cancer: Cell [57] and The Next Generation Cell [58].

Tumour-promoting inflammation: A characteristic associated with the process of angiogenesis in which the influx of blood causes inflammation.

Genome instability and mutation: Gain of genetic instability, which causes cells to increase the accumulation of mutations

2.1.2 Histology of Lung Cancer

Lung cancer is categorized into two main histological groups: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLCs are generally subcategorized into adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large cell carcinoma (LCC). Accumulating evidence suggests that lung cancer represents a group of histologically and molecularly heterogeneous diseases even within the same histological subtype [67]. On Figure 2.2, the histologic groups and the most common subtypes of lung cancer are presented along with their frequencies. The molecular heterogeneity of lung cancer makes it a difficult task to identify a set of well-defined

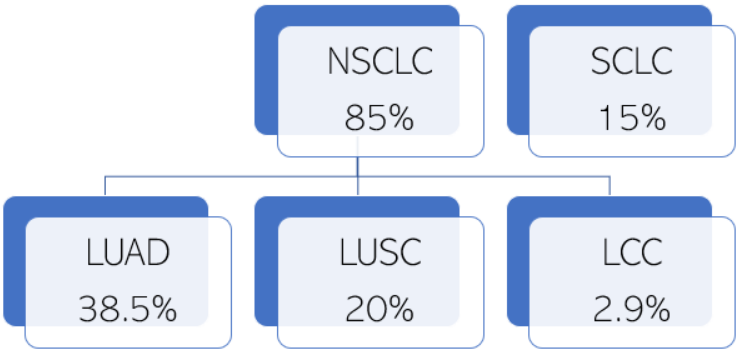


Figure 2.2: Frequency of lung cancer histologic types and subtypes, data from [31], [64].

risk genes for each subtype. The identification of the risk genes will help to define the treatment plan that can target them in therapy, so it is necessary to provide accurate diagnosis according to the histologic subtype.

2.2 Cancer Genomics

2.2.1 Mutations

Cancer is a combination of germline mutations, somatic mutations, epigenomic changes, and gene-regulatory alterations that give rise to complex phenotypes [69]. Mutations can be classified according to their origin as germline or somatic mutations. A germline mutation occurs in the germline, a specialized tissue that is set aside in the course of development to form sex cells and can be passed to offspring. Somatic mutations, however, can occur in any of the cells of the body except the germ cells (sperm and egg), which means that only tissues derived from the mutated cell are affected, and are generally triggered by environmental changes [37].

Somatic or acquired mutations are the most common cause of cancer, and are the ones represented in our data. Furthermore, we can classify cancer-causing mutations into two types: driver mutations confer an advantage to the growth of the tumour; passenger mutations do not directly contribute to the fitness of the tumour. Driver mutations are the ones we are interested in discovering, and passenger mutations can be seen as outliers because even though they are abnormalities, they do not have a direct impact on tumour growth. These driver genes responsible for inducing tumorigenesis are represented in four distinct classes:

- **Proto-oncogenes** are genes that normally promote normal cell growth; however, when mutated, they become oncogenes and stimulate overactive cell growth.
- **Tumour suppressors** usually function to slow cell division, and when mutated, they exhibit loss of function and allow for uncontrolled cell growth.
- **Mutator genes** are genes which normally regulate genomic stability.
- **Epi-mutator genes** are genes whose mutation or dysregulation leads to drastic gene-regulatory changes.

It is also important to mention the concept of a fusion gene, which is a gene resulting from the junction of two unrelated genes. Fusion genes may lead to the development of some types of cancer; for example, the BCR-ABL fusion gene and protein are found in some types of leukaemia [69].

2.2.2 Gene-regulatory pathways

A biological pathway is a series of actions among molecules in a cell that lead to a particular product or a change in the cell, gene-regulation pathways on its hand turn genes on and off [69]. These pathways are essential to cancer research as they provide a way to understand the interaction between chain mutations of genes. In Hammerman et al. [39], a Rat Sarcoma Virus (RAS) pathway

was identified which showed a potential therapeutic target in most tumours, offering new avenues of investigation for the treatment of squamous cell lung cancers.

2.2.3 Genomic Sequencing and Data Types

Gene expression or messenger RNA (mRNA) data is a result of RNA-sequencing (RNA-seq), which is a technique that can examine the quantity and sequences of ribonucleic acid (RNA) in a sample using next-generation sequencing (NGS). It analyzes the transcriptome of gene expression patterns encoded within our RNA [132]. RNA-seq can tell us which genes are turned on in a cell, what their level of expression is, and at what times they are activated or shut off [107], which enables towards a deeper understanding of molecular changes that might lead to disease.

Whole-genome sequencing (WGS), is a technique to sequence the entirety of the genome, whole-exome sequencing (WES) instead focuses on just the protein-coding sequences, covering just about 2% of the genome of an individual [96]. Both techniques collect information which makes it possible to analyze the somatic mutations present in an individual, and although WGS is most powerful, WES is more common due to its lower cost.

Chapter 3

Literature Review

In this chapter, we perform a broad literature review on machine learning applied to cancer research. The research on this topic is pervasive and has reached many different fields, including medical imaging, computer-assisted decision making, multimodal and transfer learning paradigms. In this review, we centre only on literature specific to applied computing for cancer genomics.

3.1 Machine Learning for Cancer Genomics

Although it is unattainable to do a full-coverage, we tried to diversify our research to capture the paradigm of cancer genomics research. We focused on papers from the last five years and gave more attention to papers following the same line as our work. On diagram 3.1 we try to portray the possible research lines specific to this problem. Highlighted are the topics we will approach in our research.

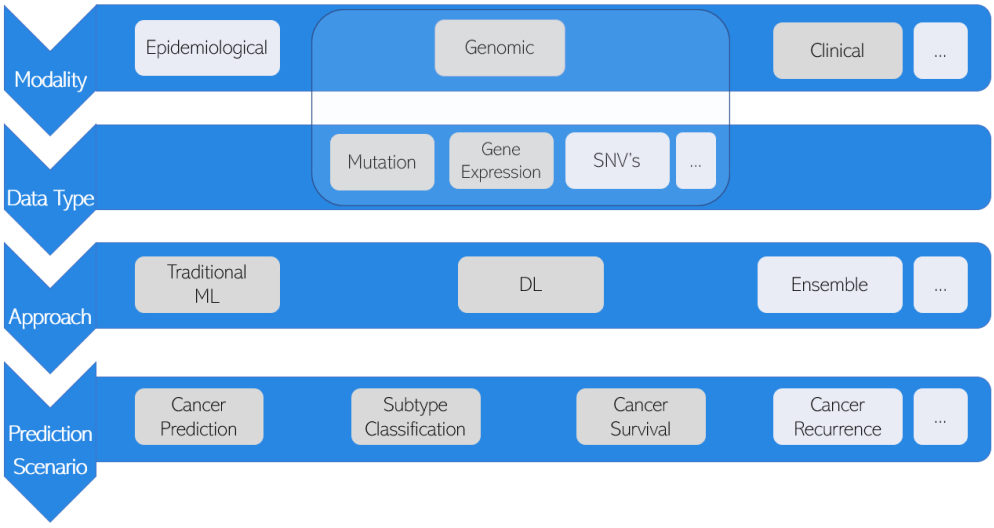


Figure 3.1: The paradigm of cancer genomics research. Topics highlighted in grey are the ones approached in this work.

To guide our research, we divide related research into three main topics:

- **Modality/Data** - It can be genomic only, but there is a noticeable trend in the integration of multiple modalities (clinical and genomic, or multi-omics). Genomic data represents a broad spectrum of data, which can be present in many forms. SNP arrays, for example, are used for detecting genetic variations within a population. Gene expression relates to the process the cell employs to produce the molecule it needs by reading the genetic code written in the DNA [36]. Mutation data is acquired by mapping the somatic mutations of an individual.

- **Approach** - We chose to divide into three separate categories: traditional machine learning models; deep learning models; and ensemble strategies which combine the multiple learners specified before.

- **Prediction Scenario** - Here, we look at three types of papers whose goal is to predict:

    - **Cancer Susceptibility** - This includes work that tries to detect cancer (cancer prediction), as well as those who try to predict cancer histologic subtypes (subtype classification).
    - **Cancer Recurrence** - Refers to predicting the re-incidence of cancer after treatment.
    - **Cancer Survival** - Applies to work whose interest is predicting the outcome for cancer patients.

On Tables 3.1, 3.2, 3.3 we summarize in tabular form the main concepts of the papers covered in the state of the art discussion for an easier understanding of each individual paper. Each table is divided as follows:

- **Focus** - A short summary, highlighting the main objective of the work.

- **Approach** - Details the specific strategies and models used to accomplish the objective.

- **Cancer Type** - Refers to the cancer type and histologic subtype (if applicable) that the research is directed.

- **Data** - Indicates the data type(s) used, and the origin of the specific datasets.

- **Metrics and Evaluation** - The type of validation performed and the metrics used for evaluation of performance.

- **Issues and Limitations** - Brief comments on possible limitations of the work.

- **Conclusions** - Explains the results obtained as well as the final remarks.

3.1.1 Dimensionality Reduction

Dimensionality reduction is widespread in pre-processing of high dimensional data analysis, visualization and modelling. One of the simplest ways to reduce dimensionality is by **Feature Selection** (FS), which refers to the process of selecting a smaller set of features from a wider set [76], normally as a pre-processing step of the modelling phase of the data mining process. **Feature Extraction** (FE) is a process that extracts a set of new features from the original features through some functional mapping [98], it is a more general method in which one tries to develop a transformation of the input space generally onto the low-dimensional subspace that preserves most of the relevant information [22].

The advantage of FS is that no information about the importance of single features is lost, as variables are not transformed. On the other hand, if a small set of features is required and the original features are very diverse, information may be lost as some of the features must be discarded. With FE techniques, the attribute space's size can often be decreased strikingly without losing much information about the original attribute space. An important disadvantage of FE is the fact that the linear combinations of the original features are usually not interpretable and the information about how much an original attribute contributes is often lost [74].

3.1.1.1 Feature Selection

Many FS methods applied to genomic data are based on a gene scoring function that assigns a score for each gene based on individual contribution, and such gene scoring functions can be the classification accuracy of individual genes. These methods generally return a number of top-ranked genes, and their quality is measured by the classification accuracy of the classifier built on them [14]. In Cai et al. [14], the authors propose clustered gene selection (CGS), a method based on the above premise; however, they propose doing gene clustering before gene selection, by applying an algorithm like K-Means. However, most genes might not contain informative point mutations and often remain normal (i.e. zero values in the data) [125], which results in very sparse data. In Yuan et al. [139], the authors apply index sparsity reduction to solve this problem, by converting the zero values to indexes of its non-zero elements. Furthermore, they propose a method for clustered gene filtering (CGF) that locates the discriminatory gene subset based on mutation occurrence frequency, by using the mean and standard deviation of the distances from each sample to the whole dataset [139], instead of to the class centre as in the CGS approach.

In Xiao et al. [136], the authors use differential expression analysis for sequence count data (DESeq)[99]. This technique estimates variance-mean dependence in count data from high-throughput sequencing assays and tests for differential expression based on a model using the negative binomial distribution [99], which managed to bring the selected genes to 3% of the total gene pool, and concluded that the selected feature set lead to better results. In Wang et al. [49]; the criteria of Max-Dependency, Max-Relevance, and Min-Redundancy (mRMR) was used; which tries to maximize the relevance between features and classification variables while minimizing relevance

between features [123]. The empirical analysis showed that mRMR criteria outperformed the simpler Max-Relevance criteria in terms of selected genes. Finally, in Chen et al. the authors used a more straightforward approach, a chi-square test to identify the top 10 rank survival correlated gene signatures for each data set [18], although a pre-process step of converting the continuous gene expression data to quintiles was necessary to reduce variability.

3.1.1.2 Feature Extraction

There are two broad categories for FE algorithms. Linear FE assumes that the data lies on a lower-dimensional linear subspace, and it projects them on this subspace using matrix factorization. In non-linear FE, a low-dimensional surface can be mapped on a high-dimensional space so that a non-linear relationship among the features can be found [65].

In Adetiba et al. [1], the authors applied a pre-processing step to convert the DNA sequences to binary indicator sequences mapping (Voss mapping) and extracted relevant features using Histogram of Oriented Gradient descriptors (HOG) [27] and Local Binary Patterns (LBP) [126], both are methods of extracting features from images. A similar strategy of embedding the gene expression values into a 2D image that considers the genes' overall features and computed features was used in De Guia et al. [29]. In Danaee et al. [28], a stacked denoising autoencoder (SDAE) was used to extract functional features from high-dimensional gene expression profiles (RNA-seq). In common among these approaches is the loss of interpretability of results compared with the feature selection strategies.

3.1.1.3 Summary

The curse of dimensionality is an inherent problem to the genetic expression research, as the datasets present a pattern of a large number of variables and a low samples count. This section discussed different FS and FE techniques, which are strategies generally used for dimensionality reduction to decrease the complexity of the modelling stage's input features and avoid overfitting.

A fundamental difference between these strategies is that FE transforms the features generally to a lower dimension space; therefore, we lose interpretability on the original features' contribution. On the other hand, FS reduces the feature dimension by selecting a subset of the original features, so part of the information must be discarded.

Method	Advantages	Disadvantages
FS	Preserving data characteristics for interpretability Shorter training times	Discriminative power Reducing overfitting
FE	Higher discriminating power Control overfitting when it is unsupervised	Loss of data interpretability Transformation maybe expensive

Table 3.4: A comparison between feature selection and feature extraction methods applied on gene expression data. Adapted from [65].

Interpretability is a significant concern from cancer biology's perspective since identifying the cancer driver genes in an individual provides essential information for treatment and prognosis. There has been a recent trend in trying to explain these non-interpretable models; however, lack of transparency and accountability of predictive models can have severe consequences when used for high-stakes decision [113]. Therefore, and mainly on this domain interpretability should be prioritized, so models that are already interpretable should be preferred.

3.1.2 Supervised and Unsupervised Learning

Machine learning methods can usefully be segregated into two primary categories: supervised or unsupervised learning methods. Supervised methods are trained on labelled examples and then used to make predictions about unlabelled examples, whereas unsupervised methods find structure in a data set without using labels [88]. The intermediate between supervised and unsupervised learning is semi-supervised learning. The semi-supervised setting is a mixture of these two approaches: the algorithm receives a collection of data points, but only a subset of these data points has associated labels [56].

3.1.2.1 Supervised and Semi-Supervised Learning

In Adetiba et al. [1], an artificial neural network (ANN) was used to classify various types of genomic mutations. Although ANN's are generally perceived as "weak" classifiers, the approach of using a simple ANN coupled with HOG yielded good results, conferring an accuracy 95.90%. The sample size was relatively small (6,406 mutations of EGFR, KRAS and TP53) which might explain why the ANN performs so well.

In Yuan et al. [139], a supervised learning approach using a deep neural network (DNN) classifier based on somatic point mutations, was used to classify 12 different types of cancer. The total dataset contained 484,463 mutations, CGF was used to filter relevantant mutations and a technique of index-sparsity reduction (ISR) was applied to shrink the data length. Furthermore, a validation was made to assess the results of the DNN model against other baseline models, as well as a validation of the ISR technique using the baseline models. The DNN classifier generated the best Accuracy 60.1% coupled with the CGF+ISR dataset, but all the baseline models also performed better after CGF+ISR pre-processing.

In Sun et al. [124], Mutect2 a method that applies a Bayesian classifier [23], was utilized to detect somatic mutations in 8,604 WES samples from 12 different types of cancer. The filtered data was fed into a four hidden-layer DNN. Exponential decay method was employed to optimize the learning rate [97], and L2 regularization to minimize overfitting [91]. It is proven in practice that pre-training the parameters in a deep architecture leads to a better generalization on a specific task of interest [129]. Greedy layer-wise pre-training is an unsupervised approach that helps the model initialize the parameters near a good local minimum and convert the problem to a better form of optimization [9]. Two types of models were created, a specific cancer type model and one to classify the mixture of 12 types. The average Accuracy of the specific model was 94.70%,

and the mixture model 70.08%. Some limitations of the work that might explain the results in the mixture model are the unbalanced distribution in-between cancer types, as well as the inclusion of only genomic data. Still, the specific models presented excellent predictive performance.

3.1.2.2 Unsupervised Learning

In Dannae et al., both supervised and unsupervised learning methods were used. In pre-training, an SDAE was used to extract features from gene expression data [28]. Support vector machines (SVM) and ANN's models were used to assess the SDAE generated features in a supervised manner. The dataset was very imbalanced, but synthetic minority over-sampling technique (SMOTE) was used to even out the samples. Validation was performed to assess the feature sets and compare them with feature sets from other unsupervised learning models like principal component analysis (PCA), on the task of gene identification. The DSAE model generated the better feature set when compared with other baseline models, performing best in Accuracy (98.26%), Sensitivity (98.73%) and F-measure (0.983) against the second-place kernel PCA (kPCA), which had excellent Specificity and Precision. Although the results are excellent, two pitfalls come from this strategy: first, the loss of interpretability of results which is generally not desirable in this domain; and also the requirement for a large dataset.

In Chaudhary et al. [15], a similar strategy was used for predicting survival in liver cancer using multi-omics data. First, an autoencoder with three hidden layers was used to transform features from 360 multi-omics samples. Then for each extracted feature, a Cox proportional hazards (Cox-PH) model [89] was built to select the features. The final reduced feature set was used to cluster the samples by K-Means. Finally, two experiments were conducted, one with multi-omics data only, and one with multi-omics and clinical data. Concordance index (C-index) [121] was used to validate these experiments, and the authors concluded that adding the clinical data did not improve the performance of the DL multi-omics based model. The authors speculate that this is due to the unique advantage of the DL neural network, which can capture the redundant contributions of clinical features through their correlated genomic features [15].

However, in Wang et al. [49] comparative research, their results contradict those of [15]. Their approach uses mRMR for feature selection and builds a similarity matrix of the multi-omics and clinical data using similar network fusion (SNF) algorithm, followed by semi-supervised learning with a graph convolutional network (GCN) classifier. The results were evaluated in two separate sets, using only multi-omics and combining it with clinical data. The validation of the GCN model was performed using baseline models: Naïve Bayes (NB), k-nearest neighbours (k-NN), decision tree (DT), LR, SVM and DNN. The results show a definite improvement in the addition of clinical data: 0.7805 against 0.9280 AUC improvement for BRCA and 0.7840 to 0.8050 in LUSC for the GCN model. This improvement was also verifiable in all the baseline models, including the DNN, the model used in [15]. This indicates that although the GCN model might be more adapted to handle this data, the SNF method to fuse the genomic and clinical data might be a crucial step that was overseen in previous research.

3.1.2.3 Summary

On this section, we discuss various approaches for supervised and unsupervised learning and even semi-supervised learning. Overall supervised learning is mostly used on the cancer susceptibility and cancer recurrence scenarios; and unsupervised and semi-supervised learning on the survival prediction scenario.

Regarding the cancer susceptibility scenario, in general, previous works show us that DL models are prime candidates for cancer classification and subtype prediction tasks. One problem raised by the use of deep learning models is the lack of interpretability. Moreover, although efforts have been made to provide interpretability on the DL learners [28] [2], we can pinpoint two downfalls of previous approaches: first, the DNN's performed better when performing feature selection a priori using variance selection techniques, which might leave out essential genes that are not selected by variance or other techniques; secondly, the algorithms used to extract weights from the network are based on leave one out technique, that might fail to capture correlations between some features post feature selection. Providing interpretability on cancer prediction and subtype classification tasks allows us to provide biologic insight that can significantly impact cancer genomics research.

Unsupervised learning finds patterns in data without a specific goal which makes it prone for cancer survival prediction, as a significant stage of it is the risk stratification of patients, making clustering algorithms suitable for this task. In the survival analysis scenario, a critical gap found in the literature is related to handling the heterogeneous data types needed for survival prediction. In Chaudhary et al. [15] and Wang et al. [49], contradictory results appear regarding the use of genomic and clinical data to predict cancer survival. To our understanding, the cause of worse results on [15] comes from the methodology applied to join the heterogeneous data, which can cause redundancy on the data. Effectively combining multimodal data can lead us towards a deeper understanding of what underlying genomic changes affect survivability.

Table 3.1: Comparison between models for predicting Cancer Susceptibility I

Reference	Focus	Approach	Cancer Type	Data	Metrics and validation	Issues and Limitations	Conclusion
Adetiba et al. (2015)[1]	Predict the presence or absence of specific genomic mutations using ensembles. Feature extraction was applied to extract representative genomic features.	Experimental comparison of ANN and SVM ensembles. A feature extraction schema with HOG and LBP was applied to extract representative genomic features.	———	EGFR, KRAS and TP53 normal and mutated genes. Data from IGDB.NSCLC[40]	20-fold cross-validation MSE, Accuracy	Only analyzes 3 biomarkers. Lacks in comparative studies. Uses only ANN's.	The ANN ensemble and HOG best fit the training dataset with an accuracy of 95.90% and mean square error of 0.0159.
Liang et al. (2015)[87]	Identification of cancer subtypes and survival risk, by clustering cancer patients, with multimodal data.	Uses DBN to cluster cancer patients. Infers the network parameters, using unsupervised learning, with Contrastive Divergence algorithm. First handles modalities separately and then fuses the resulting features.	OV, BRCA	The ovarian dataset contained gene expression, DNA methylation and miRNA expression data, and clinical data, and was taken from TCGA. The breast cancer dataset contained gene expression data from NKI	MSE	Validation is made to assess the ideal number of hidden variables in the model. But it is missing validation when comparing it to other clustering algorithms. Also lacks in cancer types.	Unlike the K-means, the proposed method remains stable under perturbation of initial states. Unlike traditional clustering techniques, it can handle multimodal data.
Yuan et al. (2016)[139]	Identify cancer types, and address the common problem of data sparsity.	Use of a clustered gene filtering technique. Perform Index sparsity reduction by removing zeroes, resulting in a new dataset. Feed the dataset into a DNN classifier.	ACC,BLCA, BRCA, CESC, HNSC, KIRP, LGG, LUAD, PAAD, PRAD, STAD, UCS	The datasets contained somatic mutations and were taken from TCGA.	10-fold cross validation Accuracy	Only evaluates with Accuracy on an unbalanced dataset. Only uses somatic point mutations. Comparison with SVM, KNN, NB only.	The formulated dataset showed an improvement in performance in every method. Overall Accuracy of 60.1% for the DNN with the enhanced dataset.
Danaee et al. (2017)[28]	Breast cancer detection from gene expression data. Feature extraction to identify risk genes.	Use of a Stacked Denoising Autoencoder to extract features. Use SMOTE to balance samples. Evaluate the impact of the extracted features using ANN and SVM.	BRCA	RNA-seq expression data from TCGA. 1097 cancer samples and 113 healthy samples.	5-fold cross validation Accuracy, Sensitivity, Specificity, Precision, F-measure	Considerably small sample. Only ANN and SVM classifiers.	The SDAE extracted important features for classification. The Accuracy was 91,74% for the ANN and 94.78% for the SVM.
Xiao et al. (2018)[136]	Prediction of cancer using a multi-model deep learning-based ensemble strategy	Feature selection using DESeq[99]. The first stage of classifiers with kNN, SVM, DT, RF, GBDT. A second stage ensemble model.	LUAD, STAD, BRCA	RNA-seq datasets from TCGA.	5-fold cross validation Precision, Recall, Accuracy, AUC	Lacks validation. Cross-validation strategy not optimal.	The results indicate the ensemble method increases the accuracy over the baseline methods. Accuracy for: LUAD 96,8%; STAD 96.59%, BRCA 95.76%

Table 3.2: Comparison between models for predicting Cancer Susceptibility II

Reference	Focus	Approach	Cancer Type	Data	Metrics and validation	Issues and Limitations	Conclusion
Krempel et al. (2018)[79]	Development of a cancer database, across multiple data types that can be used for predictive subtype and cancer classification.	Demonstrates usage of the database, using RF to predict cancer class types.	All of the existent in TCGA	Includes WXS data from 9091 patients from TCGA	Accuracy	Although it is not the core of the paper, the prediction example is missing validation as well as proper evaluation metrics.	Demonstrated that the system can be used for cross-data type queries useful for cancer prediction.
Sun et al. (2019)[124]	Distinguish cancer tissues from healthy tissues. Identify cancer type from 12 types of cancer.	Identify somatic mutations using Mutect2[23]. Feature selection by ranked points. DNN classifier. Model optimization with backpropagation and gradient descent.	BLCA, BRCA, OV, COAD, GBM, KIRC, LGG, LUSC, PRAD, SKCM, THCA	6083 WXS samples from various cancer types from TCGA. 1991 healthy samples from 1000 Genome.	10-fold cross validation Accuracy, Sensitivity, Specificity	Lacks comparative models for validation.	The accuracy, sensitivity and specificity of the total-specific model was 94.70%, 97.30%, 85.54%
Ahn et al. (2019)[2]	Identification of cancer from healthy tissue using gene expression data.	Classification of cancer or normal samples using LR, EN, LASSO, SVM and DNN. Simple feature selection strategies, limiting the number of genes.	BRCA	Gene expression data, (microarray and RNA-seq). A total of 6,703 cancer samples from TCGA, GEO, GTEx, TARGET. And 6422 healthy from the same sources.	Accuracy	Lacking in evaluation. and error measures. Feature selection by using simple measures.	The technique of limiting the number of genes showed a small improvement in performance. An accuracy of 97.22% vs 97.77%
De Guia et al. (2019)[29]	Predict cancer by using gene expression data and converting to 2D.	Transform 1D gene-expression data to 2D. Use of kNN, SVM, RF, CNN to classify.	All of the existent in TCGA	10267 gene expression samples from TCGA.	10-fold cross validation Accuraccy, Precision, Recall, F1	No feature selection.	The CNN approach was validated with an accuracy of 95.43%, showing an improvement over the baseline models.

Table 3.3: Comparison between models for predicting Cancer Survival

Reference	Focus	Approach	Cancer Type	Data	Metrics and validation	Issues and Limitations	Conclusion
Chen et al. (2014)[18]	Risk classification of cancer survival, with microarray and clinical data	Feature selection of 10 genes with a chi-square test to identify genes/survival correlation. ANN for prediction.	NSCLC	440 NSCLC samples from 4 different institutions.	Accuracy, P-value	Lacking in evaluation and error measures. Lacks comparative models. Not available data sources.	Overall accuracy of 83%. Identified potential risk genes
Chaudhary et al. (2018)[15]	Predict survival in Liver Cancer using multi-omics data.	Extract features from multiple types of data using DSAE. Select features, evaluating with Cox-PH and applying K-Means. Use SVM to predict years.	HCC	360 samples of RNA-seq, miRNA-seq, DNA methylation and clinical data from TCGA	C-Index,Brier Score, Log-rank P	C-Index<0.8. No comparative models. Small training data.	Concluded that adding clinical information did not help with multi-omics. Obtained C-indices of: HBV:0.74, HCV: 0.69, alcohol: 0.79, others: 0.77.
Wang et al. (2020)[49]	Survival prediction, using multi-omics and clinical data, based on a graph convolutional network.	Feature selection with mRMR [123]. Build a similarity matrix with SNF. Semi-supervised learning with spectral graph convolution. Compared with KNN, LR, DT and SVM. Tried all combinations of data	BRCA, LUSC	249 BRCA, 220 LUSC samples with multi-omics and clinical data from TCGA	Precision, Accuracy, Recall, AUC.	Small sample size.	Best results with multi-omics and clinical data, even with the baseline-methods. AUC for: BRCA 0.928 LUSC 0.805

Chapter 4

Dataset Analysis

This chapter introduces the two datasets we will be using and examines them by doing an exploratory statistical analysis for preliminary data understanding. A more in-depth analysis will be done onward on the data preparation section 5.1.1 but here we are interested in visualising the big picture. On this work, we use gene expression quantification and somatic mutation data from The Cancer Genome Atlas (TCGA) project. Different techniques might lead to biases in the collected data so we will focus on data coming from one data source using the same sequencing platform [1].

4.1 The Cancer Genome Atlas Project

The dataset we are using comes from the TCGA project, and it is accessible on the National Cancer Institute (NCI) Genomic Data Commons (GDC) data portal [2] upon open access. The dataset contains a total of 517 LUAD and 501 LUSC patients, and the quantity of gene expression samples outweighs the number of clinical records as the same patient can have both primary tissue samples (TP), and normal tissue samples (NT), which are non-cancerous samples taken from biopsy outside of the tumour region. We divide the statistical analysis into separate sections for better understanding of each histologic subtype:

- **Clinical data for the patients**: On Tables 4.1, 4.2, 4.3 it is presented a summary of the key sociodemographic variables, the macro survival statistics and variables of patients subject to therapy, respectively. *(N)* represents the total count for a variable and *(mean* $\pm$ *s.d.)* the average value and standard deviation for a variable's distribution.

- **Gene Expression and Mutation**: Figures 4.1, 4.6 present a gene expression subtype characterization for LUAD and LUSC subtypes, respectively. Figures 4.2 and 4.3 present an analysis of LUAD mutational data and Figures 4.7, 4.8, 4.9 for LUSC.

- **Preliminary survival analysis**: Figures 4.4, 4.5 present a preliminary survival analysis for LUAD and Figures 4.10, 4.11 for the LUSC histologic subtype.

[1] Illumina HiSeq 2000 (https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf)

[2] TCGA data portal (https://portal.gdc.cancer.gov/)

Cancer Subtype	Gene Expression Samples		Mutation Samples	Age	Gender		Race					Tumour Stage				
	TP	NT	(N)	(mean±s.d.)	Male	Female	A. Indian	Asian	Black	White	NA	I	II	III	IV	NA
	(N)				(N)		(N)					(N)				
LUAD	539	59	208,181	67.2 ± 8.6	239	278	1	8	52	390	66	277	122	84	26	8
LUSC	502	51	181,117	65.3 ± 10	371	130	0	9	30	349	113	244	162	84	7	4

Table 4.1: Summary of macro clinical variables and data sample counts for LUAD and LUSC histologic subtypes

Cancer Subtype	Diagnosis Age	Vital Status		Survival in Months
	(mean±s.d.)	Alive	Dead	(mean±s.d.)
		(N)		
LUAD	65.6±9.6	330	187	31.9±32.3
LUSC	66.9±8.6	284	217	34.5±32.8

Table 4.2: Summary of macro survival variables for LUAD and LUSC histologic subtypes

Cancer Subtype	Number of Treatments	Radiation Therapy			Neoadjuvant Therapy		
	(mean±s.d.)	Yes	No	NA	Yes	No	NA
		(N)			(N)		
LUAD	0.670± 0.481	70	370	77	2	472	43
LUSC	0.639± 0.467	42	381	79	4	486	11

Table 4.3: Summary of key therapy variables for LUAD and LUSC histologic subtypes

4.1.1 Lung Adenocarcinoma

4.1.1.1 Clinical Dataset

The LUAD dataset presents a reliable clinical sample with a small number of missing values for each parameter. A full list of the clinical parameters and its missing values is present in the appendix Table A.1, and on Table 4.1 we present the socio-demographic data for the LUAD patients.

The population presents a reasonably balanced gender distribution with a 54% female majority. All the patients are American citizens, and the population is prominently white. The average age of the LUAD patient's is 67.2 years old, and the age mean when diagnosed with cancer is 65.6. Regarding the clinical variable *Tumour Stage*, it is defined according to The Surveillance, Epidemiology, and End Results (SEER) and The American Joint Committee on Cancer (AJCC) Tumour, Nodes and Metastasis (TMN) classification [71], as a staging system to prognosticate cancer patients. The population follows a left-skewed Gaussian distribution peaking at pathologic stage I and with a minimum at stage IV.

4.1.1.2 Gene Expression Dataset

Regarding the gene expression dataset, a total of 598 LUAD samples were retrieved, with expression values for 20,531 genes, out of which 59 refer to normal tissue samples.

On Figure 4.1, it is presented the results of work from The Cancer Genome Atlas Research Network [35], [39], where the authors perform expression subtype detection to differentiate molecular subtypes of lung adenocarcinoma on the TCGA cohort. Molecular subtyping refers to the use of genomics data to find clusters of tumours within a histologic subtype of cancer, that have shared characteristics [71]. The molecular subtype is not part of the pathology report and is not used to guide treatment [68].

Results showed that the set of genes SFTPC, DMBT1, FOLR1, DUSP4, FGL1, TDG, PLAU, GOS2 and CXCL10 allow for differentiation of the three LUAD molecular subtypes: Bronchioid, Magnoid and Squamoid. The study assessed the mutation profiles, structural rearrangements, copy number alterations, DNA methylation, mRNA, miRNA, and protein expression of 230 lung adenocarcinomas. Cluster analysis was performed, enabling further refinement in sub-classification of LUAD, for the improved personalisation of treatment of this highly molecularly heterogeneous disease [35].

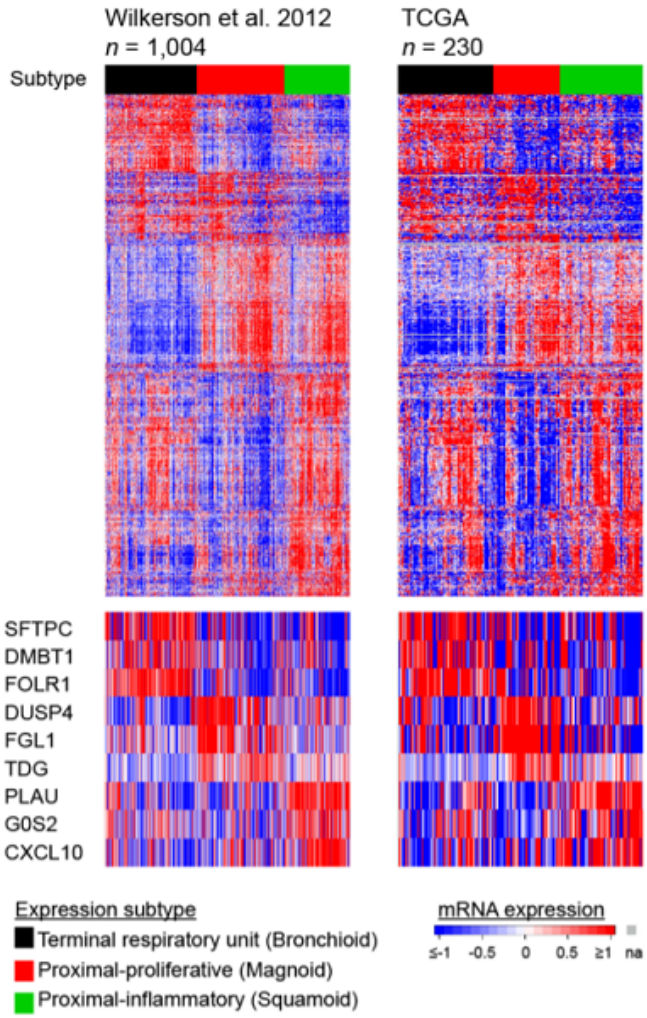
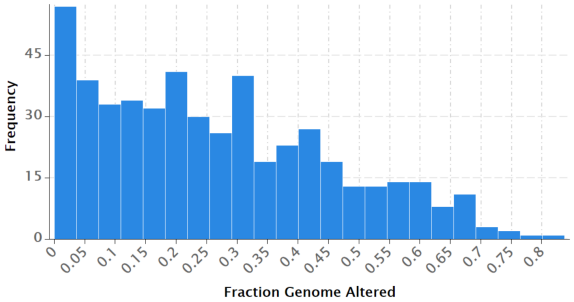


Figure 4.1: Analysis of molecular subtypes of LUAD using gene expression. Adapted from [35].

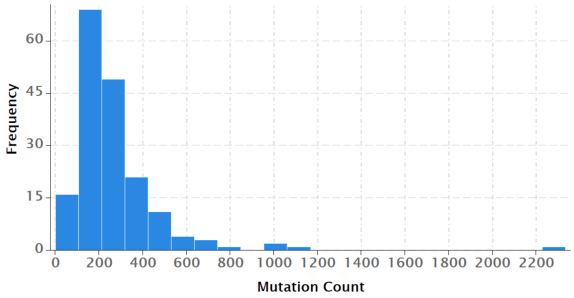
4.1.1.3 Mutation Dataset

In this section, we examine the mutational data for LUAD. Figure 4.2a presents the fraction of genome altered, which refers to the percentage of the genome affected by copy number gains or losses. Figure 4.2b presents the mutation count, that refers to the total number of mutations found in the tumour genome. On Figure 4.3 we use a Treemap diagram to show the frequency of mutations, that are considered oncogene mutations 4.3a and not oncogene mutations 4.3b, according to the OncoKB database [105].

The fraction of genome altered for LUAD patients represented in Figure 4.2a has a minimum value of 0.1; a median value of 0.24, corresponding to 24% of the genome altered; and a maximum value 80%. The total number of mutations found in the tumour genome is represented in Figure 4.2b, except for five patients with missing values; the minimum number of mutations is 2, the maximum is 2083, and the median value of mutations is 191.

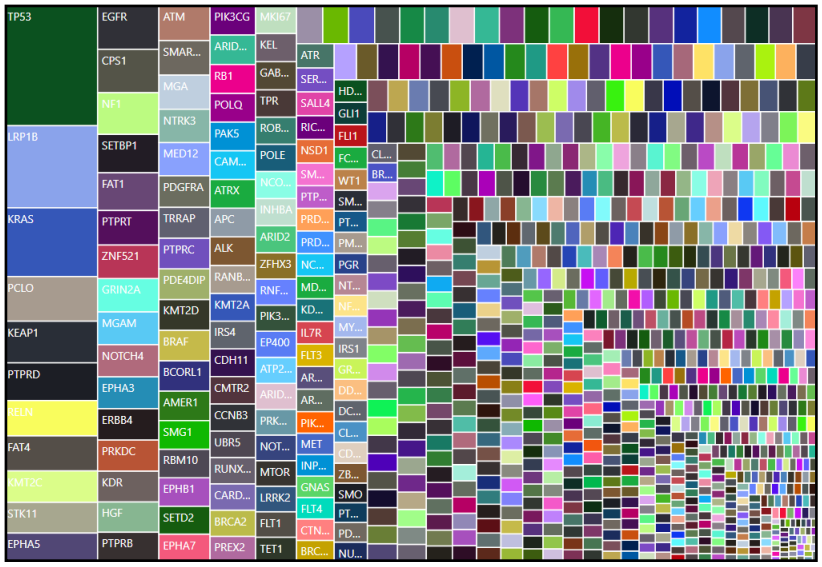


(a) Frequency of tumour genome alteration



(b) Total number of patient's mutations

Figure 4.2: Fraction of genome altered and occurence of mutations by patient.



(a) Possibly oncogenes.

(b) Mutated but not oncogenes.

Figure 4.3: Frequency analysis of mutated genes for LUAD.

Regarding the LUAD mutation data, there is a total of 208,181 mutation samples, containing a pool of 17,288 distinct mutated genes. We use the OncoKB database [105] to filter the genes according to their classification of oncogene negative or positive. This classification is based on their inclusion in various sequencing panels, and more information about it can be found in Vogelstein et al. [130]. It resulted in 1,004 genes to be considered oncogenes (5.81%) and 16,284 not. It is important to note that this oncogene classification does not necessarily tell us that the specific mutated gene contributes to cancer (driver gene) or, that it is a specific gene mutation associated with lung cancer.

Figure 4.3 presents two treemaps that show the average frequency of each mutated gene, for the oncogene positive and negative pool, respectively, and in both maps, it is not displayed the full gene count due to the extensive number of genes. On Figure 4.3a containing the classified oncogenes, the gene count is capped at ARID3A gene (0.6% frequency), and on Figure 4.3b it is capped at 4% frequency corresponding to the ABI3BP gene.

According to the National Comprehensive Cancer Network (NCCN) [101] and the European Society for Medical Oncology (ESMO) [46] clinical practice guidelines for the diagnosis of NSCLC, for LUAD, it is suggested biomarker testing for genes EGFR, ALK, ROS1, BRAF and CD274. These genes are all present in our mutation pool. EGFR shows at 12th position with frequency 12.5%; BRAF at 36th with 7.2%; ALK in 52nd with 6.9%; ROS1 at 92nd with 4.9% and CD274 in the 787th position with 0.7%. These frequencies are relative to the oncogene pool only, and their overall frequencies in LUAD and LUSC patients in comparison with literature will be reviewed on the mutation data preprocessing stage on chapter 6.

4.1.1.4 Survival Analysis

This section performs preliminary survival analysis to explore the survival expectancy of each histologic subtype. The survival data on Table 4.2 shows that 187 subjects died in the study period and 330 remain alive. The 330 alive patients represent the study's right-censored subjects, as the event of interest (time of death) is unknown. LUAD patients have an average survival time of 989 days with a standard deviation of 1,001 days, revealing a very disperse distribution of survival time. Table 4.3 presents the previous clinical therapy records of patients. Patients who have been subject to previous therapy are a minority, only 13.5% of patients have been subject to radiation therapy, and less than 1% have had previous neoadjuvant therapy. Previous therapy is an important predictive indicator of survival, and parameters such as the type of therapy received and medicine administrated should be factored in for survival prediction.

On Figure 4.4, we use a Kaplan-Meier plot to estimate the LUAD survival curve with a 5% confidence interval. The Kaplan-Meier (KM) curve, explained in detail in section 6.3, estimates the survival function of a population within a given confidence interval. Figure 4.4 shows a rapid decrease in survivability in the first three years, starting to slow down after 1,200 days. After 3,500 days (10 years approximately), the curve starts to flat down; this is due to the lack of observed deaths, remaining only alive subjects at the study. The survival rate stabilises at 21.13% after 20 years, close to the theoretical survival rate of 20% for lung cancer. On Figure 4.5, we plot the

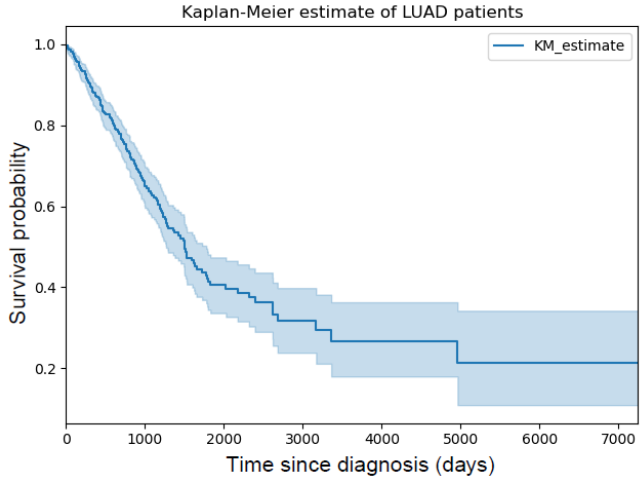


Figure 4.4: Estimate of the LUAD survival function using the Kaplan-Meier curve.

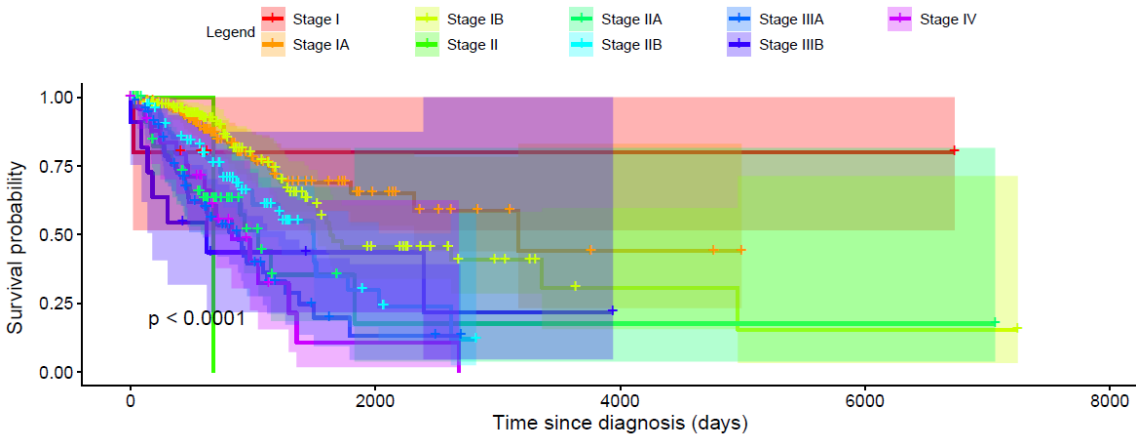


Figure 4.5: Survival analysis of the different pathological stage groups, with the KM estimate and the log-rank test with p-value of 0.0001.

impact of the covariate *ajcc_pathologic_stage* on survival groups, which represents the pathological stage of the subject according to AJCC classification. On this type of analysis, we evaluate the effect of a clinical variable on the subjects' survival, to assess if it significantly impacts patient's stratification. In this specific case, we evaluate the different pathological stage groups' survival. For each group, identified by a colour, a Kaplan-Meier curve is created that approximates the group's survival probability. The crosses in each curve indicate a right-censored (alive) observation in the group at time t, and the colour contour represents a confidence interval associated with the specific group survival. The test returns a p-value of the log-rank test that tells us that there is a difference between the pre-defined groups if it is below the 0.05 threshold, and with a p-value of 0.0001 we can conclude that the variable ajcc_pathologic_stage might be a fit candidate for the survival prediction.

4.1.2 Lung Squamous Cell Carcinoma

4.1.2.1 Clinical Dataset

The LUSC dataset presents a homogenous composition to the LUAD dataset, and there are only a few clinical variables that are mutually exclusive in both datasets, they are marked as NA when non-existent for a class in the appendix TABLE A.1, with the full clinical variables list. The gender distribution is more unbalanced than LUAD and contains a male majority, with 371 male (75%) to 127 female patients. The racial distribution is reasonably similar to LUAD, except there are no American Indian subjects. The average age of LUSC patient's is 65.3 years old, and the age mean when diagnosed with cancer is 66.9. The distribution of the pathological stage groups follows a similar distribution of the LUAD subtype, which indicates a presence of more subjects classified in lower cancer stages.

4.1.2.2 Gene Expression Dataset

There are 502 primary tumour LUSC samples and 51 normal tissue samples, with expression values for 20,531 genes that are homogeneous with the LUAD dataset. Figure 4.6 presents the results for a related work of Collison et al. [35], where the authors assess molecular differentiation of the LUSC histologic subtype. A total of 16 genes were selected as potential candidates to differentiate between the four molecular subtypes of LUSC: Classical, Primitive, Basal and Secretory.

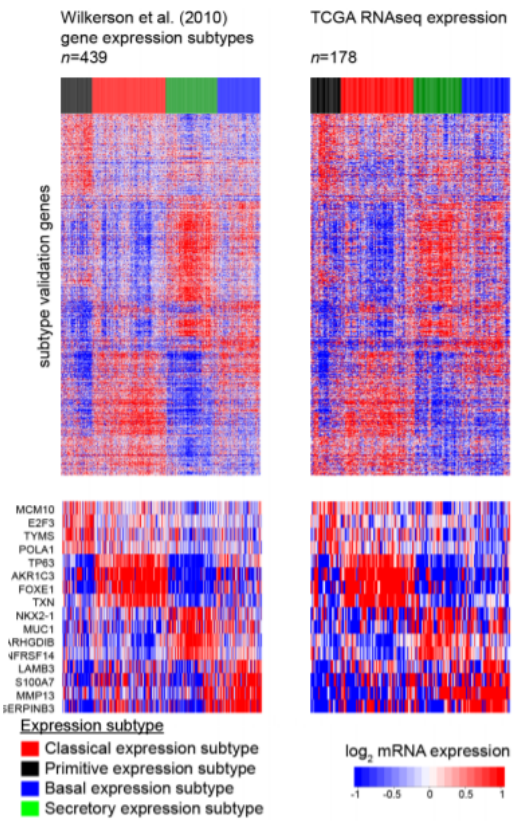
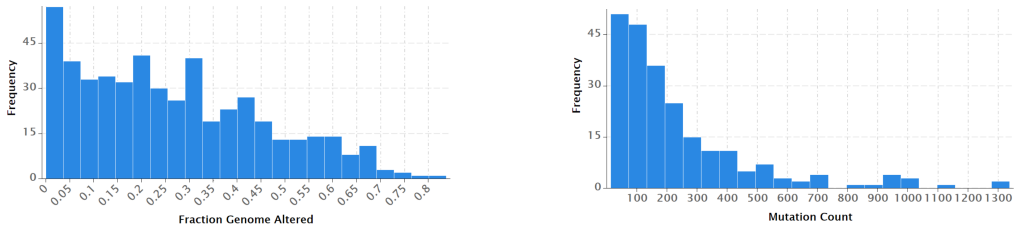


Figure 4.6: Analysis of molecular subtypes of LUSC using gene expression. Adapted from [39].

4.1.2.3 Mutation Dataset

For the fraction of genome altered on Figure 4.7a, there are no patients with missing values; the median value is 39%, which is significantly higher than for LUAD; and the maximum value for genome altered corresponds to 94%. On Figure 4.7b the minimum number of mutations is 1; the maximum is 1,591; and the median value of mutations is 222 with four patients presenting missing values, which highlights that the LUSC histologic subtype is more prone to high mutability when compared to LUAD.



(a) Frequency of tumour genome alteration

(b) Number of total patient's mutations

Figure 4.7: Fraction of genome altered and occurrence of mutations by patient.

There are a total of 181,117 mutation samples for LUSC and a gene pool of 16,970 different mutated genes. The group of oncogenes contains 1,010 (5.95%) genes and 15,960 not-oncogenes. In Figure 4.8 containing the classified oncogenes, the gene count is capped at AURKB gene (0.2% frequency), and on Figure 4.9 it is capped at 3.9% frequency corresponding to the AASS gene. All of the biomarkers suggested for testing by the NCCN and ESMO guidelines are present in the LUSC pool. ROS1 shows at 27th position with frequency 10%; ALK in 124th with 4.1%; BRAF in 152nd with 3.1%; EGFR at 186th with 2.9% and CD274 at the 1008th position with 0.2%, relative to the oncogene pool.

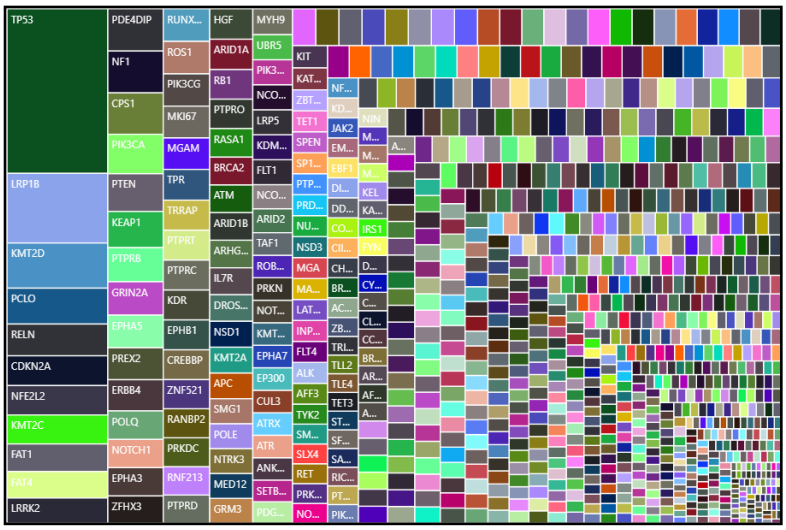


Figure 4.8: Frequency analysis of mutated oncogenes for LUSC.

Some of the main differences found between the LUAD and LUSC targeted biomarkers are:

- LUAD frequency ordering is EGFR>BRAF>ALK>ROS1>CD274 and LUSC is ROS1>ALK>BRAF>EGFR>CD274
- The targeted genes are more frequent in LUAD than in LUSC
- The most commonly mutated gene is TP53 in both cases, but in LUSC, the TP53 mutation is very prevalent 83.5% against 52.1% on LUAD.

This preliminary analysis gives us an idea of the difficulty of finding relevant biomarkers to identify lung cancer. Out of the pool of 17,288 mutated genes for LUAD; and 16,970 for LUSC, the ones currently considered viable markers for diagnosing cancer seem very infrequent even when considering only the pool of possible oncogenes, which means that most of the mutated genes, even when considered oncogenes might not act as driver genes and therefore are outliers to the problem.

4.1.2.4 Survival Analysis

For LUSC, there are 217 deceased subjects and 284 alive patients, and the outcome analysis for LUSC shows a more pessimistic scenario than for LUAD. The patients have an average survival time of 1,070 days with a standard deviation of 1,017 days, revealing a disperse distribution of survival time and patients who have been subject to previous treatment also represent a minority.

Figure 4.10, the Kaplan-Meier plot seems to confirm the more pessimistic scenario compared to LUAD. The curve is less steep in the first three years; however, the slope is more aggressive after the three year-period and stabilises at 18.58% after 13 years, showing worse survival rate than LUAD. These results seem to go in line with theoretical research, in Kawase et al. [41], where the outcomes of LUSC and LUAD were compared, and it was concluded there were considerable differences in overall survival were observed between the two histologic types. The

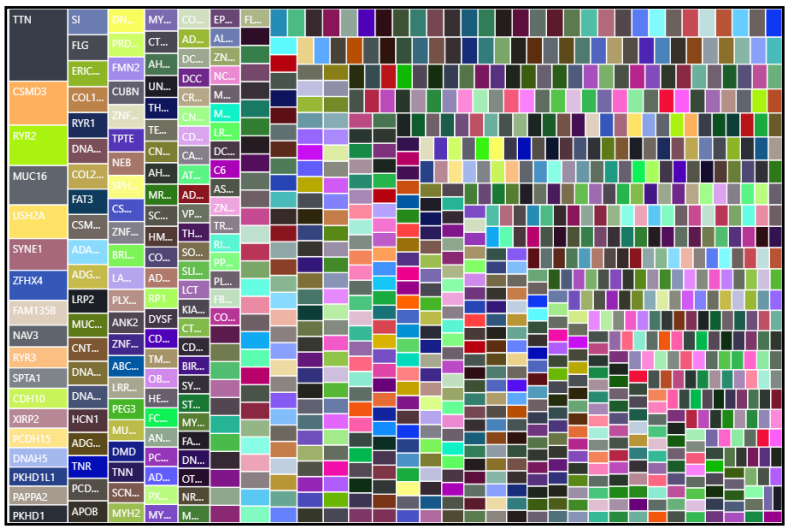


Figure 4.9: Frequency analysis of mutated but not oncogenes LUSC.

pathological stage groups' analysis on Figure 4.11 shows a similar pattern then the LUAD study, and the log-rank test returns a p-value of 0.00023, which confirms a statistical difference between the pathological stage groups' survival. As per LUAD, there is no perfect correlation between the higher pathologic stages and the lower survival; however, this is expected because the AJCC pathological stage classification is not causal to patients' survival time, it is only an estimate of the degree of severability of the disease.

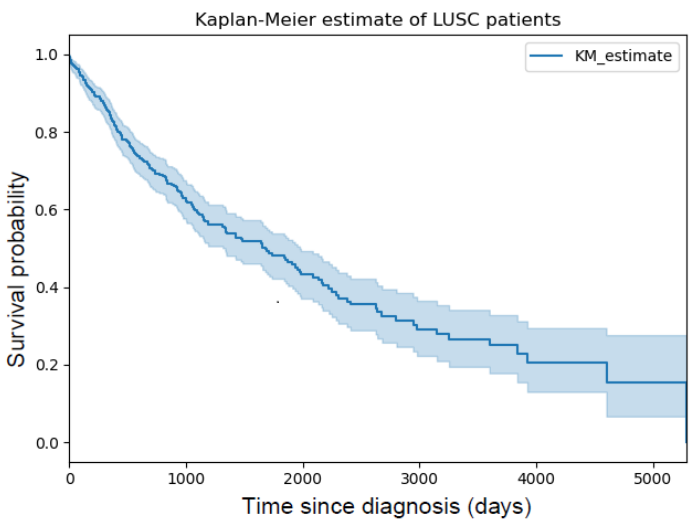


Figure 4.10: Estimate of LUSC survival function using the Kaplan-Meier curve.

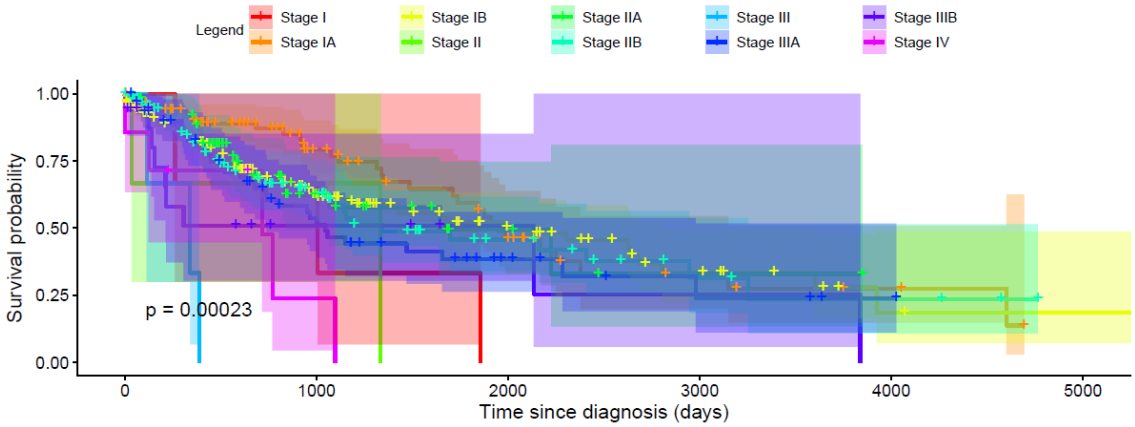


Figure 4.11: Survival analysis of the different pathological stage groups, with the KM estimate and the log-rank test with p-value 0.00023.

Chapter 5

Lung Cancer Prediction and Subtype Classification

The work developed in this chapter aims to find the best models to perform cancer prediction and subtype classification while providing interpretability on results, allowing us to identify gene expression signatures that better differentiate cancerous from non-cancerous gene expression samples and LUAD from LUSC subtypes. Cancer prediction relates to the process of differentiating cancerous from normal tissue, whereas histologic subtype classification differentiates groups within the same type of cancer, based on specific characteristics of the cancer cells.

On section 5.1, we start by giving an overview of the experimental design and describe the data preparation stage. On sections 5.2 and 5.3, we present the materials used and validate our choices for the proposed models. On section 5.4, we briefly cover the evaluation metrics used to assess the models' predictive performance. On section 5.5, we present the performance results for the cancer prediction and subtype classification problems, as well as additional analysis of the results. On section 5.6, we discuss our results comparing with state of the art and validate our research methodology.

5.1 Experimental Design

Figure 5.1 gives an overview of the experimental design for the cancer prediction and subtype classification problems. In the data preparation stage, we perform data cleaning and standardise the data for the modelling stage. In the supervised learning stage, we employ two different types of models for cancer and subtype classification: gradient boosted tree models that we provided explainability on to provide biological insight; and DL models, which are the standard strategy in state of the art. For the tree-based approach, two sets of genes were extracted: a set that differentiates normal from cancerous tissue gene expression samples (*Set 1*), and a set of genes that distinguishes LUAD from LUSC (*Set 2*).

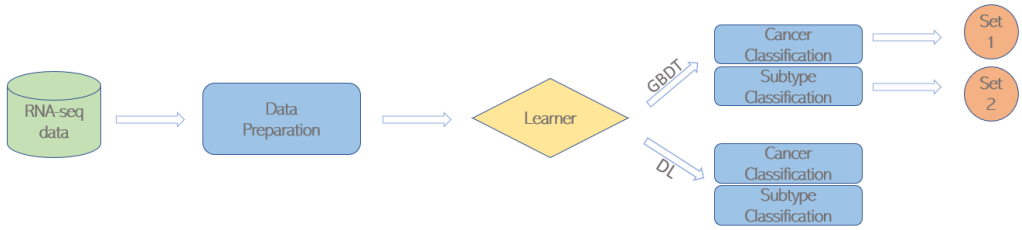


Figure 5.1: The workflow of the cancer prediction and subtype classification problems.

5.1.1 Data Preparation

Using R BioConductor framework with the TCGABiolinks package [34] [117] [43], we queried for gene expression quantification data from the TCGA project. The data was retrieved from the GDC legacy database and entailed tier 3 data of the GDC workflow, which is post-normalisation and aggregation. We proceed to add clinical information for the patients, using TCGABiolinks function *GDCprepare*, and removed duplicate patient records.

The nomenclature for gene identifiers in TCGABiolinks is the Human Genome Organisation (HUGO) symbol by default. Some errors were detected, namely duplicated gene names corresponding to different entries on distinct databases. Therefore, according to the gene metadata instance on Figure 5.1, all identifiers were renamed as a combination of their HUGO symbol and Entrez Gene, which is a unique identifier for genes across databases.

entry	seqnames	start	end	width	strand	hugo_symbol	entrezgene	ensembl_gene_id
11175	chr5	1.41E+08	1.41E+08	3274	+	PCDHB10	56126	ENSG000120324
11176	chr5	1.41E+08	1.41E+08	3274	+	PCDHB10	56127	ENSG000120324

Table 5.1: Example of duplicated gene_id entries from the TCGABiolinks generated gene's metadata file. Entrez Gene 56126 conrresponds to gene PCDHB10 and 56127 to PCDHB9.

The TCGA barcode is the primary identifier of biospecimen data within the TCGA project. It is composed of a collection of identifiers, and each identifies a TCGA data element explicitly. Figure 5.2 presents an illustration of a TCGA barcode. A vial identifies the order of sample in a sequence of samples; the portion refers to the order of portion in a sequence of 100-120 mg sample portions; the plate is the order of plate in a sequence of 96-well plates [70].

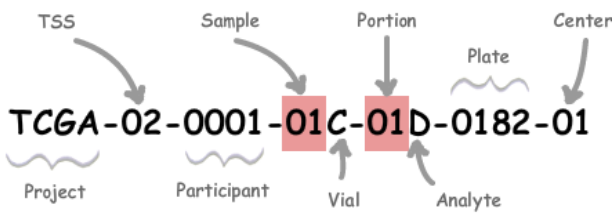


Figure 5.2: Illustration of metadata identifiers that comprise a TCGA barcode. Adapted from [70].

Upon analysis of the gene expression records, we identified two different kinds of "duplicate" samples belonging to the same patient:

1. Samples with the same vial and portion but different plate - This represents duplicate samples tested on different plates for reproducibility. Therefore we averaged these samples' expression values.

2. Samples with a different vial - Corresponds to samples from different regions of the tumour. We maintained these samples as they show far-apart expression values and can be considered different instances.

After the preprocessing steps described above, the final LUAD dataset has 575 samples out of which 60 (10.43%) are normal tissue (NT) samples. The LUSC dataset has 502 primary tumour samples (TP) and 52 (9.39%) NT samples. Table 5.2 presents an analysis of the expression values distributions' for LUAD and LUSC grouped by tissue type.

Subtype	***LUAD***		***LUSC***	
Tissue Type	***TP***	***NT***	***TP***	***NT***
Min	0	0	0	0
Max	1,432,694.16	1,324,252.01	1,737,511.59	1,036,891.47
Average	969.57	1024.19	984.87	1049.01
S.d.	695.85	353.83	678.58	375.58

Table 5.2: Analysis of minimum, maximum, averange and standard deviation of LUAD and LUSC gene expression values by tissue type.

The expression analysis on Table 5.2 highlights differences between histologic subtypes and between tissue types as it would be expected. The minimum number of expression for both subtypes and tissue types is 0, which translates to no measurable expression for that gene. An average of 4,896 genes present expression values close to 0 for NT type, and an average of 7,052 genes present close to null expression for LUAD and LUSC subtypes. The maximum expression value belongs to the LUSC subtype, relating to the ADAM6 gene and on LUAD a value of 1,432,694.00 associated with the SFTPB gene. Both LUAD and LUSC primary tumour samples present higher maximum values than their normal tissue counterparts. Regarding the aggregated expression values, normal tissue types show higher average expression values and lower standard deviations than LUAD and LUSC primary tissue samples, which tells that the distributions of expression are more disperse for the primary tumour types. The average expression value is relatively close between histologic subtypes, showing disperse distributions with similar standard deviations.

Regarding the ratio of NT to TP samples, the cancer classification problem poses a different challenge than the subtype classification problem, as there is a label skew with a 9:1 ratio class imbalance of the positive class. Label skew happens when labels in a supervised learning problem are present in different frequencies [88]. We use two different approaches to tackle this: adjusting the weights of the labels in the learners, and using ADAptive SYNthetic (ADASYN) [55] sampling approach to oversample the negative class.

5.2 Deep Learning

Representation learning is a set of methods that allow a machine to be fed with raw data and automatically discover the representations needed for detection or classification. DL methods are representation-learning methods with multiple representation levels, obtained by composing simple but non-linear modules that transform the representation at one level (starting with the raw input) into a higher, slightly more abstract level, such that very complex functions can be learned [81]. With the rapid advances in computational power, deep learning architectures have shown to outperform traditional ML methods at various tasks in the bioinformatics domain such as predicting the effects of somatic mutations in non-coding DNA [82], inferring the values of gene expression [19], predicting survival based on omics data [66]. However, most cancer datasets still pose the fundamental problem of small datasets with many variables, which can lead to overfitting, particularly on deep learning architectures due to the curse of dimensionality. For smaller data sets, unsupervised pre-training has shown to prevent overfitting, leading to significantly better generalisation when the number of labelled examples is small [8].

According to the task, there are several DL architectures capable of achieving satisfactory performance: DFF networks are widely used for the task of cancer prediction as shown in literature review section 6.2.3; CNN's are generally used in computer vision, and image recognition, but have shown potential to embed 2D representations of mRNA data [29]; Stacked Autoencoders (SA), Restricted Boltzman Machines (RBM) and Deep Belief Networks (DBN) have shown promising results for feature extraction of gene expression data [28], [15], [127].

We established a deep feedforward network, whose general architecture is represented in Figure 5.3. As input, we feed the set of selected normalised genes, then a combination of dense layers with dropouts are used to regularise the weights and avoid overfitting, and finally, the output layer represented by the grey rectangle consists of a single neuron with sigmoid activation. Previous works, highlight the use of rank-variance feature selection techniques to reduce the risk of overfitting when using DNN's for the tasks of cancer prediction [2], [136]. With no feature selection,

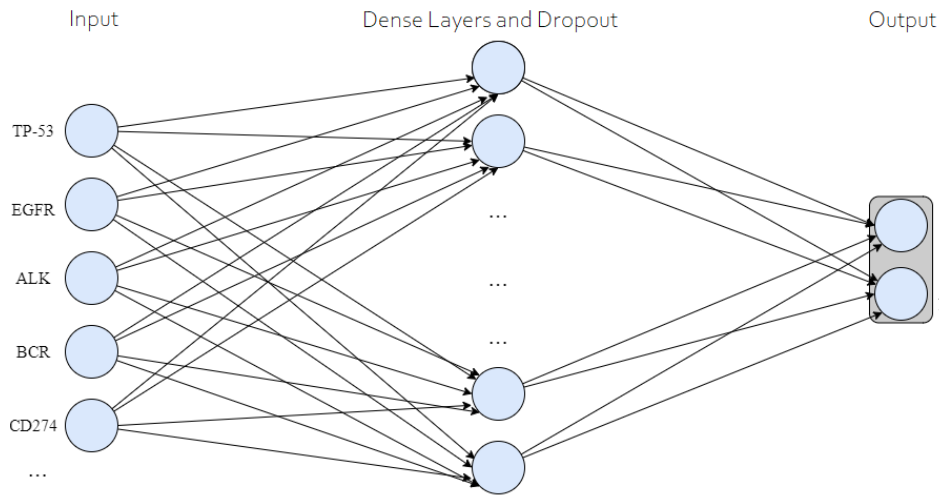


Figure 5.3: General architecture of the DFF network with dropout layers for regularisation.

there are 20,531 inputs for the neural network, to validate the effectiveness of feature selection methods, we implemented three feature selections techniques described below:

Welch's t-test - X_1 indicates the target group and X_2 the control group, σ denotes the group's standard variation and # the sample size. A significance level of α=0.05 was chosen, and all features with t value higher then critical values of the t distribution were chosen as significantly different [73].

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{[\frac{\sigma_{X_1}{}^2}{\#X_1}] + [\frac{\sigma_{X_2}{}^2}{\#X_2}]}} \tag{5.1}$$

Pearson's Correlation - The Pearson correlation calculates the degree to which two variables are correlated. We calculated the Pearson coefficient of each gene and selected the N genes with coefficients closer to 0.

$$r = \frac{[X_1 - \bar{X}][X_2 - \bar{X}]}{\sqrt{(X_1 - \bar{X})^2 (X_2 - \bar{X})^2}} \tag{5.2}$$

Fisher Correlation Coefficient - The fisher score algorithm selects each gene independently based on their scores under the Fisher criterion [85], and the top N genes with lower coefficients were selected.

$$r = \frac{\bar{X_1} - \bar{X_2}}{\sigma_{X_1} + \sigma_{X_2}} \tag{5.3}$$

We used a grid search to determine the number and size of the dense layers and their dropout rates, which are essential to control overfitting and have shown to provide improvements in predictive performance [2], [29], [15]. Furthermore, the network hyper-parameters such as the activation functions, the momentum optimiser, batch size and the learning rate were also assessed using 5-fold cross-validation.

5.3 Gradient Boosting Decision Trees

Tree-based learning algorithms represent a broad class of ML methods for regression and classification first introduced from a statistical background on Breiman et al. [12]. Some characteristics that make tree-based methods popular over other ML models like NN's or SVM's are presented in Hastie et al. [60]: robustness to outliers in the input space; ability to deal with irrelevant inputs; can handle missing values in the predictor variables; can handle both continuous and categorical variables; are insensitive to monotone transformations of inputs and can provide interpretability.

Although tree-based learning algorithms have many desirable characteristics, one aspect that prevents them from being ideal is the predictive performance when compared to NN's [26]. Tree-based ensembles, combine the predictions of many different trees to give an aggregated prediction [30], and this type of strategy allows to combine predictions of multiple weak learners which can give improved prediction accuracy over individual trees. Random Forests (RF) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [11]. Boosted Trees (BT) are a type of tree ensemble whose idea is to form an ensemble by fitting trees to weighted versions of the data which are combined by weighted voting or averaging [60].

A Gradient Boosted Machine (GBM) is a generalisation of tree boosting where a general gradient descent "boosting" paradigm is developed for additive expansions based on any fitting criterion [50]. Gradient Boosting Decision Trees (GBDT) or GBM's have various implementations such as Extreme Gradient Boosting Machine (XGBoost) [17], Parallel Boosted Regression Trees (pGBRT) [6]. Although these are popular machine learning algorithms, the efficiency and scalability are still unsatisfactory when the feature dimension is high, and data size is large [38].

The classifier used in this work is the light gradient boosting machine (LightGBM), which attempts to fix this problem by implementing two innovative techniques: Gradient-based One-Side Sampling (GOSS) which excludes a significant proportion of data instances with small gradients, and Exclusive Feature Bundling (EFB) that bundles mutually exclusive features, therefore, reducing the number of features [38]. It is well known that a specific model's predictive performance is highly dependent on the characteristics of the input data. On Wang et al. [131], a comparative study on the efficiency of XGboost and lightGBM models for miRNA classification on breast cancer patients is performed. Results showed that the lightGBM algorithm performed better overall, showing higher accuracy and precision but higher losses then XGBoost. Although the authors do not extrapolate on the possible causes of lightGBM performing better, we will describe our hypotheses to justify it outperforming other algorithms on this domain:

Optimum split points: GBM vs Pre-sorted/Histogram-based vs GOSS Algorithm

A crucial step of GBDT is finding the splits that maximise information gain. Most existing single machine tree boosting implementations, such as R's GBM [110], use the exact greedy split fiding algorithm to enumerate all the possible splits for continuous features [17]; however, this is not feasible for large-scale datasets.

Histogram-based algorithms split all of the data points of a feature into discrete bins to find the histogram's best split value. XGBoost uses a sparsity-aware split finding algorithm on top of the pre-sorted algorithm in order to reduce the training cost by ignoring the features with zero values [17]; however, GBDT with the histogram-based algorithm does not have efficient sparse optimisation solutions. GOSS addresses these limitations by excluding a significant proportion of data instances with small gradients which allows for similar information gain with a much smaller data size [38].

Reducing the number of features: Exclusive Feature Bundling

EFB provides a nearly lossless approach to reduce the number of effective features by bundling mutual exclusive features into a single feature called exclusive feature bundle [38]. Cancer genomics datasets generally present sparse characteristics and can benefit from this. Furthermore, as seen in chapter 3, the general approach in state of the art is to use various feature selection techniques prior to DL learners to improve predictive performance. We argue that this might leave out essential features not selected by feature selection techniques, and EFB allows the model to decide what features are mutually exclusive eradicating the need for apriori feature selection, therefore, addressing previous work limitations related to dimensionality reduction.

LightGBM uses the leaf-wise tree growth algorithm, while many other popular tools use depth-wise tree growth. Compared with depth-wise growth, the leaf-wise algorithm can converge much faster, however, the leaf-wise growth may be overfitting if not used with the appropriate parameters [140]. LightGBM has a wide range of hyperparameters that require tuning, and it is essential to find a combination of hyperparameters so that the model performs well while being able to generalise.

5.3.1 Bayesian Optimisation

The Bayesian Optimisation in Algorithm 1, in which a learning algorithm's generalization performance is modelled as a sample from a Gaussian Process (GP), tries to find the minimum of a function $f(x)$ on some bounded set x. The difference from traditional methods such as randomized search is that it constructs a probabilistic model for $f(x)$ and then exploits this model to make decisions about where in x_i to evaluate the function next [119], which decreases the cost of finding the solutions when the black-box function f is complex, which is the case of more elaborate ML models.

Algorithm 1: Bayesian Optimization

```
for i ← 1, 2,... do
    Find new step by optimizing acquisition function α over
        GP: x_{i+1} = arg min_x α(x | D_i);
    Sample the objective function to obtain y_{i+1} = f(x_i) + ε_i;
    Augment the training data D_{n+1} = {D_i, (x_{i+1}, y_{i+1})};
end for
```

A critical step of the optimiser is deciding on an acquisition function α that evaluates the utility of candidate points for the next evaluation of function f. The acquisition function's exploration-exploitation trade-off, and their optima are located where the uncertainty in the surrogate model is large (exploration) and/or where the model prediction is high (exploitation) [115].

$$\alpha\,\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \tag{5.4}$$

Expected Improvement (EI) represented on Equation 5.4 was chosen, as it is better-behaved than Probability of Improvement (PI), but unlike the method of GP Upper Confidence Bounds (GP-UCB), it requires less parameter tuning [119]. The expected improvement has two components, the first can be increased by reducing the mean function $\mu(\mathbf{x})$, and the second can be expanded by increasing the variance $\sigma(\mathbf{x})$. The meta-parameter ξ that defines the exploitation-exploration trade-off under the EI acquisition function was optimised by trial and error over five steps in the $[1e^{-4}, 1e^{-1}]$ search space.

The hyper-parameters were tuned using 5-fold cross-validation, and when performing Data Augmentation (DA), the oversampling was done for each fold to avoid data leakage. We optimised by minimising the objective function binary cross-entropy (logloss) represented on Equation 5.5.

$$L_{bce}(y, \hat{y}) = -(y\, log(\hat{y}) + (1-y)\, log(1-\hat{y})) \tag{5.5}$$

Cross-entropy is defined as a measure of the difference between two probability distributions for a given random variable or set of events [72]. Logloss is a calibration statistic and a measure of class separability that takes the "certainty" of classification into account, and this is especially relevant when designing a model to diagnose deadly disease, as we want to penalise uncertain decisions and not necessarily only improve performance.

5.3.2 Shapley Additive Explanations

To provide model interpretability, we used the SHapley Additive exPlanations (SHAP) technique [94]. This method explains individual predictions by estimating each feature's individual contribution to the corresponding prediction and, consequently, assigning it a SHAP value. Features with larger absolute SHAP values are more important for prediction and can positively or negatively impact the prediction depending on its sign.

TreeSHAP is a variant of SHAP for tree-based machine learning models such as decision trees, random forests and gradient boosted trees that solves the inconsistencies of previous feature attribution methods for tree ensembles, which failed to capture the positive impact of feature's importance when based solely on the additive feature attribution method [93].

5.4 Train, Test and Evaluation

Figure 5.4 represents the supervised learning pipeline followed for both problems, consisting of two different stages: an optimisation stage where we to optimise the DL model using grid-search, and Bayesian optimisation for LGBM using 5-fold cross-validation; and the classification stage, where the train/test process is repeated 100 times for model validation. The train and test splits of each problem are different due to the imbalanced nature of the cancer classification problem, so we selected the train and test sets as follows:

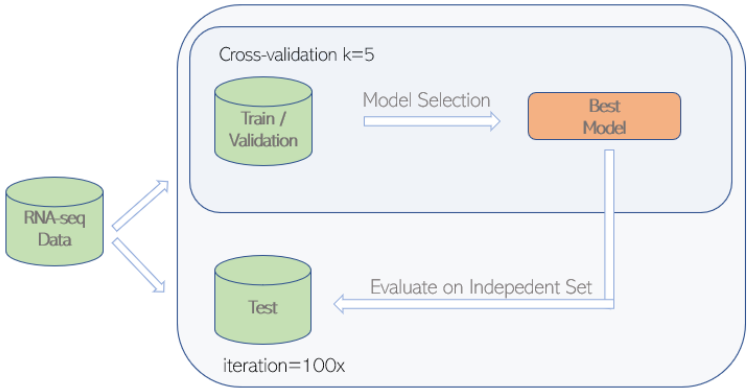


Figure 5.4: Global supervised learning pipeline for the DL and LGBM models. The optimisation is performed using 5-fold cross-validation, and the train test process is repeated 100 times for model validation.

- **Cancer Prediction**: A split ratio of {70,15,15}% was used for train, validation and test sets, respectively. Stratified sampling was used to maintain the 9:1 ratio of positive samples in the validation and test sets when not performing data augmentation. When conducting DA, standard shuffle split was employed, and the oversampling of the negative class was done for each fold of the optimisation process.
- **Subtype Classification**: A split ratio of {80,10,10}% was used for this problem. As opposed to the first problem, this one has a balanced ratio of labels, so there is no need to perform DA.

The larger size of the independent test size on the first problem is due to the labels' imbalance; therefore we need to guarantee a minimum amount of negative samples for support in the test and validation sets; otherwise, the predictive performance of the model for the negative class would incur in higher variance in between runs.

To estimate the models' performance stability, the train and test splits were executed 100 times, randomising the split seed to reduce variability and overcome skewness caused by the short sample size and class imbalance. The evaluation metrics' logloss and AUC were used to evaluate training performance and control overfitting by early stoppage. To evaluate the test set's performance, we used AUC, accuracy, precision, recall and f1-score metrics, to better assess our results against similar research.

Area Under the Curve: AUC is based on the area under the receiver operating characteristic (ROC) curve that plots the False Positive Rate (FPR) against the True Positive Rate (TPR) known as sensitivity or recall. AUC, much like logloss evaluates the model's degree of separability; however, AUC is a rank statistic that varies between minimum value of 0 and a maximum 1. AUC has desirable properties as a classification performance measure when compared to overall accuracy: its increased sensitivity in the Analysis of Variance (ANOVA) tests, it is not dependent on a decision threshold chosen, and it is mildly sensible to class imbalance [10].

Accuracy: Accuracy is given by the ratio of correct observations by overall observations. As opposed to AUC, this metric is biased when there is a label skew.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad , \quad 0 \leq Accuracy \leq 1 \tag{5.6}$$

Precision: Precision evaluates the fraction of correctly classified instances among all positive classified instances and can be seen as accuracy for the positive class. In the medical domain, its reasonably common to have negative minority classes, when assessing precision for the negative class it is called Negative Predictive Value (NPV) and allows us to understand how the model performs on the minority class.

$$Precision = \frac{TP}{TP+FP} \quad , \quad 0 \leq Precision \leq 1 \tag{5.7}$$

Recall: Recall or sensitivity is defined as the ratio between true positives and the sum of true positives and false negatives. The recall is typically used to measure the coverage of the minority class [61], and when calculated for the negative class, it is called specificity.

$$Recall = \frac{TP}{TP+FN} \quad , \quad 0 \leq Recall \leq 1 \tag{5.8}$$

F-measure: The F-measure or F1-score provides a single score to evaluate both precision and recall.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad , \quad 0 \leq F1 \leq 1 \tag{5.9}$$

Implementation: Our solution is implemented in Python 3.8 using the frameworks Keras [21], Tensorflow [42], lightGBM [38], [140], [32] and bayesian-optimization [102]. All CPU intensive tasks, namely preprocessing, were conducted on a 16x2GHz Intel Core Haswell CPU with 128Gib RAM, and all GPU taks were ran on a 2xNvidia GeForce GTX 1080 Ti with 11,178 MiB GDDR3 with the Ubuntu 18.04 LTS operating system.

5.5 Results

The work developed in this chapter aimed at finding the best models to perform cancer prediction and subtype classification. On section 5.5.1, we discuss the optimal hyper-parameters for the DL and LightGBM learners. On section 5.5.2, we present the best models' predictive performance for the cancer prediction problem, and on section 5.5.3 for the subtype classification. On section 5.5.4, we analyse the most relevant gene signatures retrieved from the LGBM model using TreeSHAP. Finally, on section 5.6 we discuss our results comparing them to the state-of-the-art and validate our results by interpreting the selected gene signatures according to mentions in relevant papers.

5.5.1 Hyperparameter Optimisation

5.5.1.1 Deep Learning

Table 5.3 presents the optimal values for the hyperparameters of the deep-feedforward network for that cancer prediction and subtype classification problems. For each problem, a 5-fold cross-validated grid-search is used to optimise the hyper-parameters over a parameter grid. The search space was defined as architectures with 1 to 6 dense layers with dropouts in the $[0, 0.5]$ range. The batch sizes were capped at 50 due to processing power, and the learning rate was optimised in the $[1e^{-6}, 1e^{-2}]$ range.

Hyperparameters	Cancer Prediction	Subtype Classification
layers	{100, 100}	{500, 200, 100}
dropout	{0.1, 0.4}	{0.2, 0.1, 0.5}
learning_rate	0.000163	0.000347
batch_size	32	40

Table 5.3: Hyperparameters for the deep-feedforward network.

The systemic assessment found that 2 hidden layers worked better for cancer prediction and 3 dense layers for subtype prediction. Rectified Layer Unit (ReLU) provided higher average performance's and Adaptive Moment Estimation (ADAM) optimiser provided better results, and the optimisation ends up going faster then using regular gradient descent and also avoids local optima [77]. Dropouts appeared to boost average performance by reducing variance between folds and the learning rate, and batch size was relatively close for both problems. Furthermore, for the cancer classification problem weights were assigned to each class according to Equation 5.10, where N_{C_0} and N_{C_1} represent the positive and negative class and dividing by 2 helps to keep the loss to the same magnitude [42].

$$w_{C_0} = \frac{1}{N_{C_1}} \cdot \frac{(N_{C_1} + N_{C_2})}{2} \quad , \quad w_{C_1} = \frac{1}{N_{C_0}} \cdot \frac{(N_{C_1} + N_{C_2})}{2} \tag{5.10}$$

5.5.1.2 LightGBM

For each problem, we ran approximately 10,000 steps of Bayesian Optimisation (BO) with 5-fold cross-validation. The optimal value for the meta-parameter ξ was fixed at $1e^{-2}$, which prioritises exploration over exploitation. Table 5.4 presents the optimal values for the hyper-parameters of the LGBM classifiers. The Bayesian optimisation's optimal step showed a cross-validation loss of 0.0005 for cancer classification and 0.0340 for subtype classification. Higher depth values showed better results for both problems, and the number of leaves was fixed at $2^{max\_depth} - 1$ in a preliminary stage and fine-tuned a posteriori.

Hyperparameters	Cancer Prediction	Subtype Classification
n_estimators	256	154
max_depth	6	8
learning_rate	0.1048	0.9173
feature_fraction	0.2673	0.7457
bagging_fraction	0.1067	0.4718
min_split_gain	0.0002	0.0141
min_child_weight	0.0057	0.0053
min_child_samples	5	17
reg_alpha	0.0140	0.0103
reg_lambda	0.1100	0.1368
scale_pos_weight	4.9604	2.3732
subsample_for_bin	256900	461045

Table 5.4: Hyper-parameters for the LightGBM model.

L1 (*reg_alpha*) and L2 (*reg_lambda*) regularisation were added to force sparsity and to diminish the value of the weights, respectively, which conferred a reduction in standard deviation across runs. The minimal number of data in one leaf (*min_child_samples*), which also prevents overfitting presented higher values in the subtype classification problem to compensate for higher depths, by avoiding to grow leaf-wise.

The parameters *bagging_fraction* that randomly selects part of data without resampling, and *feature_fraction* that selects a subset of features of each boosted tree, showed a tendency for higher values with growing depths, which is also a mechanism to force generalisation and avoid overfitting by not relying only on a small subset of variables. The *scale_pos_weight* parameter, that defines the weights of the labels in the learner, was used when not performing DA and showed higher values in the cancer classification problem considering the 9:1 ratio class imbalance of positive samples.

5.5.2 Cancer Prediction

5.5.2.1 Deep Learning

Table 5.5 presents the DL model' results when performing no feature selection, using the total of 20,531 genes for prediction. Two approaches were followed to tackle the 9:1 class imbalance of the positive class: each class's weights were corrected according to the total number of positive and negative samples at each iteration; using ADASYN to oversample the negative class. In Danaee et al. [28], SMOTE was used in a similar context to oversample non-cancerous samples; however, the results without DA are not provided for comparison. ADASYN was chosen over SMOTE as it can bias the sample space and avoid the selection of more alike neighbours that might minimise information gain [55]. When performing data augmentation, we oversampled each fold of the 5-fold cross-validation individually to avoid data leakage, and the number of neighbours was fixed at 15, which showed better optimisation results.

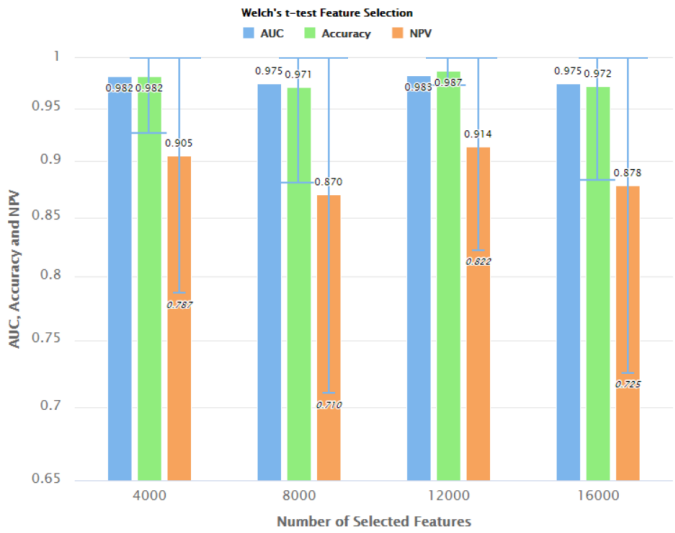
The DL model with weight correction seemed to perform better on every spectrum, with higher AUC, accuracy and even higher negative precision (NPV), which should be where the DA strategy should improve performance. We are not entirely sure why oversampling lead to worse results, even when varying the number of neighbours; however, according to the literature it is not a commonly used strategy.

	AUC	*Accuracy*		*Precision*	*Recall*	*F1-score*
DL	**0.984 ± 0.025**	**0.986 ± 0.011**	***Positive***	0.998 ± 0.004	0.989 ± 0.012	0.994 ± 0.006
			Negative	**0.912 ± 0.097**	0.977 ± 0.038	0.940 ± 0.047
DL w/ DA	0.982 ± 0.023	0.986 ± 0.016	***Positive***	0.997 ± 0.005	0.987 ± 0.018	0.992 ± 0.009
			Negative	0.908 ± 0.103	0.976 ± 0.048	0.936 ± 0.062

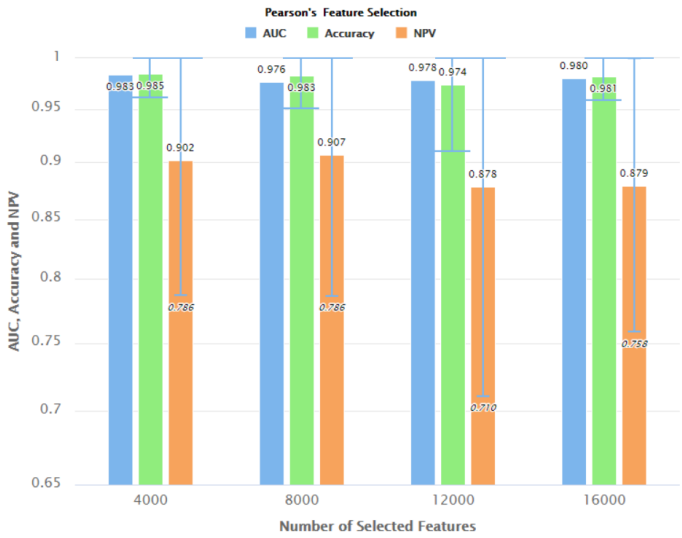
Table 5.5: Mean and standard deviations results of the DL model for cancer classifcation using the whole gene set over 100 iterations with and without data augmentation.

A general strategy as shown by state of the art is to perform feature selection to avoid overfitting and improve predictive performance [136], [2], [49]. Although dropout and early stoppage were used to avoid overfitting the training set, we employed three different feature selection techniques to evaluate the effect of dimensionality reduction on predictive performance. On Figures 5.5a, 5.5b, 5.6 we present the results for three different feature selection techniques: Welch's t-test, Pearson's correlation and Fisher's score, respectively, for the set sizes of 4000, 8000, 12000 and 16000 genes. The experiments were conducted over 100 iterations, using the model with no data augmentation that performed best.

The Welch's t-test feature selection on Figure 5.5a, provided an average AUC of 0.979±0.042, an accuracy of 0.978±0.063 and NPV 0.892±0.131 across all gene set sizes. The best gene set size was 12000 with an AUC of 0.983, which does not represent an improvement over the DL model with no feature selection. Pearson's correlation feature selection on Figure 5.5b, provided an average AUC 0.979±0.035, an accuracy of 0.981±0.036 and NPV 0.892±0.131. The best set size was the top 4000 genes with an AUC of 0.983, also not outperforming the no FS model.



(a) AUC, Accuracy and NPV results for Welch's t-test top 4000, 8000, 12000 and 16000 genes.



(b) AUC, Accuracy and NPV results for Pearson's Correlation top 4000, 8000, 12000 and 16000 genes.

Figure 5.5: Performance of the DL model, varying the FS technique and the gene set sizes.

Fisher's score on Figure 5.6, provided more stable results across all set sizes, with average AUC of 0.980±0.032, accuracy 0.980±0.043 and NPV 0.888±0.126. The best gene set size was 8000 with AUC 0.984 equaling the no FS model's performance, however presenting lower NPV of 0.905. Overall feature selection seemed not to affect the DL learner's predictive performance; Fisher's score with 8000 genes provided an equal AUC score, and the Welch's t-test with 12000 features achieved a small boost in NPV to 0.914 while showing worse AUC. This might indicate that our DL network was not complex enough to overfit the data and so could not benefit from feature selection for cancer prediction, or that the addition of dropout regularisation avoids the need for feature selection.

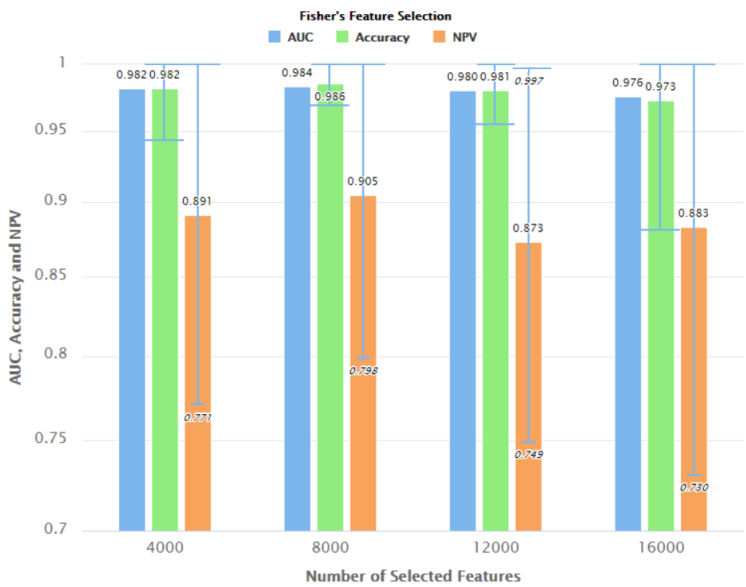


Figure 5.6: AUC, Accuracy and NPV results for Fisher's score for different gene set sizes.

5.5.2.2 LightGBM

This section discusses the gradient boosted tree learner predictive performance compared to the DFF baseline score discussed in the previous section. A similar workflow was followed then for the DFF network, the strategy with data augmentation was executed with the same number of neighbours, and when not performing DA, the hyperparameter *scale_pos_weight* was used to balance the weights of the classes. A critical difference between using LGBM and the DL models is that tree-based learners are more robust in handling correlated features. Furthermore, the specific implementation of lightGBM eradicates the need for a priori feature selection as the model removes redundant features by performing exclusive feature bundling as discussed in section 5.3. The average number of features used by the model for prediction was 19,935.3, which means that EFB discarded an average of 596 features that added no information across the 100 iterations. Table 5.6, presents the best DL model predictive performance, and the LGBM model results with and without DA. The strategy of oversampling the minority class lead to worse overall performance, similar to the DL model.

	(mean ± standard deviation)					
Metrics	***DL** (baseline)*		***LGBM***		***LGBM w/ DA***	
AUC	**0.984 ± 0.025**		0.983 ± 0.017		0.970 ± 0.013	
Accuracy	0.986 ± 0.011		**0.995 ± 0.006**		0.991 ± 0.007	
	Positive	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*
Precision	0.998 ± 0.004	0.912 ± 0.097	0.997 ± 0.005	**0.976 ± 0.036**	0.994 ± 0.007	0.969 ± 0.038
Recall	0.989 ± 0.012	0.977 ± 0.038	0.997 ± 0.004	0.969 ± 0.046	0.997 ± 0.005	0.943 ± 0.061
F1-score	0.994 ± 0.006	0.940 ± 0.047	0.997 ± 0.003	0.972 ± 0.030	0.995 ± 0.004	0.955 ± 0.031

Table 5.6: Performance of the LGBM and DL model for the cancer prediction problem across 100 iterations.

Compared with the baseline DFF, lightGBM outperformed in accuracy and false precision (NPV), and the DFF showed better AUC. However, due to the problem's imbalanced nature, it is arguable that LGBM is the best classifier as it shows higher NPV which translates to a more balanced model. The DL presented a higher AUC of 0.984; however, it is highly biased to the positive majority class with an NPV of 0.912. The average number of independent test samples' for support was 154.55 for the positive class and 17.67 for the negative, which makes NPV very relevant because a learner biased to the positive class would also present excellent AUC and accuracy. Overall it is arguable whether the DL or the LGBM model had best predictive performance; however, the LGBM model was able to reduce variance and showed to be a more balanced learner, which is important in the context of the problem.

5.5.3 Subtype Classification

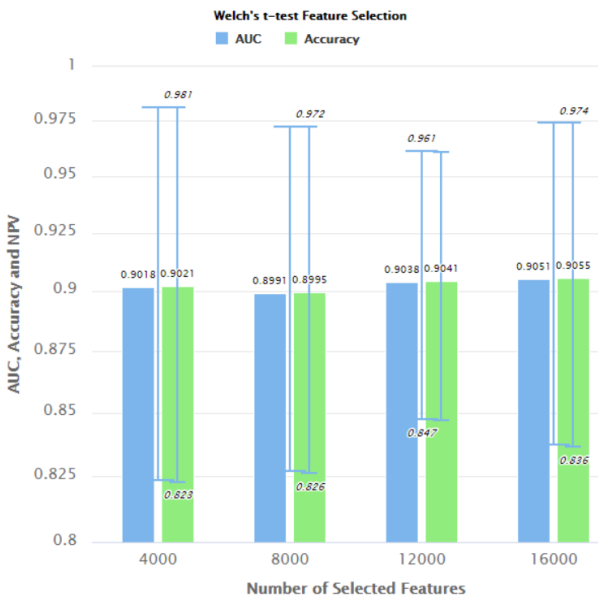
5.5.3.1 Deep Learning

The cancer subtype classification problem presents a balanced dataset as opposed to cancer prediction. The average test support for the positive class was 52 and 50 for the negative, resulting in almost 1:1 ratio. Table 5.7 presents the DL model results without feature selection over 100 iterations. The DFF network showed an AUC and accuracy of 0.907 with a very high standard deviation of 6.8%. It showed a higher precision of 0.948 for the negative class, which are LUSC samples, but a bad trade-off with recall of 0.87, showing that its predictions' are skewed to the positive class. Figures 5.7a, 5.8, 5.7b present the results after feature selection following the same methodology than in the cancer classification problem, however this time we analyse only AUC and accuracy due to the balanced nature of the problem.

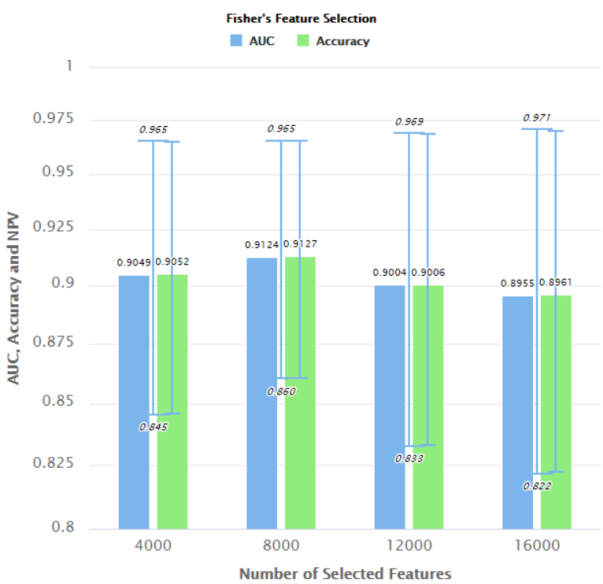
(mean ± s.d)	*AUC*	*Accuracy*		*Precision*	*Recall*	*F1-score*
DL	0.907 ± 0.068	0.907 ± 0.068	*Positive*	0.896 ± 0.089	0.942 ± 0.096	0.911 ± 0.068
			Negative	0.948 ± 0.068	0.871 ± 0.139	0.897 ± 0.092

Table 5.7: Performance of the DL model for cancer subtype classification using the 20,531 genes over 100 iterations.

Welch's FS on Figure 5.7a presents very balanced results overall, with an average AUC of 0.902±0.069 and accuracy of 0.903±0.069. The best set size by a small margin was the top 16000 genes presenting an AUC and accuracy of 0.905±0.069, and outperforming the DL model with no FS. However, the set with 12000 genes presented very similar results with AUC and accuracy of 0.904 but showed less variance with a standard deviation of 0.057. For Fisher's score selection in Figure 5.7b, the results were less balanced across all sizes with an average AUC of 0.903±0.064 and accuracy 0.904±0.063. The best set size corresponds to the top 8000 genes showing AUC



(a) AUC and Accuracy results for Welch's t-test top 4000, 8000, 12000 and 16000 genes.



(b) AUC, Accuracy and NPV results for Fisher's score top 4000, 8000, 12000 and 16000 genes.

of 0.912 and accuracy of 0.913 but presenting a significantly lower standard deviation of 0.052 compared to the remaining sets.

The results for Pearson's correlation FS in Figure 5.8, show a significant improvement in predictive performance for smaller sets. The average overall AUC and accuracy for the method were 0.916±0.052, and the 4000 and 8000 top genes significantly improved the predictive performance compared with the DL model without feature selection. The top 4000 genes showed the best performance with AUC of 0.935±0.033 and accuracy of 0.935±0.032.

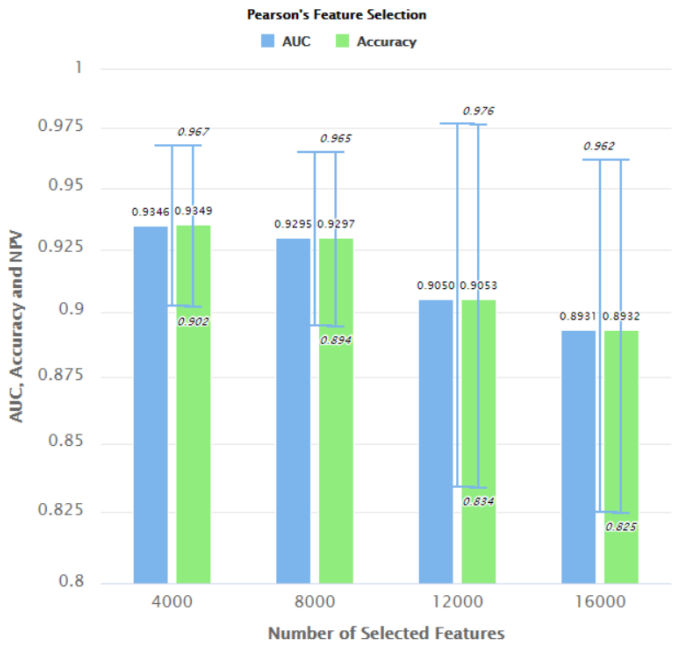


Figure 5.8: AUC and Accuracy for Pearson's Correlation top 4000, 8000, 12000 and 16000 genes.

Overall feature selection for the subtype classification problem significantly improved the predictive performance. Pearson's correlation method performed better for the top 4000 genes and reduced standard deviations considerably for the two smaller sets when compared to the DL learner with no feature selection. This result is contrary to what was verified on cancer prediction, which could result from the gene expression dataset containing more noise for the subtype classification problem than for cancer prediction.

5.5.3.2 LightGBM

Table 5.8 presents a comparison of the LGBM learner results for the subtype classification problem with the baseline DL model, using the whole gene set, and the best iteration of feature selection corresponding to Pearson's correlation top 4000 genes. Across 100 iterations an average of 19710.3 features were used by the model and approximately 821 were removed through exclusive feature bundling.

The LGBM outperformed all DL models with AUC and accuracy of 0.971 and lower standard deviations incurring in less variance. The most significant difference in the indicators presented seems to be accuracy in classifying the positive class (LUAD samples), showing precision of 0.969 against 0.919 from the best DL model. Precision was also better for the negative class, and the precision-recall trade-off remains very balanced for both classes, showing a more fit and less biased classifier than the DL learners.

	(mean ± standard deviation)					
Metrics	***DL** (baseline)*		***DL** (Pearson's top 4000)*		***LGBM***	
AUC	0.907 ± 0.068		0.935 ± 0.033		**0.971 ± 0.018**	
Accuracy	0.907 ± 0.068		0.935 ± 0.032		**0.971 ± 0.018**	
	Positive	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*
Precision	0.896 ± 0.089	0.948 ± 0.068	0.919 ± 0.052	0.960 ± 0.034	0.962 ± 0.020	0.980 ± 0.022
Recall	0.942 ± 0.096	0.871 ± 0.139	0.960 ± 0.038	0.909 ± 0.073	0.980 ± 0.022	0.961 ± 0.023
F1-score	0.911 ± 0.068	0.897 ± 0.092	0.938 ± 0.028	0.931 ± 0.040	0.971 ± 0.014	0.970 ± 0.015

Table 5.8: Performance of the LGBM model for cancer subtype classification problems over 100 iterations.

However, it is odd that although the LGBM learner for the subtype problem is more complex than the cancer prediction one, presenting larger tree depths, the number of features removed through EFB was bigger than for the cancer classification problem. The inclusion of L1 and L2 regularisation assists to force sparsity and diminishing weights, and higher values of *min_split_gain* avoid splitting with minimal gains, which contributes to avoiding overfitting the training data and are partial reasons why LGBM achieves such stable results. Larger values of *feature_fraction*, which randomly selects a subset of features on each iteration, and *min_data_in_leaf* which controls the minimum number of observations that must fall into a tree node for it to be added, control the power of generalisation of the model.

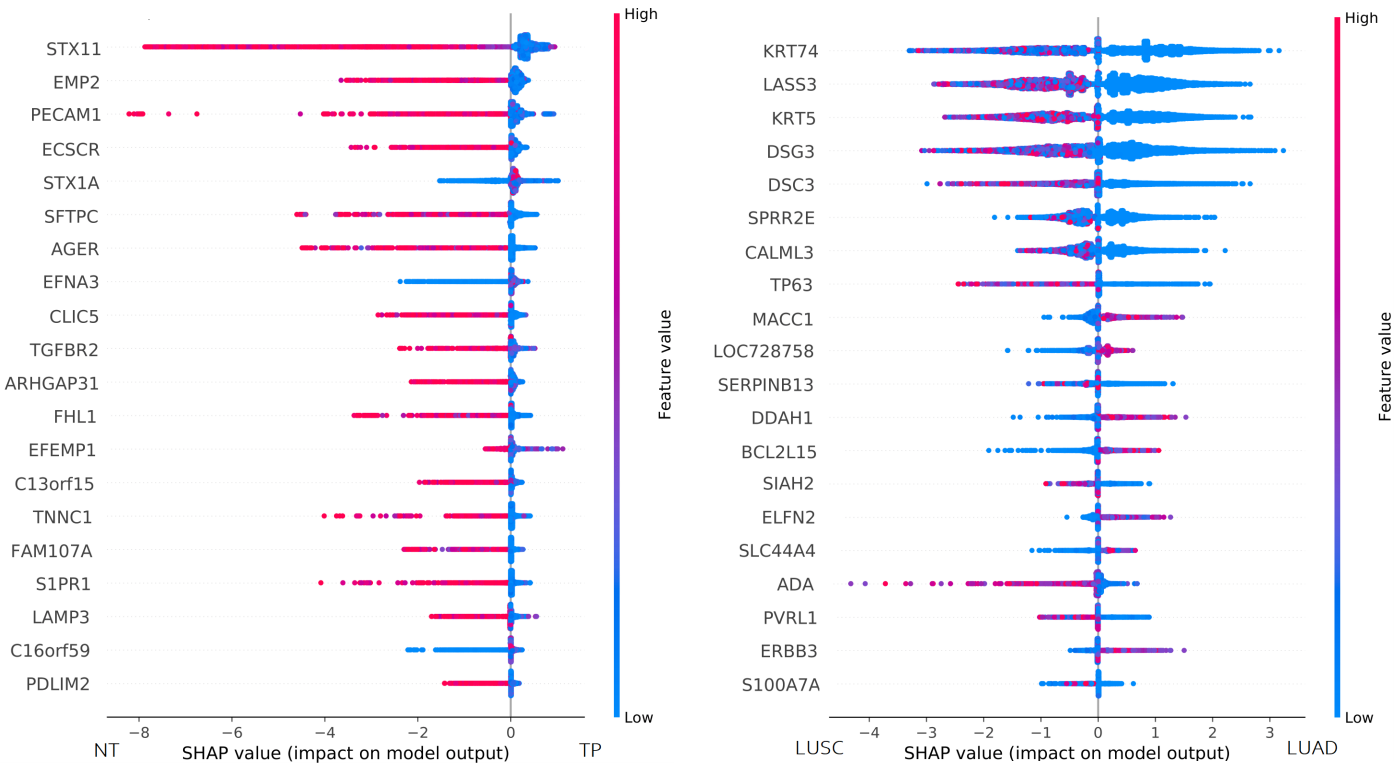
On the other hand, as seen by the DL model results, shorter gene set sizes resulted in better predictive performance, which might indicate the presence of more noise in the data for the task of subtype classification. Overall, these results seem promising and show GBDT models' capability to handle high dimensional data, while eradicating the need for apriori feature selection.

5.5.4 Most Relevant Gene Signatures

As explained in section 5.3 we used the SHAP technique to provide interpretability for the lightGBM model. This section analyses the most important features selected by the lightGBM and provides biological insight into the gene signatures that are affecting the decisions in each problem.

5.5.4.1 Feature Importance

Figures 5.9a, 5.9b present the SHAP summary plot for the cancer prediction and subtype classification problems respectively, which combines feature importance with feature effects. The plot's y-axis identifies a gene, represented by its HUGO symbol and the x-axis the corresponding SHAP values for each data instance. The genes are ordered on the y-axis by overall predictive importance, and the top 20 most important genes were selected for analysis. The colour gives us a visual representation of features' original expression value distributions, which are categorised into low or high values of gene expression. This visualisation can give us a holistic view of the model's decision as it conjugates the importance of the features with the effect on prediction while showing the value distribution of those features in the original data.



(a) Cancer Prediction (b) Subtype Classification

Figure 5.9: SHAP values over 100 iterations for the LGBM model. The y-axis identifies genes and the x-axis the corresponding SHAP values for each data instance.

For the cancer classification problem, a total of 1,183 genes were selected by the model for prediction. Figure 5.9a shows the genes that more adequately distinguish between cancerous and non-cancerous samples according to the model. The negative weights represent features with a negative effect on prediction, which equates to features whose effect helps to predict NT samples, and the positive weights bind the decision to predict cancerous tissue. Amongst these top 20 genes, we can clearly see a pattern used for prediction: most of the selected genes when over-expressed affect the prediction negatively; and when under-expressed affect the prediction positively. Exceptions to this are STX1A, EFNA3 and C16orf59 genes, which present mostly low expression in both classes and some visible over-expression, which positively impacts the prediction. Generally, the analysis of the 20 most important gene expression signatures for cancer prediction shows a pattern of selecting signatures with a high expression that are important to predict normal tissue.

For the cancer subtype classification problem, 2,685 genes presented non-null weights and therefore were used for prediction. Figure 5.9b shows that the expression value of genes is more balanced across features with a positive and negative effect on model output. Negative weights bias decision to predicting LUSC samples and positive weights bias the decision to predict LUAD samples. We can infer two groups of genes that show identical patterns: MACC1, LOC728759, DDAH1, BCL2L15, ELFN2, SLC44A4 and ERBB3 represent the first group that shows mostly over-expression when binding the decision to predict LUAD, and under-expression when negatively affecting the prediction; the second group containing the remaining 13 genes present a symmetrical pattern with mostly over-expression when important for the model to predict LUSC tissue. Overall, the feature importance analysis found two different sets of gene signatures that can be considered specific subtype gene signatures, which can help differentiate LUAD and LUSC subtypes, and a set of gene signatures that distinguishes cancerous from normal tissue at the molecular level.

Although this type of analysis is very informative from the point of view of interpreting the model decisions, it presents a limitation. It does not reflect the selected genes' real expression distributions, but the instances of those genes that were more relevant for prediction across the 100 iterations. On section 5.5.4.2 we provide a more in-depth analysis of the overall expression distribution for these genes that can provide us with more biological insight on the real expression of the genes across the populations.

5.5.4.2 Differential Expressed Genes

On Figures 5.10, 5.11 we present a heatmap showing the expression of the top 20 genes for the cancer prediction and subtype classification problem respectively. This type of analyses allows us to identify statistically significant gene expression changes across the entire samples.

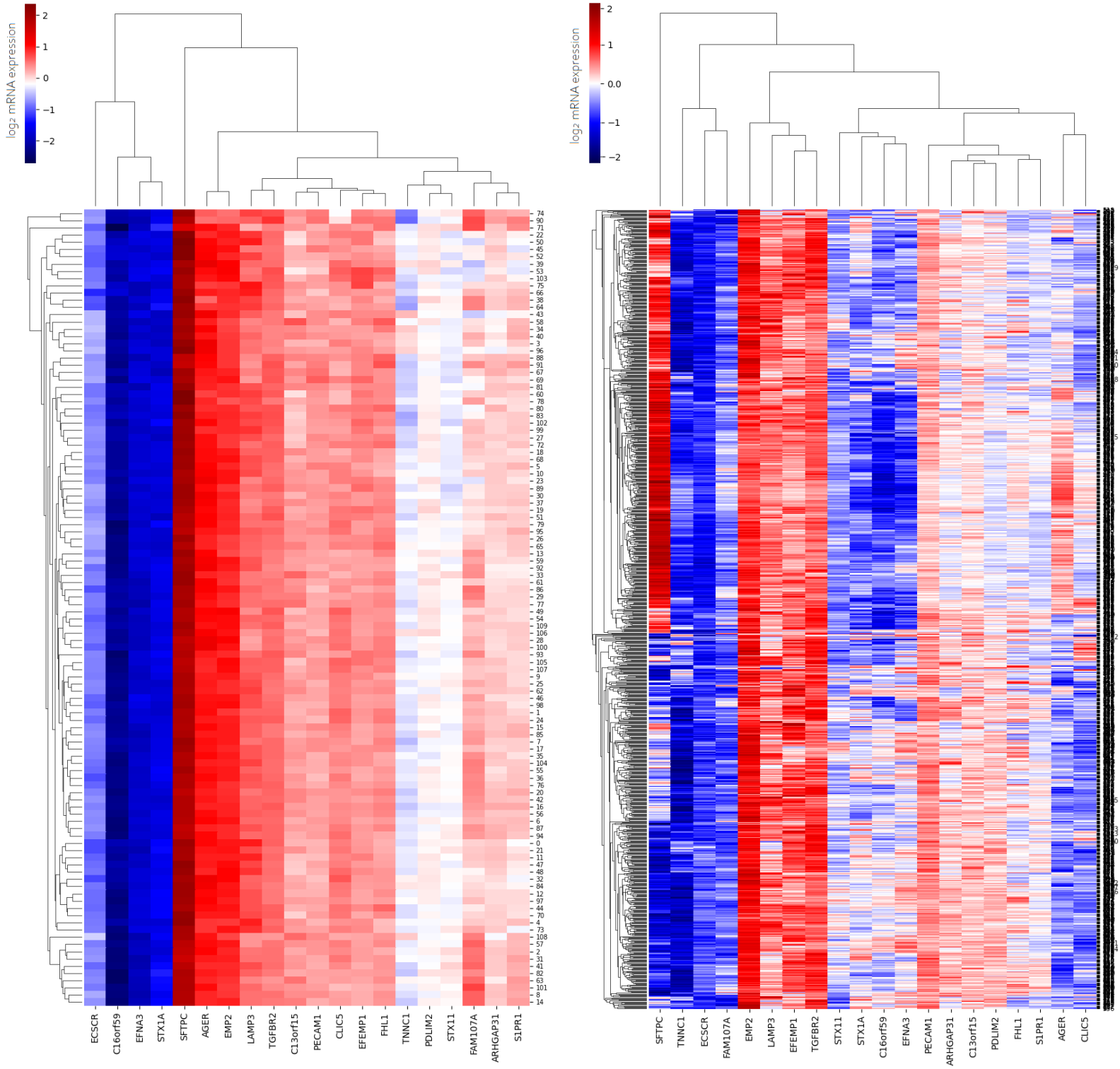
We used log_2 normalised counts of the genes' expression for differences in sequencing depth and composition bias between the samples to generate the heatmaps. They are organised as follows: the data is displayed in a grid where each row represents a sample, and each column represents a gene signature. Each individual square has a colour and intensity used to represent gene expression changes across samples, being the colour red over-expression and the colour blue under-expression, while white represents no visible changes in expression. Furthermore, there are two hierarchical clusters row-wise and column-wise that serve to group similar samples or genes together, respectively. This type of clustering mechanism based on the similarity of gene expression patterns can be useful for identifying genes that are commonly regulated, or biological signatures associated with a particular condition (e.g. a disease or an environmental condition) [53].

Cancer Prediction

Regarding the most important genes for cancer prediction, on Figure 5.10a it is presented the analysis of a total of 110 normal tissue samples and on Figure 5.10b 1,016 cancerous samples. On Figure 5.10b, it is possible to observe two well defined row-wise clusters, which refer to the LUAD samples on top and the LUSC samples on the bottom-cluster. The normal tissue samples present a more homogeneous structure with no visible row-wise clusters, which is expected as NT samples are collected from regions outside of the tumour and therefore should present similar gene expression regardless of being a LUAD or LUSC normal tissue sample.

The columns are sorted by cluster proximity, and on the NT samples it is possible to observe 4 distinct macro clusters from left to right: ECSCR to STX1A comprising mainly under-expression genes; SFTPC up to TGFBR2 containing the highest expression values; C13orf15 up to FHL1 containing more moderate values of over-expression, and TNNC1 till S1PR1 containing less visible changes in gene expression. These clusters are further divided into sub-clusters according to their inner cluster proximity: for example for the under-expression cluster, STX1A and EFNA3 are the bottom-most cluster and therefore are the closest in terms of expression levels, following C16orf59 is the closest presenting the most visible under-expression and finally, ECSCR is the furthest away presenting very moderate under-expression. Regarding the two sets of genes that presented similar patterns in the feature importance analysis on Figure 5.9a: STX1A, EFNA3 and C16orf59 were considered important for NT prediction when under-expressed, and this type of analyses validates that presumption by showing that they do display the highest relative values of under-expression of the 20 genes considered for analysis for NT.

The cancerous tissue samples provide a more heterogeneous composition with five macro clusters and less defined subclusters. Furthermore, the distinction between specific subtype genes' is evident, e.g. SFTPC is more expressed in LUAD samples, and the STX1A, EFNA3 and C16orf59 are on the middle cluster that is more under-expressed in LUAD samples.



(a) Normal Tissue (NT)

(b) Cancerous Tissue (TP)

Figure 5.10: Analysis of log_2 mRNA expression values for the top 20 most important genes for the LGBM cancer classifier. The rows represent samples and the columns genes, and each individual square represents changes in gene expresion from under-expression (blue) to over (red).

Subtype Prediction

For the subtype prediction problem, on Figure 5.11a there are a total of 515 LUAD samples represented, and on Figure 5.11b 501 LUSC samples. Both subtypes present three well-defined parent clusters, ordered from left to right by under-expressed genes first, then the ones over-expressed, and finally genes that present no visible over or under expression.

The genes in the central cluster that are more over-expressed in LUSC: KRT5, CALML3, PVRL1, DSG3, DSC3 and TP63 represent genes that when over-expressed supported the decision to predict LUSC in the feature importance analysis on Figure 5.9b. They show up in the under/average expressed clusters in LUAD except for the PVRL1 gene that also presents over-expression in LUAD.

For LUAD, the most relatively expressed genes are MACC1, PVRL1, SIAH2, SLC44A4, DDAH1 and ERBB3. All except PVRL1 and SIAH2 show up as genes that support the decision to predict LUAD when over-expressed on the feature importance analysis. PVRL1 and SIAH2 are genes that present high relative expression in-between subtypes, and by repeating the same analysis on the LUAD and LUSC populations together, we observed that the relative values of these genes were higher amongst LUSC samples. This type of analysis presents validation on the model's selected features as we can see the same type of mutual exclusion between the high and low-expressed features in the histologic subtypes.

Literature cross-reference

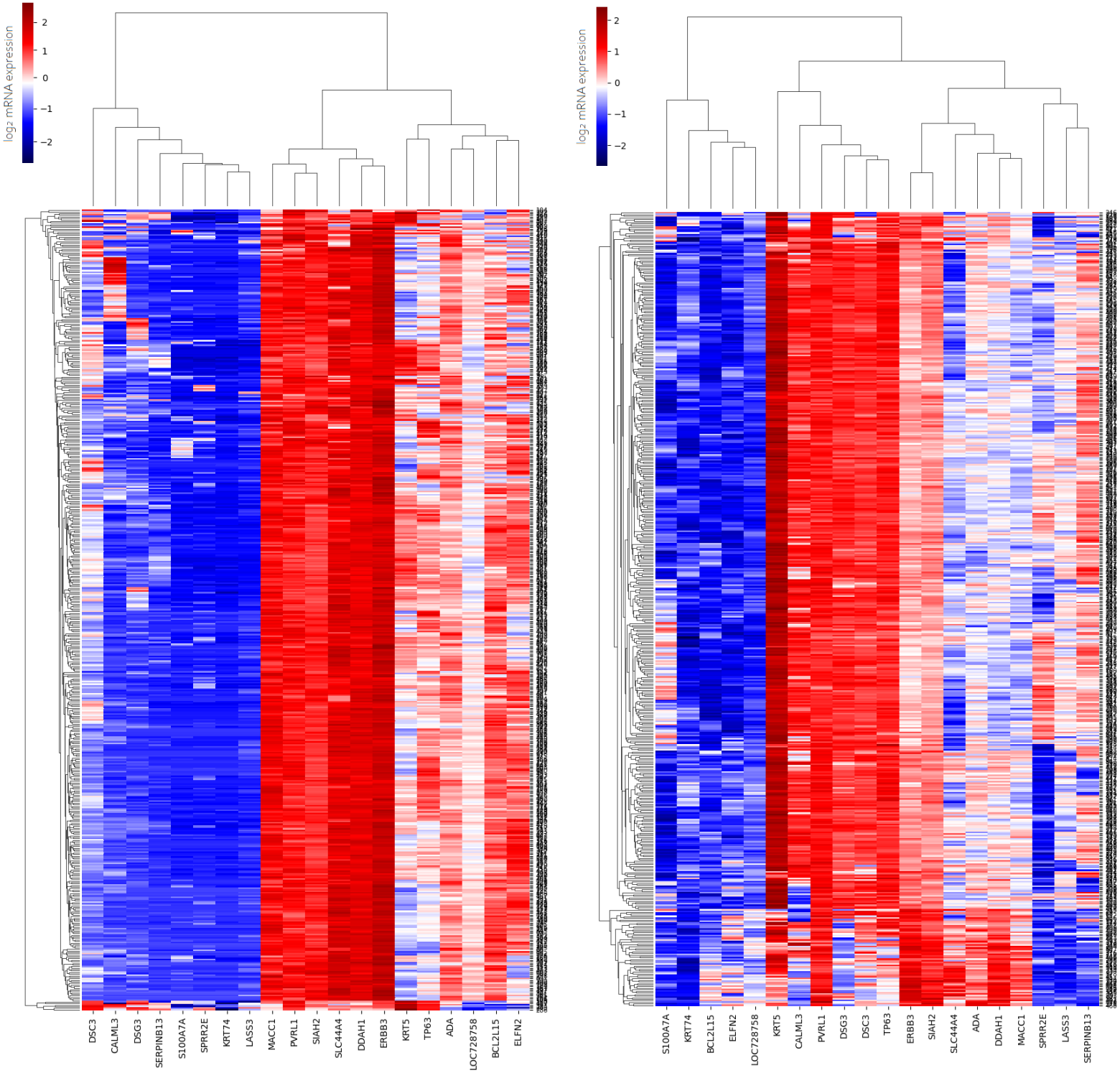
We reviewed the cancer research corpus for mentions of the selected genes to further validate the results obtained. We highlight some gene signatures selected by the models that are already extensively documented in literature:

Cancer Prediction

- SFTPC encodes the pulmonary-associated Surfactant Protein C (SPC) and its deficiency is associated with interstitial lung disease [45]. SFTPC downregulation might be involved in the progression of lung cancer [83].
- C16orf59 or TEDC2 is prognostic, and high expression is unfavourable in lung cancer [47].
- PDLIM2 which is a putative tumour suppressor protein, and decreased expression of this gene is associated with several malignancies including breast cancer and adult T-cell leukaemia [45]

Subtype Classification

- Alterations in genes with known roles in LUSC were found, including over-expression and amplification of TP63 [39]
- Elevated MACC1 expression has been implicated in the progression of LUAD [86]
- ERBB3 or HER3 is an important paralog of gene ERBB2, whose aberrations have been found to drive LUAD tumours [35]



(a) LUAD Tissue

(b) LUSC Tissue

Figure 5.11: Analysis of log_2 mRNA expression values for the top 20 most important genes for the LGBM subtype classifier. The rows represent samples and the columns genes, and each individual square represents changes in gene expresion from under-expression (blue) to over (red).

5.6 Discussion

This chapter aimed at finding the most suitable solutions for cancer prediction and subtype classification using gene expression data. Furthermore, the importance of providing interpretability on the results was a key consideration taken into account as the identification of gene expression signatures that better differentiate histologic subtypes can further impact the cancer research and potentially help to discover new possible targetable genes for personalised therapy. In this section, we compare our results with state of the art and summarise some pros and cons of our approach against similar research.

5.6.1 Cancer Prediction

On Table 5.9, we present a comparison of the predictive performance of our main method LGBM with previous works which try to predict lung cancer. The selection of comparable research was made according to the following criteria:

1. The focus was given to research published after 2015, and using gene expression data only.

2. The selected works must use data from the same source TCGA[1], which eliminates bias from different methods and sequencing platforms.

3. The works must have at least one individual metric score for classification of cancer versus non-cancerous tissue using TCGA-LUAD or TCGA-LUSC; ideally, they use both.

In Li et al. [84], a total of 10,663 TCGA samples were used out of which 856 (8%) were normal tissue, including 515 LUAD and 501 LUSC samples. Two strategies were used for dimensionality reduction: feature extraction via an autoencoder and using prior knowledge of Gene Ontology (GO) enrichment analysis, following multiple DNNs were used for classification and a 5 layer DNN performed better with features extracted from an autoencoder. Our LGBM classifier outperforms the DNN with an f1 score of 0.997 against 0.900. Furthermore, a considerable drop in train performance from 0.940 to 0.900 in test might indicate that either the autoencoder or the DNN was overfitting the training data.

In Xiao et al. (2018) [136], 162 TCGA LUAD samples were used, out of which 37 (23%) corresponded to NT. The authors use a multi-model approach by employing a deep neural network to ensemble multiple learners' outputs, after feature selection by applying DESeq [99]. The best individual learner was DTs with an accuracy of 0.968 ± 0.023. Our LGBM approach outperformed the ensemble results in every aspect, and no apriori feature selection was required. However, the ensemble strategy provides the unique advantage of minimising the overall variance by combining multiple learners' decisions, and so the ensemble method shows lower standard deviations.

[1] TCGA data portal (https://portal.gdc.cancer.gov/)

	Performance Metrics (mean ± standard deviation)			
Methods	***Accuracy***	***Precision***	***Recall***	***F_1 score***
Li et al. (2017)[84]		0.900	0.900	0.900
Xiao et al. (2018)[136] *	0.988 ± 0.003	0.985 ± 0.002	0.974 ± 0.03	0.979 ± 0.003
Xiao et al. (2019)[137] *	0.999 ± 0.002	0.998 ± 0.002	0.999 ± 0.002	**0.998 ± 0.004**
Ahn et al. (2019)[2]	0.979			
LGBM w/ BO	0.995 ± 0.006	0.997 ± 0.005	0.997 ± 0.004	0.997 ± 0.003

Table 5.9: Comparison of LGBM predictive performance with previous works for cancer prediction using TCGA LUAD and LUSC RNA-seq datasets. * Results only for TCGA-LUAD.

In Xiao et al. (2019) [137], the same 162 TCGA-LUAD samples were utilised. DESeq2 [92] was used for feature selection, and 1,385 genes out of 20,531 were selected for the modelling stage. A semi-supervised learning approach was used by extracting relevant features and using a stacked sparse autoencoder (SSA) for classification. The SSA classifier provided an accuracy 0.9987±0.018, which outperforms our LGBM in terms of accuracy; nevertheless, accuracy is not an optimal indicator for unbalanced problems and AUC score is not provided. However, the difference in F_1-score is minimal, and our approach showed less variance and more balanced precision-recall trade-off. Furthermore, considering that these and the previous research only consider a total of 162 LUAD samples out of the 598 available on TCGA, and the inclusion of LUSC samples strenuous the task by increasing the heterogeneity of gene expression profiles across cancerous and non-cancerous tissue, it is arguable that our method is as good if not outperforming in a similar context. However, we decided to include these papers as they present up to par validation with our work, and provide excellent predictive performance even with the considered limitations. A key difference between our strategy and SSA is the need for two pre-processing steps; first DESeq2 was used to select differential expressed genes, then SSA was used to extract relevant gene signatures out of the 1,385 pool for classification. Our results also show the DL model necessity for prior feature selection, mostly on the subtype problem, while the LGBM model handles feature selection internally, which presents as an advantage.

In Ahn et al. [2], a total of 8,839 TCGA samples were used out of which 645 (7.3%) were NT samples, retrieved from 24 different cancer types, including all TCGA LUAD and LUSC samples. Five gene sets were created with the top 3,000 genes using various mean and rank-variance feature selection techniques, following a DNN was employed for classification and compared to 5 baseline models. Our LGBM approach outperforms the described method, although the comparative research presents the unique advantage of including samples from different cancer types. It highlights an interesting opportunity for future work to extend our learner to other cancer types to assess how well it generalises.

5.6.2 Subtype Classification

Table 5.10 presents a comparative study of the predictive performance of our best performing method LGBM for the task of lung cancer subtype classification. A similar selected papers' criterion was used then for cancer classification, except point 3) where papers must include both LUAD and LUSC datasets for subtyping. Accuracy is used as the primary indicator of predictive performance as subtype classification presents a problem of balanced nature.

In González et al. [109], a total of 230 TCGA samples were used, with a 2:1 ratio imbalance of LUAD samples. Preliminary feature selection was done using the Mann-Whitney test and selecting genes gene which present statistical different distributions ($p_{value} < 0.05$), following the second stage of feature selection was performed using pGBRT to minimise the difference between genes, evaluated using MSE. Finally, wkNN, kNN, SVM and NB were used to distinguish LUAD from LUSC samples using various gene set sizes. Our approach outperformed in every aspect; however, the LGBM learner had a less balanced precision-recall trade-off which translates to less precision in predicting LUAD samples. Nonetheless, the work analysed presents the possibility of using more complex models for feature selection which could present an opportunity for future work of using LGBM as a feature selection method and on top an unsupervised or supervised strategy to separate the histologic subtypes.

	Performance Metrics (mean ± standard deviation)			
Methods	***Accuracy***	***Precision***	***Recall***	***F_1 score***
González et al. (2017)[109]	0.965	0.924	0.983	0.953
De Guia et al. (2019)[29]	0.955	0.957	0.954	0.955
Smolander et al. (2019)[118]	0.960 ± 0.012	0.936 ± 0.012	0.984 ± 0.010	0.959 ± 0.011
Ye et al. (2020)[138]	0.921	0.946	0.901	0.923
LGBM w/ BO	**0.971 ± 0.018**	0.962 ± 0.020	0.980 ± 0.022	**0.971 ± 0.014**

Table 5.10: Comparison of LGBM predictive performance with previous works for cancer subtype classification for TCGA LUAD and LUSC subtypes.

In De Guia et al. [29], 515 LUAD and 501 LUSC samples were used. A different strategy of embedding the gene expression values into a 2D image, and using a CNN to find differential expressed genes. Baseline methods SVM and RF were used to assess the CNN performance, and SVM performed better with 0.931 F_1-score. The CNN showed better precision in predicting LUAD samples, which is not the case for our LGBM model, which although it outperforms the CNN, it shows better accuracy in the classification of the LUSC class.

Smolander et al. [118], a total of 976 samples was used with 1:1 ratio. The authors evaluate the effect of coding and non-coding proteins in the classification of NSCLC by employing a deep belief network for classification while comparing it to baselines SVM and RF. The impact of feature selection was studied DBN outperformed the SVM with 500 features and FS that showed an accuracy of 0.939 ± 0.01. The DBN was also used as a feature selection method prior to SVM classification, which resulted in a worse overall accuracy of using only the DBN. According to our revision of state of the art, Smolander et al. [118] provided the best results for subtype classification of lung cancer using gene expression TCGA data with an accuracy of 0.960. Our approach outperforms the related work, which also presented worse precision for the positive class, but nevertheless showing very low variance overall. The work conclusions follow the same line with the results drawn from our work, which showed that feature selection was very important for deep learning models mainly in the subtype classification problem, in which the use of simple variance-based techniques provided significant predictive performance gains and reduced overall variance.

In Ye et al. [138], a total of 1,013 samples were enrolled, out of which 512 corresponding to LUAD samples. The authors followed a different strategy to minimise the number of selected genes to find relevant signatures. Preliminary feature selection was performed using Edger [112] which selected 5,469 differentially expressed genes. Following hierarchical clustering was used to separate the genes in sub-groups further, each sub-group was evaluated individually, and the top 20 genes from the best performing group were selected. SVM, kNN and DT were used to predict solely using combinations of the top 20 genes, and kNN using top 17 genes yielded the best accuracy of 0.922. The precision was also better for the LUSC class similar to our classifier but considering that only 17 genes used for prediction the related work presents excellent results.

5.7 Summary

This chapter presented our solutions for cancer and subtype classification using gene expression data, which we tackled using two approaches: a DL model that is the standard strategy in state of the art, and a tree-based learning model that we provided interpretability to retrieve biological insight. Cancer prediction using gene expression data poses a different problem then subtype classification, as the composition of medical datasets generally has a positive majority class, while the subtype problem presented a balanced class problem.

To solve the class imbalance problem in the cancer classification problem, we explored data augmentation using ADASYN to oversample the minority negative class. Although we found one reference to oversampling using SMOTE in the covered literature [28], the research shows no comparative results for DA. Our research showed no performance gains for both the DL and the LGBM method when performing oversampling, and a strategy of adjusting the weights of each learner as an optimisation step yielded better results. Hyperparameters optimisation ran for both problems, using grid search for DL models and bayesian optimisation for the LGBM model to determine the best architectures and find combinations of hyperparameters that minimise overfitting and improve generalisation.

A systematic review of the literature highlighted the importance of feature selection to reduce the high dimensionality of gene expression profiles, which is particularly relevant when using deep learning methods as they are prone to perform poorly due to the curse of dimensionality. We designed an experimental study to assess the impact of feature selection on DL models by using different FS techniques with several gene set sizes. Our study's empirical results showed a clear performance gain in the subtype classification problem when performing FS but not a statistically significant result in the cancer prediction one. Our hypothesis for this is the following:

1. The cancer classification DNN had a 3 hidden layer architecture versus 4 layers from the subtype prediction, and the lower complexity of the network might not get the full benefits from feature selection because it was not complex enough to overfit the data.

2. The task of predicting subtypes has the same dataset characteristics (number of genes) but represents a subset of the cancer prediction problem (only TP samples corresponding to LUAD and LUSC). The objective is also very different, and while the heterogeneity between TP and NT samples is huge, the heterogeneity for LUAD and LUSC samples is not that much; therefore there is more noise in the data when the task is cancer subtyping as the difference is focused in a minimal set of genes.

The GBDT method proved to excel at handling high dimensional data while eradicating the need for feature selection. A comparative study on the effectiveness of the LGBM versus the XGboost implementation was performed on breast cancer mRNA data in Wang et al. [131], which showed that lightGBM algorithm performed better overall. Although the authors do not conclude why we deduce that the unique advantage of using the GOSS algorithm over the Pre-sorted/Histogram-based algorithm gives the capability of LGBM to handle large, sparse datasets better as instances

with small gradients vanish. Furthermore, EFB provides a means of the model of removing redundant features by eliminating mutually exclusive features, which eradicates the need for feature selection. Preliminary results show that our methods outperform previous work's results using DL methods for subtype classification, and show the potential for usage as main predictive models or tools for identifying differentially expressed genes.

To validate our results, we used the Shapley technique to provide interpretability on the LGBM model. We analysed the top 20 most relevant gene expression signatures for the cancer and subtype prediction problem. First, we analysed the importance of the features from the model point of view, using a feature importance visualisation that correlates feature importance with feature effect that gives us an idea of what features are more critical for the model when deciding. Furthermore, we analysed the same set of 20 genes independently from the model to have a more in-depth analysis of the overall expression distribution changes for the genes in the target populations of each problem. We performed hierarchical clustering of the genes to select sub-groups of genes whose distribution was most similar in each population. This type of clustering mechanism allows for identifying common-regulated genes, which can serve as a mechanism to identify groups of gene signatures associated with specific diseases. As a result of this work, two sets of gene signatures were extracted: a set that differentiates normal tissue from cancerous tissue (cancer prediction), and a set of that distinguishes LUAD from LUSC samples (subtype classification). Furthermore, we cross-referenced some staple works such as those of The Cancer Genome Atlas Network [35], [39] and found the same patters of over and under-regulation for genes specific to each histologic subtype, e.g. over-regulation of TP63 in LUSC, which provides additional validation of our results and presents the prospects to have found new, un-covered gene signatures in literature.

Chapter 6

Survival Analysis and Outcome Prediction

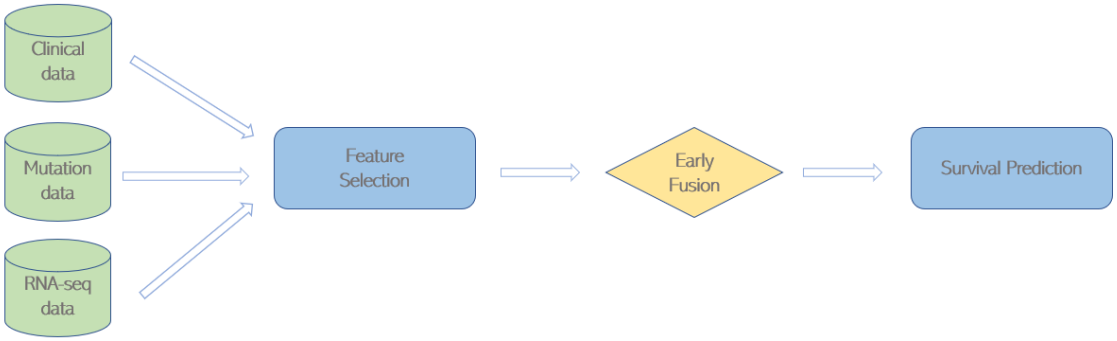
In this chapter, we describe our solution for the survival outcome prediction problem. The TNM staging system originated from the need for an accurate, consistent, universal cancer outcome prediction system. Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified, and new methods for integrating prognostic factors have been developed [13]. The work in this chapter aims to integrate multiple genomic and clinical data to capture the complex associations needed for accurate outcome prediction.

On section 6.1, we give an overview of the experimental workflow and describe the data preparation stage. On section 6.2, we present the materials used and explain the different strategies to handle multimodal data properly. On section 6.3, we briefly cover the evaluation metrics used to assess the models' predictive performance. On section 6.4, we present the predictive performance results and the post-analysis. Finally, on section 6.5, we discuss our results comparing with state of the art and validate our research methodology.

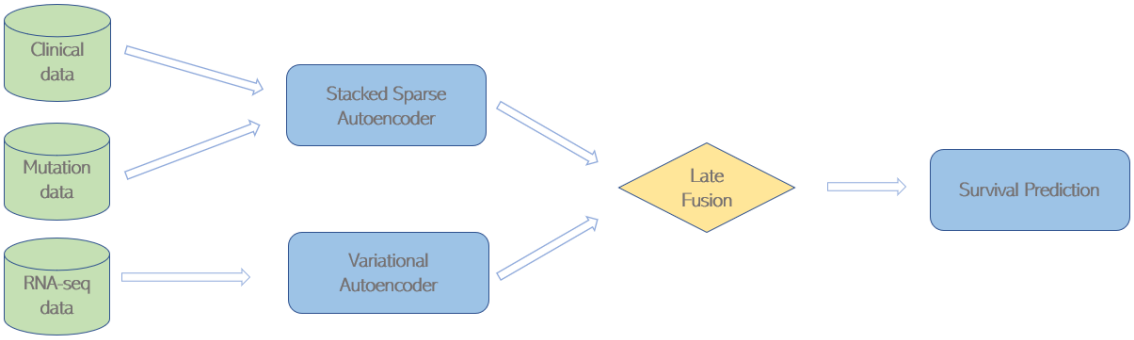
6.1 Experimental Design

Figure 6.1 presents an overview of the experimental design employed to predict cancer patients' outcome. We integrate three modalities: gene expression, somatic mutation data and clinical variables in order to maximize the information gain needed for accurate survival prediction.

To effectively handle the multimodal data, we use early and late fusion techniques to combine the heterogeneous data types. Figure 6.1a presents the early fusion experiment design, where the unprocessed modalities go through feature selection to select for relevant information and then are fed into a DL network for survival prediction. Figure 6.1b presents the late fusion design, where each modality is first processed in individual learners to extract relevant features, which then merged into a DL network. LUAD and LUSC represent heterogeneous molecular disease even within the NSCLC subtypes and are known to present different overall survival [41]; therefore, the described methodology was followed for the LUAD and LUSC cohorts separately.



(a) In early fusion the features of each modality are merged into a DL network after feature selection.



(b) In late fusion, each modality is handled separately in an unsupervised learning stage to extract representative features, that are merged into a DL network.

Figure 6.1: Methodology of early and late fusion techniques for survival outcome prediction of the LUAD and LUSC cohorts.

6.1.1 Data Preparation

On this section, we explain the data preprocessing steps individually of each data type. The data acquisition and preprocessing steps for RNA-seq have already been covered in section 5.1.1 and are similar.

6.1.1.1 Mutation Data

Using R BioConductor framework with the TCGABiolinks package [34], [117], [43] we queried for somatic mutation from the TCGA project. The somatic data is aligned according to the human reference genome (GRCh38) using the MuTect2 [23] somatic mutation calling pipeline. We added the clinical records for the patients using TCGABiolinks' GDCPrepare, and a total of 208,181 mutation calls were retrieved for LUAD and 181,117 for LUSC.

We consider a mutation to be identified by a combination of its Entrez Gene and by the Human Genome Variation Society (HGVS) protein sequence name. There are multiple "duplicate" mutation entries for the same patients, we grouped these mutations by patients and filtered for clinically relevant cancer biomarkers as targets for therapy, as shown in Villalobos et Al. [48]. In Table 6.1, we show the mutation count, as well as their frequencies, for the considered target genes in the

LUAD and LUSC datasets, and the mutation data analysis shows that the TCGA LUAD and LUSC populations show mutational frequencies that are mostly within the ranges presented in [48].

	LUAD			LUSC		
Gene	***Counts***	***Frequency (%)***	***[48] Frequency (%)***	***Counts***	***Frequency (%)***	***[48] Frequency (%)***
ALK	38	6.94	4-7	20	4.07	None
BRAF	41	7.23	1-3	15	3.05	0.3
DDR2	24	4.23	0.5	17	3.46	3-4
EGFR	71	12.52	10	14	2.85	3
ERBB2	21	3.70	1.6-4	14	2.85	None
FGFR1	15	2.65	3	11	2.24	5
FGFR2	13	2.29	3	16	3.25	5
FGFR3	4	0.71	3	13	2.64	5
FGFR4	21	3.70	3	8	1.63	5
KRAS	150	26.46	25-35	9	1.83	5
MET	24	4.23	8	8	1.63	4
NTRK1	21	3.70	3	23	4.67	None
PIK3CA	27	4.76	2	62	12.60	7
RET	24	4.23	1-2	25	5.08	None
ROS1	28	4.94	1-2	49	9.96	None

Table 6.1: Analysis of frequency of known biomarkers of LUAD and LUSC for the TCGA mutation dataset. Comparison with the values in Villalobos et al. [48].

6.1.1.2 Clinical Data

As seen in the data analysis in chapter 4, the clinical datasets for LUAD and LUSC present missing values for some variables. Furthermore, there are mutually exclusive variables between the datasets, marked with *'NA'* in the list of the clinical parameters and its missing values in appendix Table A.1. The following steps of preprocessing described below were followed for each dataset:

- **Transformations** - For categorical types with no meaningful order, we convert the variables to binary using one-hot encoding. Ordinal encoding is used to convert variables with a natural rank-ordering, e.g. *paper_Tumorstage*. As a final step, we scaled the features into the [0, 1] range using min-max normalisation.

- **Handle missing data** - For categorical variables with missing values, we create an extra column representing missing values when doing one-hot encoding. For labels, more specifically the survival status, we drop these samples in the survival prediction problem. However, we generally impute missing values according to the average value for numerical types when applicable.

6.2 Survival prediction

This section focuses on explaining the details of our solution for survival prediction. On section 6.2.1, we explain our late fusion model's architecture, highlighting the differences to an early fusion technique. Section 6.2.2 discusses the unsupervised learning stage of the problem, explaining the methods used for feature extraction and section 6.2.3 explains the Cox Proportional-Hazards model used for survival prediction.

6.2.1 Multimodal Data: Early and Late Fusion

Survival analysis of cancer patients represents a more difficult task, where the integration of multiple data types can improve predictive performance by capturing complex multimodal associations required for accurate outcome prediction. In this work, we use three different data types: somatic mutation, which after preprocessing contains a binary composition representing the presence of specific mutations; clinical data, which contains a mixture of binary and continuous variables that are normalised into the [0, 1] range; and gene expression data that is continuous and scaled to fit into the 0 to 1 range.

We propose a method of Late Fusion (LF) to join the clinical and genomic modalities. Regarding the data heterogeneity problem, we will refer to two research papers that use both clinical with genomic data, which inspired our solution. In Chaudhary et al. [15], the authors combined clinical features with genomic features in the same model, which did not lead to an improvement in performance. However, in Wang et al. [49], clinical and genomic data was integrated by building a similarity matrix using the SNF algorithm [142], and a significant statistical improvement in the results was observed. This leads us to believe that the contradictory results obtained in [15] are related to the way the authors fed the original clinical and genomic features to the same model (Early Fusion), which may have induced redundancy in the overall features.

Late fusion represented on Figure 6.2b, integrates the modalities by sending each of them through an individual learner, joining their last layer's results. The fusion of each model's output can work, for instance, like majority voting (or weighted averaging, in the case of regression) if all models have a final classification step, or by feeding the intermediate layers containing latent features into a new model, thus allowing more flexibility. It is easier to handle a missing modality as the predictions are made separately; however, because LF operates on inferences and not the raw inputs, it is not effective at modelling signal-level interactions between modalities. Early fusion (EF) represented on Figure 6.2a, creates a joint representation of input features from multiple modalities; it uses one single model to make predictions, which assumes that the model is well suited for all the modalities [90].

We chose to use LF as there should not be signal-level interactions between the clinical and genomic modalities, so a technique of EF, as applied in [15], might induce data redundancy. Figure 6.3 presents the general architecture used for the late fusion of different modalities. Each modality is handled separately: we use a Variational Autoencoder (VAE) to extract a latent representation of high dimensional gene expression profiles and a more simple AE to extract features from the

(a) Early fusion of different modalities.

(b) Late fusion of different modalities.

Figure 6.2: Architecture to combine two distinct modalities using early and late fusion techniques.

mutation and clinical datasets. The autoencoder's outputs containing latent representations of the features are merged into a cox-regression layer that performs the regression of survival groups. To test our late fusion model's predictive performance, we implement an early fusion method similar to the one used in [15] to serve as a baseline. Furthermore, we compare the effectiveness of the early and late fusion strategies with a single modality strategy, implementing Cox PH models that rely solely on single modalities for prediction.

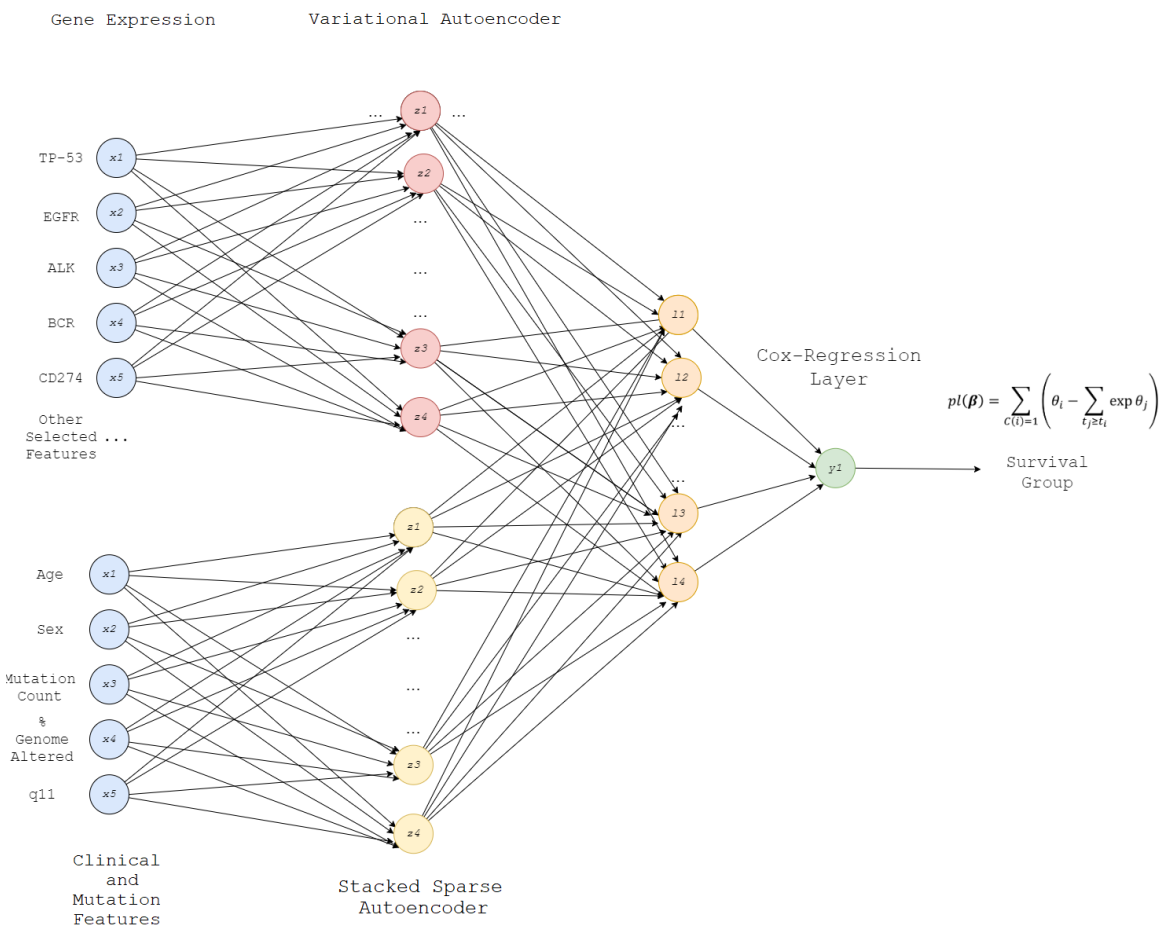


Figure 6.3: The architecture of the late fusion model used to combine the modalities for survival analysis. The modalities are fused in a Cox PH model after going through individual learners.

6.2.2 Dimensionality Reduction: Feature Extraction

Feature extraction provides some unique advantages over feature selection, as discussed in section 3.1.1, and has been widely used in the field of bioinformatics for dimensionality reduction. Due to the complexity of genomic datasets and the high amount of noise in the gene expression data, there is a need to analyse the raw data and exploit the important subsets of genes [28]. Linear FE techniques such as PCA that have shown satisfactory results for dimensionality reduction of gene expression data [75], however, this might fail to extract some non-linear relationships of the data. Non-linear FE methods techniques can capture more complex relationships lying on the inter-dimensional space, with the downside of the complexity and computational power required when compared to linear FE techniques. Autoencoders are a type of unsupervised learning technique that can be used to extract functional features from high dimensional gene expression profiles, and in Danaee et al. [28] the use of a stacked denoising autoencoder to extract latent features from gene expression has shown to outperform those extracted by PCA in the task of cancer prediction.

Autoencoders have been first introduced in [114] as a neural network that is trained to reconstruct its input. Their main purpose is to learn in an unsupervised manner, an informative representation of the data that can be used for various implications such as dimensionality reduction or clustering [5]. The problem can be formally defined as the objective is to learn the functions $A : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $B : \mathbb{R}^p \rightarrow \mathbb{R}^n$ so that the expectation E over the distribution of x in Equation 6.1

$$arg\,min_{A,B}\, E[\Delta(x,\, B \circ A(x))] \tag{6.1}$$

can be minimised, which is dependent on the distortion function Δ, as well as the presence of additional constraints, such as regularization [4]. In the particular case of deep learning, A and B are generally neural networks which leads to deep autoencoders.

6.2.2.1 Gene Expression Modality: Variational Autoencoder

In this work, we propose using a variational autoencoder to extract a latent representation of gene expression profiles, to reduce the data's dimensionality to be integrated into the survival prediction learner. VAEs first introduced in Kingma et al. [77], are generative models that attempt to describe data generation through a probabilistic distribution that can capture an underlying data manifold from input data.

Figure 6.4 presents the general architecture of the implemented VAE. First, we start by transforming the original gene expression sets by dividing them by their max so that expression is capped in the [0, 1] range, which does not affect the original scales between genes as normalisation using min-max or standard scaling would do. Batch Normalisation (BNorm) is used to transform the gene expression input dimension (x) so that the dense layers maintain a mean output close to 0 and the output standard deviation close to 1 [42], which enables for more stable training and faster convergence. VAEs are stochastic and learn the distribution of explanatory features over samples by learning two distinct latent representations: a mean (μ) and standard deviation vec-

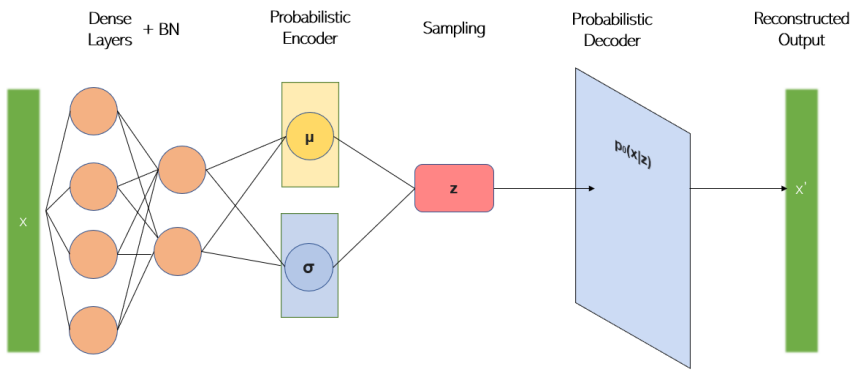


Figure 6.4: The variational autoencoder's architecture, X is the input RNA-seq profiles connected to dense layers with BNorm, which are sampled in the lambda layer and reconstructed into X'.

tor encoding (σ) [133]. A lambda layer is created where we define a custom activation function to perform the probabilistic sampling (z) on Equation 6.2 with ε being drawn from a Gaussian

$$z = \mu + \sigma \odot \varepsilon \tag{6.2}$$

distribution with 0 mean [133]. We assume an approximate posterior distribution over the latent variable (z) given a sample x_i with some known distribution $p_\theta(z||x)$ that can be ascertained. Symmetrically, a probabilistic decoder is assumed for each sample x_i conditioned on an unobserved random latent variable z_i , where θ are the parameters governing the generative distribution. To calculate the loss associated with the reconstruction two factors need to be considered: the loss associated with the reconstruction error, and the Kullback-Leibler (KL) that regularises weights by constraining the latent vectors to match a Gaussian distribution [133]. The loss function represented in Equation 6.3 is known as variational lower bound, and through minimising the loss, we are maximising the lower bound of the probability of generating new samples [5].

$$L(\phi, \theta : x) = E_{z \sim q_\phi(z|x)}(log(p_\theta(x \mid z)) - D_{kl}(q_\phi(z \mid x)||p_\theta(z \mid x)) \tag{6.3}$$

To calculate the reconstruction loss we use Mean Squared Error (MSE) in Equation 6.4, to average the squares of the errors between the gene expression values and the decoded values, and created a layer connecting the inputs with the decoded layer with a custom loss function, defined as the mean of the KL loss and the reconstruction loss given my MSE.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(d_i - y_i)^2 \tag{6.4}$$

The ReLu activation with He uniform initialization [62] was used for dense layers, except the probabilistic decoded layer which uses sigmoid for prediction. The hyper-parameters were optimised using grid search: Nesterov-accelerated Adaptive moment estimation (NAdam) optimisation with a learning rate of $1e^{-3}$ and batch size of 50 seemed to minimize the loss for latent dimension size's (Z) of 120.

6.2.2.2 Clinical and Mutational Modalities: Stacked Sparse Autoencoder

The mutational and clinical modalities present a more simplistic dataset composition when compared with the gene expression data. After preprocessing, the clinical dataset comprises two types of continuous and binary variables, and the mutational dataset contains only binary variables representing specific mutations. To extract features from these modalities, we used shallow autoencoders with an L1 penalty that was added to the encoder layers to encourage sparsity and avoid bad representations by creating features with large weights [21]. Grid search was used to find the optimal architecture and the hyper-parameters for each modality, by doing 5-fold cross-validation with MSE as the reconstruction loss for the clinical modality and binary cross-entropy for mutational data. Table 6.2 presents the results of grid-search optimisation: the activation ReLu with He uniform initialisation [51] was selected, and NAdam optimisation seemed to reduce the loss for both autoencoders. The optimal number of encoder and decoder layers was one for the clinical SSA and a more deep architecture with two encoder, and three decoder layers seemed optimal for the higher dimensional mutation dataset. L1 regularisation was also individually optimised for each encoder layer to allow for more differentiation between layers' ability to induce sparsity on the features. The grid search space defined for the architecture was 1 to 4 layers for the encoder and decoder, respectively, with a number of nodes between 10 and 500. The batch sizes were capped at 50 due to processing power; the learning rate was optimised in the $[1e^{-6}, 1e^{-2}]$ range, and logarithmic scale was used to search for L1 values in the $[1e^{-1}, 1e^{-10}]$ gamma.

Hyper-Parameters	Clinical	Mutation
Encoder	40	{120, 80, 90}
Decoder	60	{80, 50}
Bottleneck	25	50
L1 regularisation	$5e^{-6}$	$\{ 2e^{-7}, 1e^{-4}, 8e^{-5} \}$
Learning Rate	0.00263	0.0000347
Batch Size	40	60

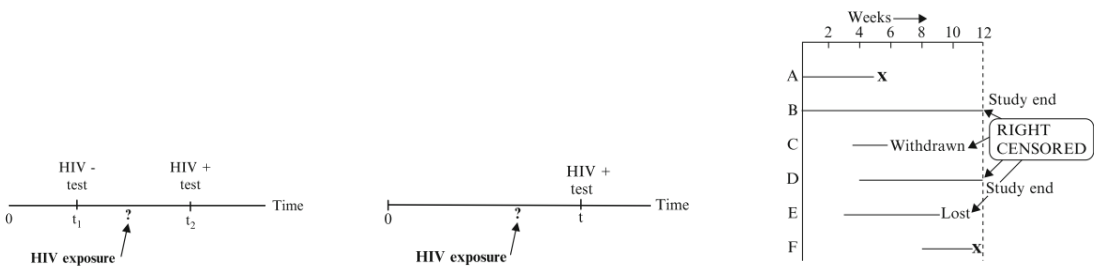
Table 6.2: Optimal hyper-parameters for the clinical and mutation SSAs, resulting from grid-search optimisation.

6.2.3 Cox Proportional-Hazards

Survival analysis is a term coined for the collection of statistical procedures for data analysis, for which the outcome variable of interest is time until an event occurs [24]. In the survival analysis of cancer patients, the event of interest is death, and the time variable is the survival time. In section 6.2.3.1, we start by giving a background on some key survival analysis concepts that must be understood before delving into the specifics of the Cox Proportional Hazards [25] model used for survival analysis in section 6.2.3.2.

6.2.3.1 Background

Censoring: Censoring is a common problem in medical domain survival analysis, and it occurs when we have partial information on the subject survival time, but we do not know the exact survival time. Figure 6.5 presents the three different types of censoring existent: interval--censoring occurs when the event of interest is within a known time interval; left-censoring when



(a) Interval-Censoring (b) Left-Censoring (c) Right-Censoring

Figure 6.5: Type of censoring in survival analysis. Adapted from [78].

the real survival time is less than or equal to the observed survival time; right-censoring happens when the real survival time is equal to or greater than observed survival time, e.g. if a patient leaves the study or the study ends. In our specific case, only right-censoring is relevant as it is the only type of censoring present in the data, which translates to patients with unknown survival time or patients still alive.

Survivor and Hazard Functions: The survivor function $h(t)$ gives the probability of subject surviving over time. It is theoretically a continuous function starting at $t = 0$ with a probability of survival $h(0) = 1$ that decreases over time, which empirically translates to a step function rather than a smooth curve. The hazard function $h(t)$ in equation 6.5 equals the limit, as

$$h(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t \mid T \geq t)}{\Delta_t} \tag{6.5}$$

Δ_t approaches zero, of a probability statement about survival, divided by Δ_t which denotes a small interval of time [78]. This roughly translates to the instantaneous risk of the event occurring in an expeditious time period Δ_t.

6.2.3.2 Cox Proportional Hazards Model

The Cox Proportional Hazards model (1972) [25] it is still the most widely used method for multivariate survival analysis due to its solid statistical background. The idea behind Cox's PH model is that the log-hazard of an individual is a linear function of their covariates and a population-level baseline hazard that changes over time [33], as represented on Equation 6.6.

$$h(t|x) = b_0(t) \ \exp\left(\sum_{i=1}^{n} b_i(x_i - \overline{x_i})\right) \tag{6.6}$$

The formula translates to the hazard at time t for a predictor variable x is given by the baseline hazard which is a time-dependent function giving the hazard for an individual at time zero ($b_0(t)$), multiplied by the partial hazard, or Hazard Ratio (HR), which is a time-invariant scalar factor that only increases or decreases the baseline hazard of the covariates (x_i) [33].

A fundamental assumption made by this model concerns the proportional hazards assumption, in which the baseline hazard is a function of t, but does not involve the hazard of the covariates; in contrast, the hazard ratio is given by the effect of the time-independent covariates. This method is considered semi-parametric as it contains a parametric set of covariates (partial hazards) and a non-parametric component (baseline hazard) [25]. On Equation 6.7, it is presented a fully-parametric version of the Cox PH model, called the Cox time-varying proportional hazard model which as opposed to the original model considers that the hazard of a covariate can change over time ($x_i(t)$).

$$h(t|x) = b_0(t) \ \exp\left(\sum_{i=1}^{n} \beta_i(x_i(t) - \overline{x_i})\right) \tag{6.7}$$

Although the parametric version of the model is theoretically more powerful, in practice generally the semi-parametric version of the model is more fit if the goal is to maximise the score of survival prediction [78]. This will be discussed more in-depth in section 6.2.3.3, and in Table 6.3 it is presented a summary of the statistics describing the fit, the coefficients, and the error bounds of a Cox-PH execution for 5 clinical covariates. The clinical covariates along the rows from top to bottom represent the shortest and intermediate dimension of the patient's tumour, the initial weight when enrolled in the study and the AJCC metastasis and pathologic stage classifications.

Covariate	C	***HR(C)***	$\sigma(C)$	$\underline{HR(C)}$	$\overline{HR(C)}$	z	p	-$\log_2(p)$
short._dim	-0.042	0.959	0.220	-0.474	0.390	-0.189	0.850	0.234
int._dim.	0.240	1.272	0.197	-0.146	0.627	1.218	0.223	2.164
init._weight	0.926	2.524	0.249	0.437	1.415	3.712	0.0005	12.249
ajcc_m	0.711	2.037	0.157	0.405	1.018	4.544	0.0001	17.467
ajcc_stage	1.028	2.794	0.252	0.533	1.522	4.075	0.0003	14.409

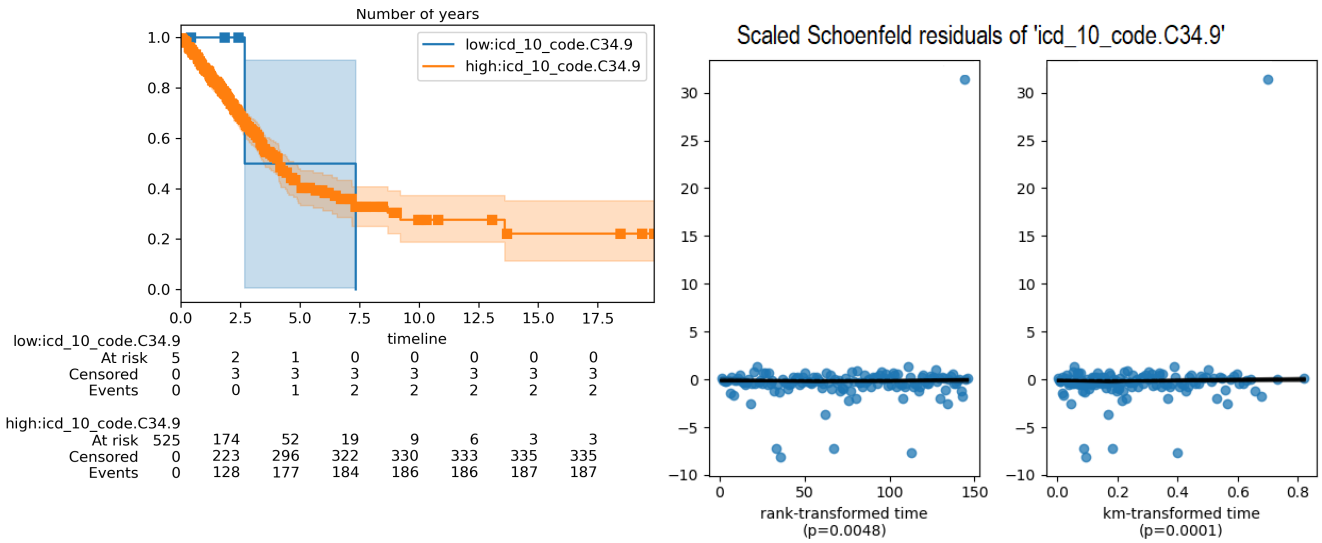
Table 6.3: Summary statistics describing the fit, the coefficients, and the error bounds of a Cox-PH execution.

Across the columns, C if the regression coefficient of a specific covariate, $\exp(C)$ gives the Hazard Ratio of the variable, $\sigma(C)$ is the standard error of the estimated regression coefficient, and the fourth and fifth columns are the lower and upper bounds of the 95% confidence interval for the hazard ratio of the covariate C. The z is the value of the Wald statistic which evaluates if the coefficient of a given variable is statistically significantly different from 0 and p is the result of the log-rank p-value that tests for the significance of each coefficient on the separation of survival groups. The last variable is the log_2 value of the Likelihood Ratio (LR) which can be used to assess the impact of covariate without model interaction [78].

6.2.3.3 Evaluating Proportional Hazards Assumption

This section is related to the validation of the Cox Proportional Hazards assumption, and it evaluates the advantages and disadvantages of using the Cox PH semi-parametric or the Cox time-varying PH model according to the literature.

When using the Cox PH semi-parametric version, it is required to verify the proportional hazards assumption in order to validate the model statistically. As discussed in the previous section, the Cox PH premise assumes that a covariate contribution for the partial hazards is constant throughout time. If a covariate breaks the premise, it implies that the contribution to partial hazards changes over time, making it a time-dependent covariate. The scaled Schoenfeld residuals test, visually represented in Figure 6.6, is used to test whether a covariate violates the proportional hazards premise by creating a time-varying coefficient in a (fictional) alternative model that allows for time-varying coefficients [108]. A non-failed test would present a disperse distribution over the x-axis with no clear trend. Variable *icd_10_code.C34.9*, which stands for a type of unspecified



(a) KM estimate of *C34.9* subjects and control group. (b) Schoenfeld residuals test on covariate.

Figure 6.6: Assessment of the Cox PH premise on covariate *icd_10_code.C34.9* using the scaled Schoenfeld residuals test.

lung neoplasm according to the World Health Organization (WHO) International Classification of Diseases (ICD-10) classification [106], failed the non-proportional test with a global p-value of 0.0001 and a rank-transformed time p-value of 0.0048 . However, failing the test does not imply that the variable is time-dependent. In Figure 6.6a, we plot the difference in survival groups between subjects classified as *icd_10_code.C34.9* and the control group using the KM estimate, and it is possible to see that the low cardinality induces the failed test, which is caused by some arbitrary dependency of the covariates' partial hazard when analysed in a time-varying model, given that all uncensored subjects that have the covariate positive, die in a short period.

The Cox time-varying model is more fit for situations when hazard ratios change over time, such in the case of follow-up trials, and even if the hazards were not proportional, altering the model to fit a set of assumptions fundamentally changes the scientific question which leads to poor results [33]. Given a large enough sample size, even minimal violations of proportional hazards will show up [122], therefore when the objective is to maximise a score such as the survival prediction score it is wiser to chose the best predicting model even if some violations occur.

6.2.3.4 Dimensionality Reduction: Feature Selection

The curse of dimensionality is severe when modelling high-dimensional discrete data: the number of possible combinations of the variables explodes exponentially [7]. Genomic datasets are generally composed of sampled high-dimensional vectors with very sparse characteristics, which increases the risk of overfitting and making it very difficult to identify patterns in the data without having plenty of training data [54].

The Cox Proportional Hazards model by evaluating each covariate's fitness for survival prediction makes it possible to use the penalised likelihood of covariates, that induces sparsity in fitted parameters, for better estimation of variables log-likelihood [78]. Therefore, it is possible to rank the features in descending order of their log-likelihood or select them by their p-value, which is a necessary step so that the Cox PH model can converge.

6.3 Train, Test and Evaluation

On figure 6.7, we show the full pipeline followed for the unsupervised and supervised learning parts of the problem. We use stratified sampling in order to maintain the same proportion of non-censored and right-censored patients across the train and test sets, with a split of 60 to 40 per cent so that there is covariate variability across the test set. Since the estimation of the coefficients in the Cox proportional hazard model is done using the Newton-Raphson algorithm, there are sometimes problems with convergence [33], variables with low variability which completely separate censored from non-censored subjects must be dropped so that the Cox PH model can perform adequately. To handle low variability covariates, we perform two preprocessing steps:

1. We run the Variance Inflation Factor (VIF) algorithm to find multicollinearity in the predictor variables [111], and drop all variables above the 8.0 VIF threshold

2. At each train/test split, we check for variables that completely separate censored from non-censored subjects, for example, for all "death" events in the dataset, there exists a covariate that is constant amongst all of them, we drop these variables so that the model converges.

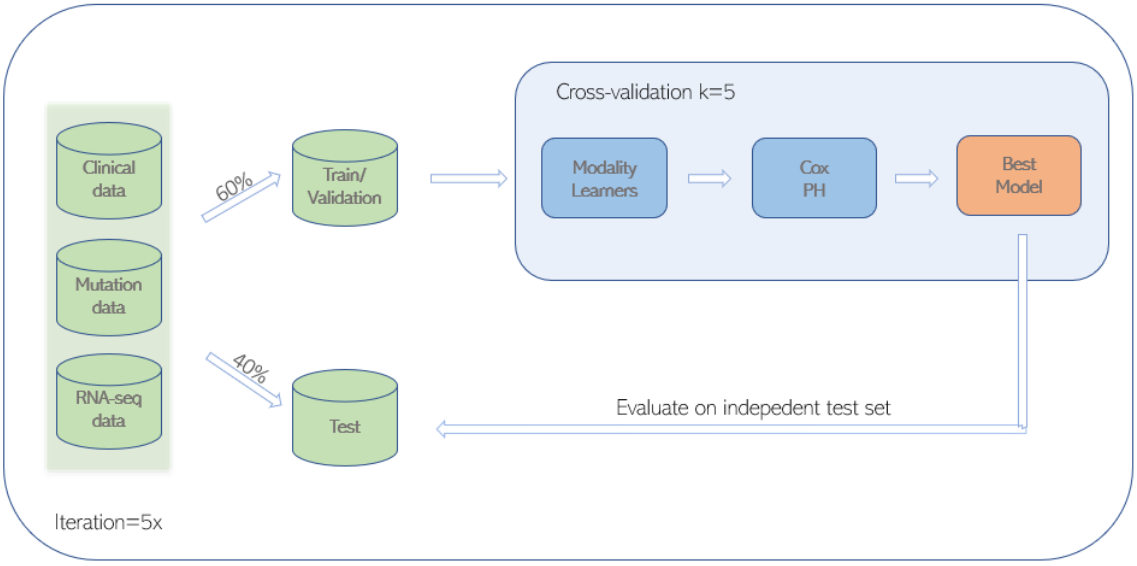


Figure 6.7: Global pipeline for LUSC and LUAD survival prediction. The autoencoders and the Cox PH model are optimised using 5-fold cross-validation, and the train and test process is repeated 5 times for model validation.

Modality Learners: This comprises the unsupervised stage of the problem, where we run each modality learner using a 80 to 20 per cent split for train and validation purposes, as explained in the section 6.2.2, which is used to train the autoencoders for feature extraction.

Cox PH: This stage comprises two different stages: first feature selection is performed by performing a log-rank test to evaluate the fitness of each variable for survival prediction for each separate

modality features; following, we feed all the features with a p-value below a pre-defined threshold for the late fusion model as explain in the LF section 6.2.1. To optimise the Cox PH deep network, we use 5-fold cross-validation in order to get the best selection of network hyper-parameters, using the Akaike Information Criterion (AIC) to compare the predictive performance of different survival models. Finally, we used the best performing cox model to evaluate the performance on the independent test set, and the full process was repeated five times by randomising the train test seed to account for the variability induced by the small survival sets. The evaluations metrics Concordance Index and the p_{value} of the log-rank test were used to assess the model' predictive performance and the covariates' fitness for survival analysis, in order to best compare our results against similar research, and the Kaplan-Meier estimate was used for the analysis of survival curves.

Concordance Index: The C-Index or Harrell's C on Equation 6.8, is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. It estimates the probability of concordance between predicted and observed responses [59].

$$C\text{-}Index = \frac{\sum_{i,j} I(T_i > T_j)(\eta_j > \eta_i)\Delta_j}{\sum_{i,j} I(T_i > T_j)\Delta_j} \tag{6.8}$$

Where i and j refer to pairs of observations, η refers to the prediction and T to the actual survival time. The pair (i, j) is concordant if the survival prediction for observation j is greater $(\eta_j > \eta_i)$, then the survival of j is longer $(T_i < T_j)$. The overall C-index can be translated as the ratio of concordant pairs by the total number of observations multiplied by the factor Δ_j that discards pairs of observations that are not comparable when observation T_j is right-censored. Harrell's C can also be interpreted as a summary measure of the area(s) under the time-dependent ROC curves [63], where a C-index of 0.5 determines no predictive discrimination, and 1.0 is a perfect separation between patient's outcomes.

Kaplan-Meier Estimate: Kaplan–Meier (KM) estimator plots the KM survival curve, which is defined as the probability of surviving in a given length of time while considering the time in many steps [3]. It involves computing of probabilities of occurrence of an event at a certain point of time and multiplying these successive probabilities by any earlier computed probabilities to get the final estimate [52].

Log-rank $\mathbf{p}_{value}$ of Cox-PH regression: The log-rank test is a hypothesis to assess the difference between two survival curves, and assign it a confidence interval. It calculates the chi-square (X^2) on formula 6.9 for each event time, and for each group and sums the results. It is further

possible to create strata of the population. On Equation 6.10, it is presented a variation

$$\chi^2 = \frac{(\sum_{i=1}^{m} d_i - \sum_{i=1}^{m} e_i)^2}{\sum_{i=1}^{m} \sigma_i^2} \tag{6.9}$$

$$O_i - E_i = \sum_s \sum_j (m_{i,j,s} - e_{i,j,s}) \tag{6.10}$$

of the log-rank test when stratifying by a covariate, where instead of calculating (X^2) for group i, we calculate it for each stratum (s). The hypothesis returns a p_{value} used to test if the survival between two groups presents a statistically difference ($p_{value} < 0.05$).

Akaike Information Criterion: AIC on Equation 6.11, provides an approach for comparing the fit of models with different underlying distributions, making use of the -2 $log_2\ likelihood$ statistic, and the addition of 2 times p can be thought of as a penalty if non-predictive parameters are added

$$\text{AIC(model)} = \text{-2 LR} + 2\,k \tag{6.11}$$

to the model. Where k is the number of parameters (degrees-of-freedom) of the model and LR is the maximum log-likelihood. A smaller AIC statistic suggests a better fit since it is a trade-off between maximising LR with the fewest parameters possible [78].

Implementation

Our solution is implemented in Python 3.8 programming language using the frameworks Keras [21], Tensorflow [42], KerasTuner [104] and lifelines [33]. All CPU intensive tasks, namely preprocessing, were conducted on a 16x2GHz cores Intel Core Haswell CPU with 128Gib RAM running on the Ubuntu 16.4 operating system. All GPU taks were run on a 2xNvidia GeForce GTX 1080 Ti with 11,178 MiB GDDR3 running on Ubuntu 18.04 LTS operating system.

6.4 Results

On this section we compare the results of the early and late fusion techniques with single modality learners, and perform explorative data analysis to assess the fitness of the different survival models' features. The results are presented separately for LUAD and LUSC cohort as they show different overall survival curves

Figure 6.8 presents a KM plot of the LUAD and LUSC cohorts survivability. The bottom left corner's risk counts show the total number of subjects enrolled in the study, 530 for LUAD and 496 for LUSC. For each time step (years) the number of deaths is updated, which corresponds to the at-risk patients minus the total number of right-censored patients. The shadow surrounding each curve corresponds to the 95% confidence interval associated with the KM estimate, and the squares represent the presence of a right-censored observation at time t. The most evident differences differences between the histologic subtypes survival curves are: LUSC survival is more aggressive in the first three years, and shows a worse perspective of survival in the long term, stabilising at 19% while LUAD stabilises at 21%, and LUAD survival seems to have higher mortality between the 4 and 6 years period, after which it slows down. Although this type of analysis can give a visual estimate of each group's survival tendency, it should be complemented with a long-rank test that evaluates if there is a statistically significant difference between the groups' survival. The log-rank test returns a p-value of 0.24 that rejects the null hypothesis under alpha of 0.05, which means it is not conclusive that the LUAD and LUSC populations present different hazard functions according to the hypothesis.

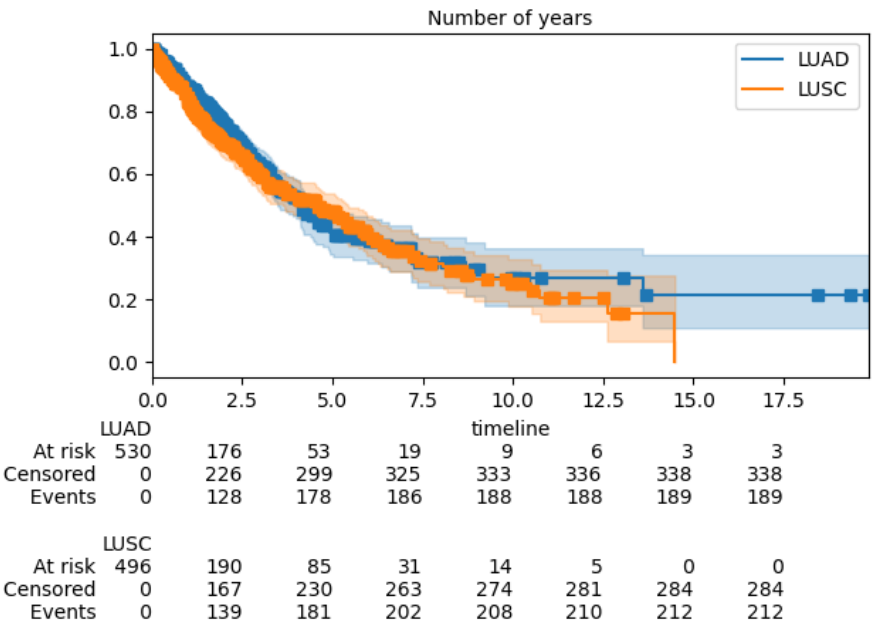


Figure 6.8: Kaplan-Meier estimate of the LUAD and LUSC survival curves with log-rank test p-value of 0.24.

6.4.1 Lung Adenocarcinoma

Table 6.4 presents the results of the different strategies for predicting LUADs cohort survival across five iterations, where we randomised the train and test splits to get a more reliable estimate of predictive performance and to account for variance induced by the shorter sample sizes.

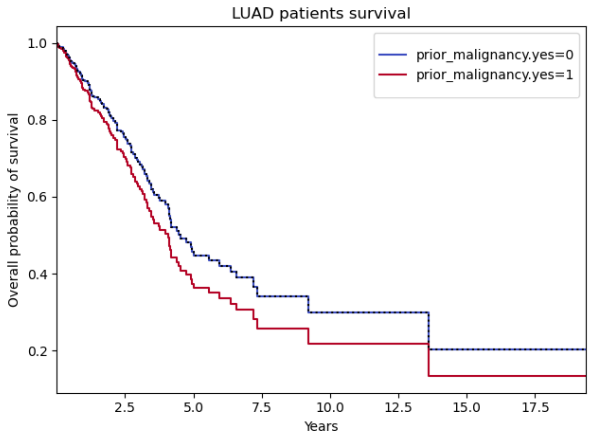
Out of the single modality models, the clinical model performed better with a C-Index of 0.638 and a median AIC indicating less model stability across the five iterations, while also presenting slightly higher deviations than the gene expression model. A total of 18 covariates showed p-values below 0.05, and the fittest variable was *prior_malignancy_yes*, with a p-value of 0.00009, and a full list of the variables with p-value's below 0.05 can be found in the Appendix section B on Tables B.1, B.3 for each modality respectively. The gene expression modality was the second-best, with a C-Index of 0.616; however, it showed the best overall log-likelihood score, translating to a higher average number of better-fitted covariates with 19 features presenting p-values below the 0.05 threshold. The mutational model showed more instability and worse predictive performance, having four features below the p-value threshold.

	Performance Metrics (mean ± standard deviation)		
Methods	***C-Index***	***AIC***	***-log$_2$(p) ll-ratio***
EF (Clinical + Genomic)	0.666± 0.015	1285.657 ± 9.271	8.366 ± 2.213
LF (Clinical + Genomic)	**0.701 ± 0.009**	**1039.991 ± 10.624**	8.397 ± 2.409
Clinical	0.638 ± 0.048	1497.859 ± 30.277	9.480 ± 3.920
Mutation	0.597 ± 0.052	2047.684 ± 2.028	9.978 ± 0.020
Gene Expression	0.616 ± 0.045	1169.246 ± 19.272	**8.365 ± 3.268**

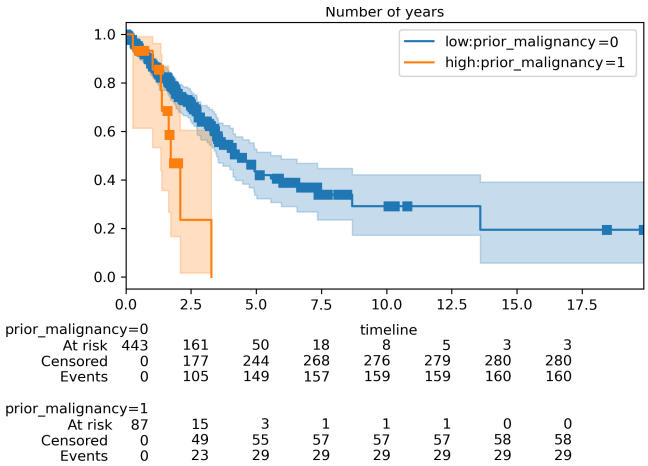
Table 6.4: Results of different survival analysis strategies for the prediction of LUAD survival over 5 iterations.

Both types of fusion methods showed gains in overall predictive performance; however, the late fusion presented the best overall score, with a C-Index of 0.701 and the lowest AIC showing the best trade-off between the log-likelihood and number of parameters. Furthermore, it showed a reduction in variance when compared to the early fusion technique.

On Figure 6.9, we present two types of analysis used to interpret the effect of the variable *prior malignancy* on survival, which indicates a malignancy identified prior to the diagnosis of the specimen submitted for TCGA [70]. In the first analysis on Figure 6.9a it is represented the effect of the covariate on the cox-ph model, which translates to varying the value of the covariate while holding everything else equal. This analysis is useful to understand the impact of the covariate on the model; however, it is only a conjecture on the covariate's contribution to the partial hazards regression and does not show the real impact of the variable on the subject's survival. Figure 6.9b presents the KM estimate of the LUAD subjects with prior malignancy and the control group, representing all subjects with no known prior malignancy or missing values for the prior malignancy variable. A total of 53 patients have had previous malignancies and show a more accelerated failure in the first three years, after which only right-censored patients remain.



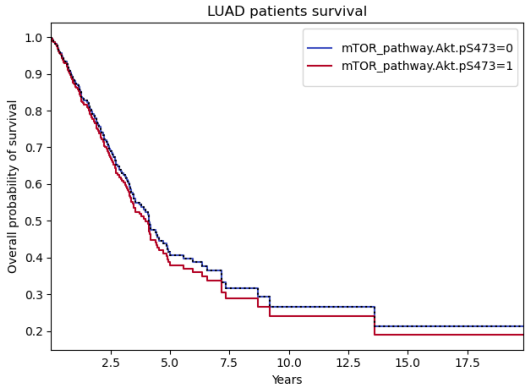
(a) Effect of the covariate on the Cox PH model.



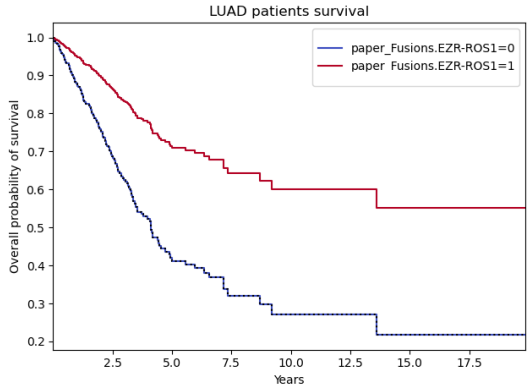
(b) KM estimate of the prior malignancy and control groups.

Figure 6.9: Analysis of the subjects with prior malignancy in LUAD population and log-rank test with p-value 0.00009

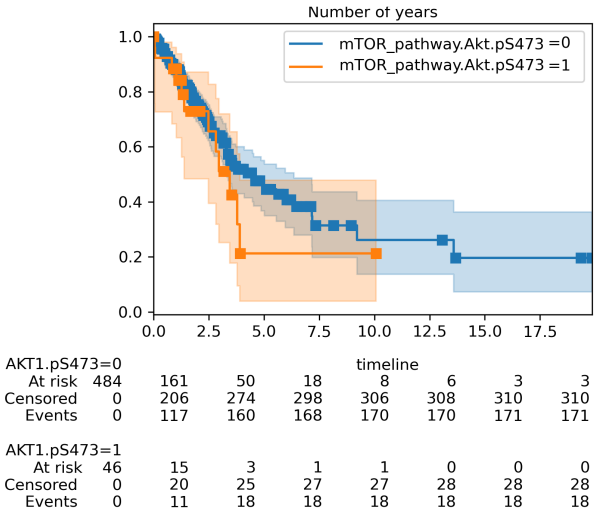
Although the predictive performance results were not satisfactory concerning the mutational model, achieving the lowest C-Index, two interesting patterns were found on the data, upon analysis of differentiable mutations. Figure 6.10 shows the analysis of two types of patterns found in the mutational data, where we analyse mutations and fusions with opposing effects on overall survival. Figures 6.10a, 6.10c show the effect of the AKT1 p.S473 mutation on the model and the KM estimate of the mutated and non-mutated subjects with a log-rank test p-value of 0.47, and although it fails the null-hypothesis test, it is possible to see a faster failure rate after three years. This type of patterns where mutations lead to accelerated failure is key to identify pronto-oncogenes, which in the case of AKT1 is a known oncogene present in 0.63% of NSCLC patients [45]. The second type of pattern found in Figures 6.10b, 6.10d related to the EZR-ROS1 fusion, presenting a contrary effect by showing an improvement in overall survival in subjects containing the fusion. However, the KM estimate shows that although there is a big difference in survival of the groups, the at-risk



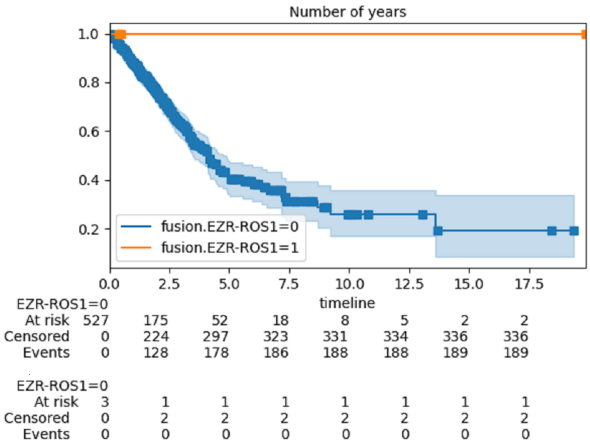
(a) AKT1 mutation effect on Cox PH model.



(b) EZR-ROS1 fusion effect on Cox PH model.



(c) KM estimate of AKT1 mutated and control group.



(d) KM estimate of EZR-ROS1 fusion and control group.

Figure 6.10: Analysis of AKT1 p.S473 mutation and EZR-ROS1 fusion on LUAD patients.

group presents a very low cardinality of only three subjects, being that 2 remain right-censored till the end of the study, which results in a log-rank test p-value of 0.19. The log-rank test and the Cox proportional hazards work account for the right-censored observations when calculating the population hazards; however, this example highlights a critical problem in the mutational dataset: it is full of low cardinality covariates which is partly the reason why its predictive performance is very unstable.

6.4.2 Lung Squamous Cell Carcinoma

Table 6.5 presents the results for the survival analysis of the LUSC cohort. The overall results show worse predictive performance for the LUSC subtype than for LUAD, which might be connected to the lower sample count of the LUSC cohort, as the same methodology was followed for both histologic subtypes.

Out of the isolated modality models, the mutational presented the best predictive performance with C-Index of 0.568, however, it was the more unstable model showing higher standard deviations and a worse trade-off between log-likelihood and number of selected features, and with a total of 5 features presenting p-values under 0.05 present on Appendix Tables B.4, Tables B.2 for each modality. The clinical model performed second-best with better predictive performance and a better trade-off than the mutational model, probably due to the latter's larger dimension and sparsity. The gene expression model performed worst, however, it showed to be the most stable and with the best overall score of AIC, indicating the best trade-off between the number of variables and the log-likelihood.

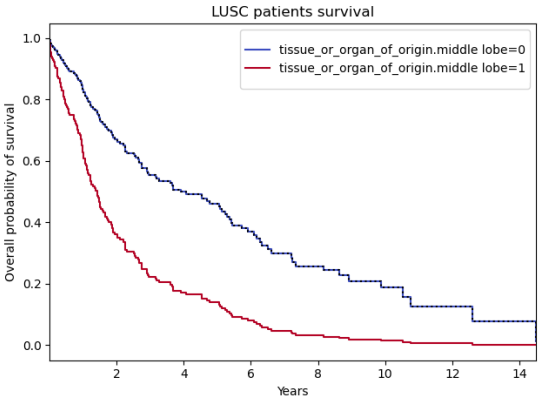
	Performance Metrics (mean ± standard deviation)		
Methods	***C-Index***	***AIC***	***-log$_2$(p) ll-ratio***
EF (Clinical + Genomic)	0.571 ± 0.032	1507.677 ± 60.278	**8.479 ± 4.914**
LF (Clinical + Genomic)	**0.622 ± 0.010**	1452.501 ± 60.868	8.489 ± 4.928
Clinical	0.564 ± 0.046	1640.654 ± 59.751	9.624 ± 4.901
Mutation	0.568 ± 0.071	2307.187 ± 4.313	10.11 ± 1.109
Gene Expression	0.552 ± 0.002	**1314.199 ± 72.057**	**8.651 ± 5.171**

Table 6.5: Results of different survival analysis strategies for the prediction of LUSC survival over 5 train/test iterations.

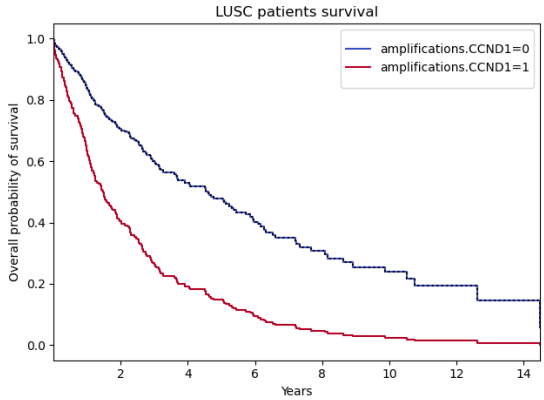
The late fusion method significantly improved predictive performance with a C-Index of 0.622 and fairly lower variances than the other models. Although the early fusion model presented the best average log-likelihood, showing the higher average number of fitted variables, it did not show a statistical improvement over the single modalities and incurred in higher variances, with a maximum test score of 0.722 and a minimum of 0.525. We suspect that the reason early fusion did not show a statistical improvement overall is related to the integration of the mutational data that presents a very sparse structure for the LUSC dataset with a high number of low cardinality variables that penalise overall performance, nevertheless, it still was able to achieve lower standard variations than the single modalities. To our knowledge, the late fusion model was able to further reduce the variance due to the extraction of features from the mutational data that was able to eliminate redundant features as opposed to the early fusion model. The late fusion model had a total of 28 variables with p-value under 0.05 which are going to be examined onward on section 6.4.3, where we analyse the non-interpretable features.

On Figure 6.11, we present the analysis of the most significant variables for LUSC: the clinical variable *tissue_or_organ_of_origin.middle lobe* on Figures 6.11a, 6.11c that identifies subjects

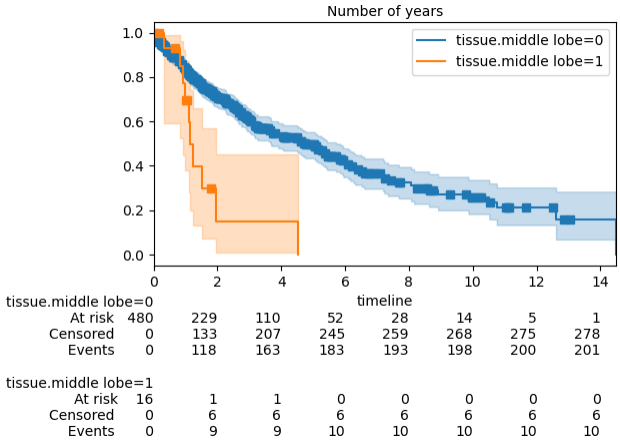
whose tumour developed on the middle lobe of the lung, showing very high separability and a p-value of 0.000037, which leads to accelerated failure rate in the first two years; and the mutational variable *amplifications.CCDN1* on Figures 6.11b, 6.11d, signalling an increase in the number of copies of the CCDN1 gene showing a p-value of 0.028 and leading to worse overall survival in the first three years.



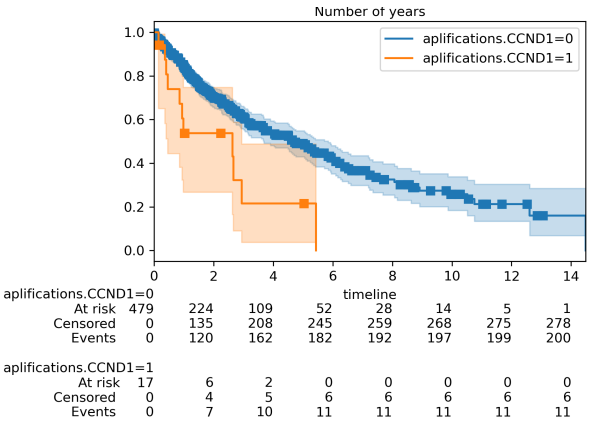
(a) Effect of tissue of origin tumour is middle lobe on the Cox PH model.



(b) Effect of CCDN1 amplification on Cox PH model.



(c) KM estimate of subjects with origin tumour on middle lobe and control group.



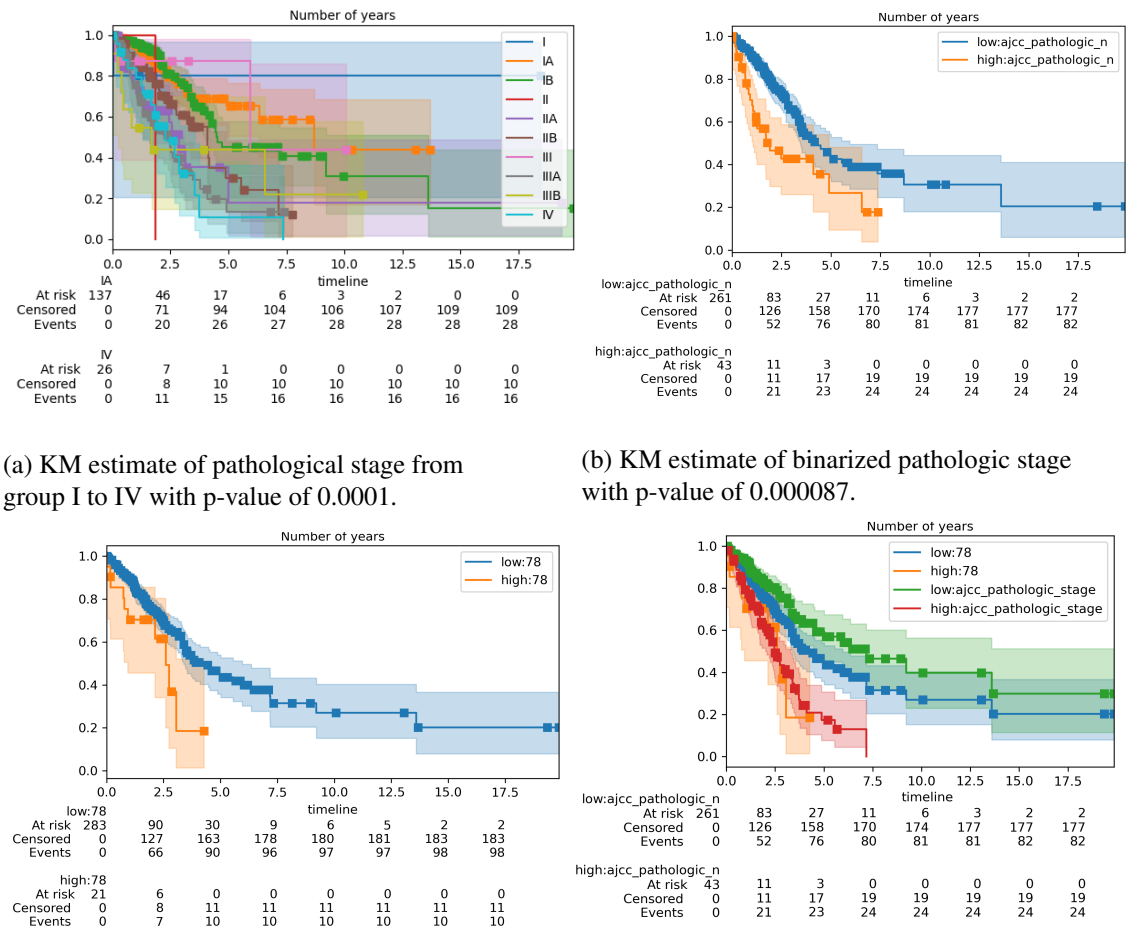
(d) KM estimate of subjects with amplification on CCND1 and control group.

Figure 6.11: Analysis of origin of tumour tissue is middle lobe and amplification of CCDN1 on LUSC patients.

6.4.3 Most Separable, Non-Interpretable Features of the Late Fusion Model

To validate the late fusion model's effectiveness and analyse the features generated by the unsupervised learning methods, on this section we compare the results of these non-interpretable features with clinical features that were selected as statistically significant in the previous sections for both histologic subtypes.

Staging for lung cancer based on the TNM classification is a system to stratify cancer patients which serves as a guide to treatment and an indicator of prognosis [103]. For NSCLC according to the last AJCC revision [71] there are four distinct stage groups: stage I to stage IV, with subgroups A and B to further stratify between levels, being level B more severe than A, and stage IV the worst prognostic stage. Figure 6.12 presents a comparative analysis between the AJCC pathological stage and a non-interpretable feature from the late fusion model for the LUAD cohort. Figure 6.12a shows the KM estimate of the LUAD patients stratified by the pathological stage and sub-level, that was previously analysed in the data analysis chapter 4, and with a p-value of 0.0001 shows



(a) KM estimate of pathological stage from group I to IV with p-value of 0.0001.

(b) KM estimate of binarized pathologic stage with p-value of 0.000087.

(c) KM estimate of late fusion non-interpretable feature with p-value of 0.000026.

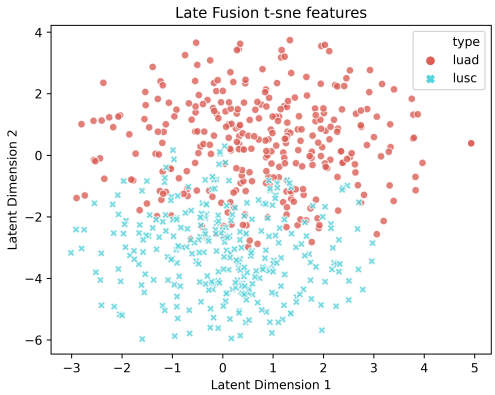
(d) KM comparison of the binarized pathologic stage and the late fusion feature groups.

Figure 6.12: Comparison of separability between AJCC pathologic stage and late fusion feature (#78) for LUAD patients.

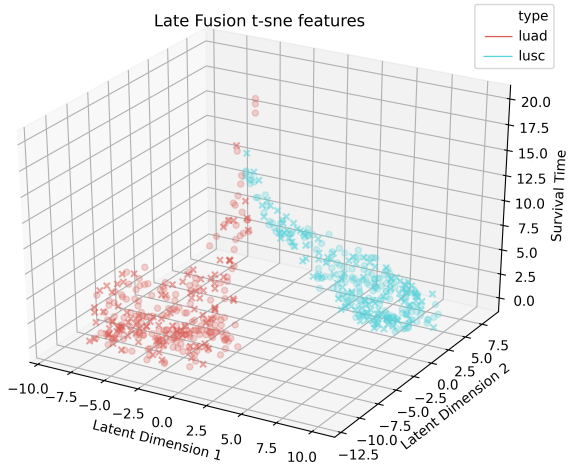
a clear statistical separation between the different stage groups. On Figure 6.12b, we binarise the subjects into two groups: low pathological stage comprising stage I to stage III, and high pathological stages from stage IIIA to stage IV, resulting in further differentiation of the survival curves with log-rank test p-value of 0.000087 and with an apparent higher failure rate in the first five years for higher stages, which is strikingly visible in the first months of survival time.

Figure 6.12c presents the KM estimate of a continuous feature ranging from 0 to 1 prevenient from the feature extraction stage used by the late fusion model for survival prediction (feature #78). To analyse its impact, we binarised the variable using quantile-based discretisation to separate the data into two bins according to its mean value, resulting in greater separability with a p-value of 0.000026. Due to the lack of feature interpretability, it is impossible to measure its quality for survival prediction, however, to quantify its impact on survival, in Figure 6.12c we compared the survival groups of the LF variable with the pathological stage strata. The analysis shows that higher values of the LF variable seem to lead to worse survival failure than high pathological stages, particularly between the 2 and 4 years-period, while low values of the feature lead to somewhat worse survival than the low cancer stages. Furthermore, the confidence intervals, given by the margin of error of the KM estimate, seem to overlap more on the pathological stages, especially after two years, while for the LF feature they overlap more before the 2 years-period. On Appendix section B, the Tables B.5, B.6 present the full list of late fusion covariates with p-values under 0.05 for LUAD and LUSC, respectively.

On Figure 6.13 we present a visualisation based on the t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm [128] that depicts high-dimensional data by giving each datapoint



(a) 2D visualisation of the set of 195 features extracted in the unsupervised learning stage based on t-SNE.



(b) 3D visualisation where X and Y are t-sne dimensions and Z is the survival time in years. Crosses (x) represent decesead subjects and circles (o) alive patients.

Figure 6.13: 2D visualisation of the set of 195 latent features, extracted in the unsupervised learning stage, based on t-sne for LUAD and LUSC cohorts.

a location in a two or three-dimensional map. It is based on the Kullback-Leibler divergence, much like the VAE used in this work that minimises the difference between two probability distributions. To better understand the set of 195 features prevenient from the unsupervised learning stage, we ran 2,000 optimisation iterations of the t-SNE algorithm with 50 nearest neighbours to visualise it the two-dimensional (2D) t-SNE space.

Figure 6.13a presents the visualisation in the 2D space, where the red labels represent LUAD features and the blue's LUSC. The resulting visualisation shows an almost perfect separation between the LUAD and LUSC features, demonstrating the ability to distinguish between classes within the extracted features. Figure 6.13b presents a three-dimensional (3D) representation with an extra dimension given by the patients' survival time. The X and Y dimensions are the same as in 6.13a, describing the t-SNE representation of each sample, and the Z gives the overall survival time recorded for the patient, with crosses being deceased subjects and circles alive patients. The visualisation shows the same separation between the classes as in the 2D; however, we can see an intersection between classes in later survival times, which may signify less differentiability between LUAD and LUSC features for longer survival times. Relating to the inner-class distinction between alive and deceases subjects it is possible to see that features that represent alive patients are more clustered in the middle, while features describing dead patients are generally in the outer margins. Although not conclusive, this kind of visualisation allows us to better understand these non-interpretable features and understand why they are able to stratify between patients in survival prediction.

6.5 Discussion

Table 6.6 presents a comparative study of predictive performance among related work for lung cancer survival prediction. The table presents the C-index score according to the histologic subtype predicted and type of approach used, and in the single modality cells, the results for each data type are ordered top to bottom according to the subtitle. To select comparable work, we used the following criteria:

1. The selected works must use data from the same source[1] and should perform survival analysis for the LUAD or LUSC cohort.

2. At least one type of genomic or clinical data should be used.

3. Concordance index must be used to assess predictive performance.

In Oberije et al. [44], a total of 548 NSCLC samples were used; however, no distinction between LUAD and LUSC was made. The authors used only clinical parameters for prediction and used stratified Cox regression for prediction, and the best performing model showed a C-Index of 0.620 and an AIC of 3945. There is no direct comparison with our work that can be done in terms of predictive performance because the authors used a stratified approach and did not divide between

[1]TCGA data portal (https://portal.gdc.cancer.gov/)

	C-Index (mean ± standard deviation)			
	LUAD		**LUSC**	
Methods	***Single***	***Multimodal***	***Single***	***Multimodal***
Oberije et al. (2015)[44] *	0.620		**0.620**	
Zhu et al. (2016)[141] **	0.613 0.660	0.691		
Ching et al. (2018)[20] *		0.630		0.549
Cheerla et al. (2019) [16] ***	0.630 0.613	**0.725**	0.510 0.615	**0.655**
Our Work *	**0.638** 0.616	0.701	0.564 0.568	0.622

Table 6.6: Comparison of predictive accuracy on NSCLC survival outcome using Harrell's C.
* Clinical and Genomic, ** Imaging and Genomic, *** Imaging, Genomic and Clinical

subtypes. However, the model AIC is somewhat higher than our clinical LUAD and LUSC models showing less fitness and a worse trade-off between log-likelihood and the number of parameters.

In Zhu et al. [141], a total of 112 LUAD records with image and RNA-seq samples were used. The authors use a supervised conditional Gaussian graphical model (CGGM) strategy to map pathological images and genetic data, feeding it to a Cox PH model. The imaging modality provided the best individual C-Index of 0.660 over the gene expression data with 0.613. The CGGM integration of both modalities improved the predictive performance to 0.691. The imaging modality performed better on the single modalities than our single models and the genomic performed worse, and on the multimodality, our approach of late fusion outperformed the CGGM, and although this work presents imaging data which deviates from our domain it still presents very interesting results.

In Ching et al. [20], 324 LUAD and 277 LUSC samples were used, including clinical and gene expression data. The authors perform an empirical study to assess different survival models' predictive performance, including Cox-nnet, Cox PH, CoxBoost and Random Forests Survival over ten different TCGA cancer types. Early fusion was used to join the modalities, and for LUAD and LUSC, the best performing model was Cox-nnet with C-index of 0.630 and 0.549. Both of our early fusion and late fusion models outperformed the related work, however, the results for single modalities are not provided for comparison.

In Cheerla et al. [16], the authors use four different data types: 7512 clinical records, 10,914 Whole-Slide Images (WSI), 10,125 RNA-seq samples and 10,198 miRNA samples for 20 different cancer types, and although the results are included individually for LUAD and LUSC there is no mention of the sample sizes. Similarly to our approach, the authors use unsupervised learning for dimensionality reduction and feature representation. Three different architecture were used: a fully-connected network for the clinical data, a CNN for the images and deep highway networks [120] for the miRNA and RNA-seq data. A further pre-processing step was executed to

represent and encode pathologist annotation of Regions of Interest (ROI) on the images. The feature vectors from each modality are aggregated into a single representation, and a final set of 512 features per modality was fed into a cox regression layer for prediction. The authors provide the results for all combinations of modalities: clinical, miRNA, RNA-seq and WSI provided the best performance with C-Index of 0.715 for LUAD and 0.655 for LUSC. However, when using only genomic data and clinical data, the best score was 0.690 for LUAD and 0.620 for LUSC, which our method outperforms. The inclusion of miRNA data seemed to provide a striking gain in performance for the LUSC class, similarly to our work predictive performance is worse for the LUSC class, but there was a significant improvement in accuracy for the LUSC class after the inclusion of miRNA data.

6.6 Summary

This chapter aimed at finding appropriate solutions to integrate multimodal data for lung cancer survival prediction. The task of joining different modalities poses the difficult task of handling heterogeneous data types, which can induce data redundancy if proper techniques are not used. To inspire our solution, we based our approach on some gaps found in the literature regarding the use of early [15], [20], and late fusion [49], [16] techniques to combine different modalities which showed contradictory results, that to our understanding result from the methodology used to join modalities in [15], [20]. In this work, we propose a method of late fusion to join RNA-seq, mutation and clinical modalities in order to validate the results of [49], [16] by using a distinct approach.

Our approach consists of an unsupervised learning stage, where we use two types of network architectures to extract latent representations from each type of data: for the RNA-seq data, we use a variational autoencoder due to its capability of reducing small-sample-sized high-dimensional datasets, without suffering from overfitting and high-variance vanishing gradients [95], due to its optimisation mechanism to minimise the KL divergence; and for the clinical and mutational we used an autoencoder with sparsity constraints and simpler architectures with smaller bottleneck layer to not over over-represent these lower-dimensional modalities. On the supervised learning stage, we merge the latent representations of the modality learners and use a Cox PH regression layer to perform survival prediction. To assess our approach's predictive performance, we designed an empirical study to compare our method's predictive performance with an early fusion model and single-modality learners on survival prediction for the LUAD and LUSC cohorts.

The results showed a statistically significant improvement in predictive performance for the late fusion strategy for LUAD and LUSC cohorts, being the fittest model overall. The early fusion strategy showed substantial gains in performance for the LUAD class, however, there was not a statistical improvement for the LUSC class. Overall predictive performance was best for the LUAD class than for the LUSC class in strategies, which seems to be consistent with the state of the arts' results that use genomic data for survival prediction. The EF model is very similar to the single modality learners as it combines subsets of features from the different modalities

for prediction, which presents the advantage of combining some of the best features from each modality, however, it also incurs the risk of inducing data redundancy as the model might not be apt to capture multimodal relationships. The LF technique presents the unique advantage of reducing each modality's high dimensionality to capture only relevant information from each data type. However, this presents a significant drawback, which is the loss of interpretability of the original features. To validate the late fusion model's effectiveness and analyse the features generated by the unsupervised learning methods, we compare the results of these non-interpretable features with clinical features that were selected as statistically significant in the previous sections for each histologic subtype. We were able to validate the late fusion features' effectiveness by showing that its selected features showed log-rank test p-values smaller than the best single-modality features, allowing for better risk stratification of LUAD LUSC patients, and consequently better overall survival prediction performance.

Chapter 7
Final Remarks and Future Work

Cancer is a genetic disease caused by specific changes to genes that control the way cells function, especially how they grow and divide [68]. The use of computational genomics can impact cancer research in many ways, but maybe the most significant is understanding the underlying causes of cancer. Cancer prediction relates to the process of differentiating cancerous from normal tissue, whereas histologic subtype classification differentiates groups within the same type of cancer, based on specific characteristics of the cancer cells. Interpretable subtype classification approaches allow us to identify gene expression signatures that better differentiate subtypes of cancer, leading to potentially discovering new targetable genes for personalised therapy. The AJCC staging system originated from the need for an accurate, consistent, universal cancer outcome prediction system. Although the analysis of clinical parameters can provide useful prognostic insights, they fail to capture the complex associations needed for accurate prognosis. Combining multimodal genomic data can lead us towards a deeper understanding of what underlying genomic changes affect survivability, which will improve the risk stratification of patients and has the potential to enhance the prognostic and treatment mechanism.

We proposed two main objectives in this work: lung cancer and subtype classification using gene expression data, and predicting cancer patients' survival using clinical and multi-omics data. For the first problem, we used two different approaches: deep learning models that are the standard in state of the art, and gradient boosted trees with the lightGBM implementation, that we provided explainability on using SHAP to retrieve valuable biological insight. The high dimensionality of the gene expression data coupled with the small sample sizes tends to induce overfitting in DL strategies, so we implemented a deep feed-forward network with dropout regularisation, and feature selection was used to remove redundant information. Gradient boosted trees methods are more capable of handling correlated data, and through exclusive feature bundling lightGBM removes redundant features, thereby eliminating the need for apriori feature selection and demonstrating an excellent aptitude of handling the high dimensional data. The DL model obtained an AUC of 0.984 on cancer prediction, and the LGBM model an AUC of 0.971 on subtype classification leading to an improvement over the previous state of the art's results. Furthermore, our interpretable approach allowed us to identify two sets of genes that distinguish cancerous from normal samples and LUAD from LUSC samples. Additionally, we performed hierarchical clustering to identify common-regulated genes and found some patterns of over and under-regulation for genes specific to each histologic subtype which were validated according to relevant literature,

e.g. over-regulation of TP63 in LUSC which is described in The Cancer Genome Atlas Network research [39], and presents as a validation to our results.

For the survival prediction problem, we aimed to find the best solution to integrate gene expression, mutation, and clinical modalities in order to capture the complex data associations required for accurate survival prediction. The task of joining different modalities poses the difficult effort of handling heterogeneous data types, which can induce data redundancy if proper techniques are not used. We proposed a late fusion method with an unsupervised stage, where we use two types of network architectures to extract latent representations from each type of data: a variational autoencoder for the high dimensional gene expression profiles and stacked sparse autoencoders with simpler architectures to extract clinical and mutation features. Following the supervised learning stage, we combined the learned representations of each modality into a deep learning network with a Cox PH regression layer used for survival prediction. To validate our approach, we compared our late fusion model's predictive performance with single-modality Cox PH models and significant gains in performance were observed for the LUAD and LUSC cohort's survival prediction. Our approach presents a drawback which is the loss of interpretability of the original features. To validate the late fusion model's effectiveness and analyse the features generated by the unsupervised learning methods, we compare the results of these non-interpretable features with the fittest single-modality features in the ability to stratify between cancer patients. The results showed a significant performance gain of the late fusion model compared with single modalities, with a concordance index of 0.701 for LUAD survival prediction and 0.622 for LUSC's. The inclusion of extracted features from multiple modalities led to the selection of prognostic factors fitter for survival prediction, which allows for better risk stratification of lung cancer patients and can improve the treatment and prognostic mechanism.

According to our literature review, there are still interesting gaps left to explore in the state of the art of cancer genomics. The LGBM classifier used in this work outperformed state of the art results for subtype classification with an AUC of 0.971 over the previous best score of 0.960, and in future work, we intend to further validate these results by extending it to other cancer types and performing validation to datasets outside of the TCGA scope. Our unsupervised approach for feature extraction of high-dimensional data also demonstrated excellent results for survival prediction, and it could be adapted for cancer subtype classification as it showed the ability to separate between histologic subtypes. Due to the unique architecture of variational autoencoders that are generative models that attempt to describe data generation through probabilistic distributions that can capture an underlying data manifold [77], they can be used as a generative method much like Generative Adversarial Networks (GAN), and it would be interesting to see its effectiveness in generating high-dimensional gene expression profiles.

Overall, this work's efforts showed significant cancer classification results, and presents exciting prospects to have found new, uncovered gene signatures in the literature that can be potential targets for personalised subtype therapy. Our multimodal late fusion model showed great survival prediction results and opened the possibility of extending it to more data types to further enhance cancer patients' risk stratification, which can improve the treatment and prognostic mechanism.

Appendix A

Datasets

Clinical attribute	LUAD TP % missing	LUAD NT % missing	LUSC TP % missing	LUSC NT % missing
days_to_collection	77.4	NA	77.9	NA
oct_embedded	77.4	NA	74.5	NA
shortest_dimension	23.2	0	25.7	0
sample_type_id	0	0	0	0
state	0	0	0	0
is_ffpe	0	0	0	0
tissue_type	0	0	0	0
intermediate_dimension	23.2	0	25.7	0
initial_weight	76.6	94.9	74.5	NA
longest_dimension	23.2	0	25.7	0
submitter_id	0	0	0	0
ajcc_pathologic_m	1.1	1.7	0.8	3.9
ajcc_pathologic_stage	1.5	1.7	0.8	0
age_at_diagnosis	5.8	11.9	2	0
progression_or_recurrence	0	0	0	0
synchronous_malignancy	0	0	0	0
days_to_last_follow_up	23.4	30.5	22.9	35.3
treatments	0	0	0	0
site_of_resection_or_biopsy	0	0	0	0
prior_treatment	0	0	0	0
year_of_diagnosis	1.9	0	3.4	0
ajcc_pathologic_t	0	0	0	0
diagnosis_id	0	0	0	0
days_to_diagnosis	3.5	0	5	0
morphology	0	0	0	0

classification_of_tumor	0	0	0	0
last_known_disease_status	0	0	0	0
tissue_or_organ_of_origin	0	0	0	0
icd_10_code	0	0	0	0
tumor_stage	0	0	0	0
primary_diagnosis	0	0	0	0
ajcc_staging_system_edition	4.1	1.7	19.7	15.7
tumor_grade	0	0	0	0
ajcc_pathologic_n	0.2	0	0	0
prior_malignancy	42.4	30.5	25.7	15.4
alcohol_history	0	0	0	0
exposure_id	0	0	0	0
years_smoked	61.4	72.9	55.8	54.9
cigarettes_per_day	31.4	44.1	15.3	19.6
pack_years_smoked	31.4	44.1	15.3	19.6
demographic_id	0	0	0	0
year_of_death	77	62.7	69.7	52.9
race	0	0	0	0
ethnicity	0	0	0	0
year_of_birth	3.5	0	5	0
gender	0	0	0	0
vital_status	0	0	0	0
days_to_birth	5.8	11.9	2	0
age_at_index	3.5	0	1.8	0
days_to_death	64.9	55.9	57.8	37.3
bcr_patient_barcode	0	0	0	0
releasable	0	0	0	0
project_id	0	0	0	0
primary_site	0	0	0	0
disease_type	0	0	0	0
released	0	0	0	0
name	0	0	0	0
paper_patient	54	100	64.3	100
paper_Sex	54	100	64.3	100
paper_Age.at.diagnosis	54	100	65	100
paper_T.stage	54	100	64.3	100
paper_N.stage	54	100	64.3	100
paper_Tumor.stage	54	100	NA	NA

paper_Smoking.Status	54	100	65.5	100
paper_Survival	54.2	100	NA	NA
paper_Transversion.High.Low	54	100	NA	NA
paper_Nonsilent.Mutations	54	100	NA	NA
paper_Nonsilent.Mutations.per.Mb	54	100	NA	NA
paper_Oncogene.Negative.or.Positive.Groups	54	100	NA	NA
paper_Fusions	98	100	NA	NA
paper_expression_subtype	54	100	NA	NA
paper_chromosome.affected.by.chromothripsis	99	100	NA	NA
paper_iCluster.Group	54	100	NA	NA
paper_CIMP.methylation.signature	65	100	NA	NA
paper_MTOR.mechanism.of.mTOR.pathway	64	100	NA	NA
paper_Ploidy.ABSOLUTE.calls	54	100	NA	NA
paper_Purity.ABSOLUTE.calls	547	100	NA	NA
paper_M.stage	NA	NA	65.1	100
paper_Pack.years	NA	NA	70.9	100
paper_Nonsilent.Mutatios	NA	NA	64.3	100
paper_Nonsilent.Mutatios.per.Mb	NA	NA	64.3	100
paper_Selected.Mutation.Summary	NA	NA	65.9	100
paper_High.Level.Amplifications	NA	NA	81.3	100
paper_Homozygous.Deletions	NA	NA	85.5 100	100
paper_Expression.Subtype	NA	NA	64.3	100

Table A.1: Missing values of raw data for LUAD and LUSC primary tumour (TP) and normal tissue (NT) clinical datasets.

Appendix B

Survival Analysis

covariate	C	HR(C)	$\sigma(C)$	up 95%	low 95%	z	p
prior_malignancy.Yes	0.7009	2.0155	0.149	1.505	2.6992	4.7034	0
ajcc_pathologic_stage	0.3982	1.4891	0.1491	1.1116	1.9947	3.6696	0.0001
oct_embedded.True	0.0639	1.066	0.0161	1.0328	1.1003	3.9617	0.0001
tissue_or_organ_of_origin.nos	0.333	1.3951	0.1124	1.1192	1.7391	2.9615	0.0031
ajcc_pathologic_t	0.4091	1.5055	0.1393	1.1457	1.9782	2.9365	0.0033
paper_Oncogene.Group.Oncogene Negative	0.2026	1.2246	0.0744	1.0585	1.4169	2.7235	0.0065
days_to_birth	-0.3992	0.6709	0.1502	0.4998	0.9005	-2.6582	0.0079
morphology.8260/3	-0.3362	0.7145	0.1306	0.5532	0.9229	-2.5748	0.0100
treatment_or_therapy0.yes	0.059	1.0608	0.023	1.0141	1.1098	2.5667	0.0103
paper_CIMP.methylation.signature.high	0.1833	1.2011	0.0745	1.038	1.3899	2.4602	0.0139
initial_weight	0.3184	1.375	0.1321	1.0613	1.7812	2.4106	0.0159
gender.male	0.0874	1.0913	0.0375	1.0141	1.1745	2.3329	0.0197
icd_10_code.C34.9	-0.3178	0.7278	0.1437	0.5492	0.9645	-2.2119	0.0270
is_ffpe.True	-0.3347	0.7155	0.1567	0.5263	0.9728	-2.1358	0.0327
race.asian	-0.4982	0.6076	0.2462	0.375	0.9844	-2.0238	0.0430
treatment_or_therapy1.yes	0.1129	1.1195	0.0558	1.0036	1.2488	2.0241	0.0430
morphology.8253/3	-0.4819	0.6176	0.2407	0.3854	0.9899	-2.0021	0.0453
morphology.8550/3	-0.2607	0.7705	0.1318	0.5951	0.9976	-1.9779	0.0479

Table B.1: Clinical variables with p-value<0.05 for the LUAD survival analysis.

covariate	C	HR(C)	$\sigma(C)$	up 95%	low 95%	z	p
tissue_origin.middle lobe	-0.3675	0.6925	0.2004	0.4675	1.0257	-1.8334	0
ajcc_pathologic_m	-0.8279	0.437	0.469	0.1743	1.0955	-1.7654	0.0002
tissue_origin.nos	0.4839	1.6224	0.2755	0.9455	2.7839	1.7564	0.0490
ajcc_pathologic_n	-0.8279	0.437	0.469	0.1743	1.0955	-1.7654	0.0010
ajcc_pathologic_stage	-0.8279	0.437	0.469	0.1743	1.0955	-1.7654	0.0330

Table B.2: Clinical variables with p-value<0.05 clinical variables for LUSC survival analysis.

covariate	C	HR(C)	$\sigma(C)$	up 95%	low 95%	z	p
mut.PIK3CA	0.7164	2.047	0.2097	1.3572	3.0874	3.4165	0.0006
mTOR_pathway_PI3K_mut	0.059	1.0608	0.023	1.0141	1.1098	2.5667	0.0013
chromothripsis.chr19	0.0362	1.0369	0.0126	1.0117	1.0628	2.8818	0.0041
mTOR_pathway.unaligned	0.3982	1.4891	0.1491	1.1116	1.9947	2.6696	0.0076

Table B.3: Mutational variables with p-value<0.05 for LUAD survival analysis.

covariate	C	HR(C)	$\sigma(C)$	up 95%	low 95%	z	p
mut.TP53 p.E271K	0.0599	1.0618	0.0293	1.0024	1.1246	2.0431	0.0310
mut.PIK3CA p.E545K	0.7105	2.0351	0.3482	1.0284	4.0271	2.0404	0.0413
amplifications.CCND1	-0.4982	0.6076	0.2462	0.375	0.9844	-2.0238	0.0280
mut.ASCL4 p.P14T	-0.2607	0.7705	0.1318	0.5951	0.9976	-1.9779	0.0479

Table B.4: Mutational variables with p-value<0.05 clinical variables for LUSC survival analysis.

covariate	C	HR(C)	$\sigma(C)$	up 95%	low 95%	p
78	0.0592	1.061	1.0333	1.0895	4.3886	0
32	-0.0379	0.9628	0.9435	0.9825	-3.6663	0
38	0.0287	1.0291	1.0125	1.046	3.4531	0.0006
69	0.0784	1.0816	1.0311	1.1346	3.2126	0.0013
3	0.0366	1.0372	1.0131	1.062	3.0437	0.0023
62	-0.0482	0.9529	0.923	0.9838	-2.9657	0.0030
75	-0.017	0.9831	0.972	0.9944	-2.9335	0.0034
96	-0.0618	0.9401	0.9012	0.9806	-2.8697	0.0041
74	0.0653	1.0675	1.02	1.1172	2.8111	0.0049
71	0.0379	1.0386	1.0097	1.0684	2.6282	0.0086
8	-0.0526	0.9488	0.9119	0.9871	-2.6038	0.0092
41	0.0411	1.0419	1.0096	1.0754	2.5502	0.0108
48	-0.0505	0.9507	0.9141	0.9888	-2.5242	0.0116
116	0.056	1.0576	1.0126	1.1046	2.5227	0.0116
30	-0.0242	0.976	0.9577	0.9947	-2.51	0.0121
117	-0.0541	0.9473	0.9079	0.9885	-2.493	0.0127
0	-0.0287	0.9717	0.9496	0.9944	-2.4339	0.0149
99	-0.059	0.9427	0.8985	0.989	-2.411	0.0159
118	-0.0755	0.9273	0.872	0.986	-2.4103	0.0159
36	-0.0445	0.9565	0.9225	0.9918	-2.4059	0.0161
51	-0.041	0.9598	0.9278	0.9928	-2.3762	0.0175
20	-0.0224	0.9778	0.9597	0.9963	-2.3427	0.0191
1	0.057	1.0586	1.0092	1.1106	2.3337	0.0196
81	0.0428	1.0437	1.006	1.0828	2.2769	0.0228
85	-0.0406	0.9602	0.927	0.9945	-2.2653	0.0235
95	0.0245	1.0249	1.003	1.0471	2.2361	0.0253
10	0.048	1.0492	1.004	1.0964	2.1385	0.0325
43	0.0602	1.0621	1.005	1.1224	2.1379	0.0325
40	0.0325	1.0331	1.0025	1.0646	2.1224	0.0338
106	0.0418	1.0427	1.0032	1.0837	2.1216	0.0339
119	-0.0376	0.9631	0.9294	0.998	-2.0722	0.0382
22	-0.0297	0.9707	0.9426	0.9997	-1.9831	0.0474
11	-0.0232	0.9771	0.9549	0.9998	-1.977	0.0480

Table B.5: Late fusion covariates with p-value<0.05 for LUAD histologic subtype.

covariate	**C**	**HR(C)**	$\sigma(C)$	**up 95%**	**low 95%**	**p**
52	-0.1006	1.1043	0.0645	0.7969	1.0262	0
22	-0.0445	0.9565	0.0185	0.9225	0.9918	0
29	-0.041	0.9598	0.0173	0.9278	0.9928	0.0175
20	-0.0224	0.9778	0.0096	0.9597	0.9963	0.0191
36	0.057	1.0586	0.0244	1.0092	1.1106	0.0196
44	0.0428	1.0437	0.0188	1.006	1.0828	0.0228
33	-0.0406	0.9602	0.0179	0.927	0.9945	0.0235
95	0.0245	1.0249	0.011	1.003	1.0471	0.0253
10	0.048	1.0492	0.0225	1.004	1.0964	0.0285
43	0.0602	1.0621	0.0282	1.005	1.1224	0.0295
40	0.0325	1.0331	0.0153	1.0025	1.0646	0.0308
108	0.0418	1.0427	0.0197	1.0032	1.0837	0.0319
119	-0.0376	0.9631	0.0182	0.9294	0.998	0.0382
22	-0.0297	0.9707	0.015	0.9426	0.9997	0.0474
11	-0.0232	0.9771	0.0117	0.9549	0.9998	0.0480
117	-0.0541	0.9473	0.9079	0.9885	0.987	0.0482

Table B.6: Late fusion covariates with p-value<0.05 for LUSC histologic subtype.

References

[1] Emmanuel Adetiba and Oludayo O. Olugbara. Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. *Scientific World Journal*, 2015, 2015.

[2] 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. *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 1748–1752, 2019.

[3] Douglas G Altman and Hall Crc. Medical research. *Public Health*, 51(C):199–200, 1937.

[4] Pierre Baldi. Autoencoders, unsupervised learning and deep architectures. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, page 37–50. JMLR.org, 2011.

[5] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. 03 Accessed: February 1, 2021.

[6] Yael Ben-Haim and Elad Tom-Tov. A streaming parallel decision tree algorithm. *Journal of Machine Learning Research*, 11:849–872, 02 2010.

[7] Samy Bengio and Yoshua Bengio. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks*, 11(3):550–557, 2000.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

[9] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, (January):153–160, 2007.

[10] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, 30(7):1145–1159, July 1997.

[11] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[12] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification And Regression Trees*. 10 2017.

[13] Harry B. Burke, Philip H. Goodman, David B. Rosen, Donald E. Henson, John N. Weinstein, Frank E. Harrell Jr., Jeffrey R. Marks, David P. Winchester, and David G. Bostwick. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857–862, 1997.

[14] Zhipeng Cai, Lizhe Xu, Yi Shi, Mohammad R. Salavatipour, Randy Goebel, and Guohui Lin. Using gene clustering to identify discriminatory genes with higher classification accuracy. *Proceedings - Sixth IEEE Symposium on BioInformatics and BioEngineering, BIBE 2006*, (January):235–242, 2006.

[15] Kumardeep Chaudhary, Olivier B. Poirion, Liangqun Lu, and Lana X. Garmire. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259, 2018.

[16] Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pan-cancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 07 2019.

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[18] Yen Chen Chen, Wan Chi Ke, and Hung Wen Chiu. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in Biology and Medicine*, 48(1):1–7, 2014.

[19] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 02 2016.

[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, apr 2018.

[21] François Chollet et al. Keras. keras.io, 2015.

[22] Nikolay Chumerin and Marc M. Van Hulle. Comparison of two feature extraction methods based on maximization of mutual information. *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, MLSP 2006*, (May 2014):343–348, 2006.

[23] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, 2013.

[24] T G Clark, M J Bradburn, S B Love, and D G Altman. Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238, 2003.

[25] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[26] Adele Cutler, D Richard Cutler, and John R Stevens. High-Dimensional Data Analysis in Cancer Research. *High-Dimensional Data Analysis in Cancer Research*, (May 2014), 2009.

[27] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, I:886–893, 2005.

[28] Padideh Danaee, Reza Ghaeini, and David A. Hendrix. A deep learning approach for cancer detection and relevant gene identification. *Pacific Symposium on Biocomputing*, 0(212679):219–229, 2017.

[29] Joseph M. De Guia, Madhavi Devaraj, and Carson K. Leung. DeepGX: Deep learning using gene expression for cancer classification. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, pages 913–920, 2019.

[30] Thomas G Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2):139–157, 2000.

[31] Altekruse et al. SEER Cancer Statistics Review 1975-2007 National Cancer Institute SEER Cancer Statistics Review 1975-2007 National Cancer Institute. *Cancer*, pages 1975–2007, 2010.

[32] Barbier et al. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 424–432. Curran Associates, Inc., 2016.

[33] Cameron Davidson-Pilon et al. Lifelines @ `lifelines.readthedocs.io`, Accessed: February 1, 2021.

[34] Colaprico et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71–e71, 12 2015.

[35] Collisson et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*, 511(7511):543–550, 2014.

[36] Dunn et al. Next generation sequencing methods for diagnosis of epilepsy syndromes. *Frontiers in Genetics*, 9:20, 2018.

[37] Griffiths AJF et al. *An Introduction to Genetic Analysis.* W.H.Freeman & Co Ltd, 7th edition, Accessed: February 1, 2021.

[38] Guolin Ke et al. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. Curran Associates, Inc., 2017.

[39] Hammerman et Al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

[40] Kao et al. IGDB.NSCLC: integrated genomic database of non-small cell lung cancer. *Nucleic Acids Research*, 40(D1):D972–D977, 12 2011.

[41] Kawase et al. Differences Between Squamous Cell Carcinoma and Adenocarcinoma of the Lung: Are Adenocarcinoma and Squamous Cell Carcinoma Prognostically Equal? *Japanese Journal of Clinical Oncology*, 42(3):189–195, 12 2011.

[42] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems. `tensorflow.org`, 2015.

[43] Mounir et al. New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex. *PLOS Computational Biology*, 15(3):1–18, 03 2019.

[44] Oberije et al. A validated prediction model for overall survival from stage III non-small cell lung cancer: Toward survival prediction for individual patients. *International Journal of Radiation Oncology Biology Physics*, 92(4):935–944, 2015.

[45] O'Leary et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–45, jan 2016.

[46] Postmus et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 28(Supplement 4):iv1–iv21, 2017.

[47] Uhlén et al. Tissue-based map of the human proteome. *Science*, 347(6220), 2015.

[48] Villalobos et al. Lung Cancer Biomarkers. *Hematology/oncology clinics of North America*, 31(1):13–29, feb 2017.

[49] Wang et al. A Cancer Survival Prediction Method Based on Graph Convolutional Network. *IEEE Transactions on Nanobioscience*, 19(1):117–126, Accessed: February 1, 2021.

[50] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001.

[51] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256, 2010.

[52] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274–278, Oct 2010. 21455458[pmid].

[53] Gregory R Grant, Elisabetta Manduchi, and Christian J Stoeckert. Analysis and management of microarray gene expression data. *Current protocols in molecular biology*, Chapter 19:Unit 19.6, January 2007.

[54] Aurlien Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 1st edition, 2017.

[55] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.

[56] Zeyad Hailat, Artem Komarichev, and Xue Wen Chen. *Deep Semi-Supervised Learning*, volume 2018-August. 2018.

[57] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer Review Douglas. *Cell*, 100(7):57–70, 2000.

[58] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[59] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Prognostic/Clinical Prediction Models: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Tutorials in Biostatistics, Statistical Methods in Clinical Studies*, 1:223–249, 2005.

[60] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. *Aug, Springer*, 1, 01 2001.

[61] Haibo He and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 1st edition, 2013.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[63] Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

[64] Roy S. Herbst, John V. Heymach, and Scott M. Lippman. Lung cancer. *New England Journal of Medicine*, 359(13):1367–1380, 2008. PMID: 18815398.

[65] Zena M. Hira and Duncan F. Gillies. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, 2015(1):1–13, jun 2015.

[66] Zhi Huang, Travis S Johnson, Zhi Han, Bryan Helm, Sha Cao, Chi Zhang, Paul Salama, Maher Rizkalla, Christina Y Yu, Jun Cheng, Shunian Xiang, Xiaohui Zhan, Jie Zhang, and Kun Huang. Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations. *BMC Medical Genomics*, 13(5):41, Accessed: February 1, 2021.

[67] Kentaro Inamura. Lung cancer: understanding its molecular pathology and the 2015 wHO classification. *Frontiers in Oncology*, 7(AUG):1–7, 2017.

[68] National Cancer Institute. About cancer: Causes and prevention @ cancer.gov, Accessed: February 1, 2021.

[69] National Cancer Institute. Cancer topics: What is cancer @ cancer.gov, Accessed: February 1, 2021.

[70] National Cancer Institute. Encyclopedia: TCGA barcode @ docs.gdc.cancer.gov, Accessed: February 1, 2021.

[71] Ruhl J, Adamo M, and Dickie L. Section V: Stage of disease at diagnosis. *Seer*, (February), 2016.

[72] Shruti Jadon. A survey of loss functions for semantic segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct 2020.

[73] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[74] Andreas Janecek, Wilfried N Wn Gansterer, Michael Demel, and Gerhard Ecker. On the Relationship Between Feature Selection and Classification Accuracy. *Fsdm*, 4(May 2014):90–105, 2008.

[75] Prasoon Joshi, Seokho Jeong, and Taesung Park. Cancer Subtype Classification based on Superlayered Neural Network. pages 1988–1992, Accessed: February 1, 2021.

[76] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*, pages 372–378, 2014.

[77] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[78] D.G. Kleinbaum. Survival analysis, a self-learning text. *Biometrical Journal*, 40(1):107–108, 1998.

[79] Rasmus Krempel, Pranav Kulkarni, Annie Yim, Ulrich Lang, Bianca Habermann, and Peter Frommolt. Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB). *BMC Bioinformatics*, 19(1):1–10, 2018.

[80] The Lancet. The global burden of disease: cancer @ ourworldindata.org, Accessed: February 1, 2021.

[81] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[82] Michael K K Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, 30(12):i121–9, jun 2014.

[83] Bin Li, Yu-Qi Meng, Zheng Li, Ci Yin, Jun-Ping Lin, Duo-Jie Zhu, and Shao-Bo Zhang. Mir-629-3p-induced downregulation of sftpc promotes cell proliferation and predicts poor survival in lung adenocarcinoma. *Artificial Cells, Nanomedicine, and Biotechnology*, 47(1):3286–3296, 2019. PMID: 31379200.

[84] Chao Li and Meng Zhang. Deep Learning in Pan Cancer Early Detection Based on Gene Expression, 2018.

[85] Chengzhang Li and Jiucheng Xu. Feature selection with the Fisher score followed by the Maximal Clique Centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma. *Scientific Reports*, 9(1):17283, 2019.

[86] Zhuoshi Li, Tao Guo, Lei Fang, Nan Li, Xiaochao Wang, Peng Wang, Shilei Zhao, Fengzhou Li, Yanwei Cui, Xin Shu, Lei Zhao, Jinxiu Li, and Chundong Gu. MACC1 overexpression in carcinoma-associated fibroblasts induces the invasion of lung adenocarcinoma cells via paracrine signaling. *Int J Oncol*, 54(4):1367–1375, 2019.

[87] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4):928–937, 2015.

[88] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[89] D. Y. Lin and L. J. Wei. The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

[90] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to Combine Modalities in Multimodal Deep Learning. 2018.

[91] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[92] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[93] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2019.

[94] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[95] Mohammad Sultan Mahmud, Joshua Zhexue Huang, and Xianghua Fu. Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. *International Journal of Computational Intelligence and Applications*, 19(01):2050002, Accessed: February 1, 2021.

[96] Janine Meienberg, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. Clinical sequencing: is WGS the better WES? *Human genetics*, 135(3):359–362, mar 2016.

[97] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9(1):1–12, 2018.

[98] Hiroshi Motoda and Huan Liu. Feature selection, extraction and construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan)*, 5:67–72, 01 2002.

[99] Steven G. Nadler, Douglas Tritschler, Omar K. Haffar, James Blake, A. Gregory Bruce, and Jeffrey S. Cleaveland. Differential expression and sequence-specific interaction of karyopherin α with nuclear localization sequences. *Journal of Biological Chemistry*, 272(7):4310–4315, 1997.

[100] NCCN. Mestastic Lung Cancer 2019. *European Journal of Cancer*, 51, Supple:S610–S649, 2015.

[101] NCCN Guidelines for Patients. Lung Cancer: Early and Locally-Advanced Non-Small Cell Lung Cancer. pages 1–56, 2019.

[102] Fernando Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014–.

[103] Morihito Okada, Noriaki Tsubota, Masahiro Yoshimura, Yoshifumi Miyamoto, and Reiko Nakai. Evaluation of tmn classification for lung carcinoma with ipsilateral intrapulmonary metastasis. *The Annals of Thoracic Surgery*, 68(2):326 – 330, 1999.

[104] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner. `github.com/keras-team/keras-tuner`, 2019.

[105] OncoKB. Oncogene list @ `oncokb.org`, Accessed: February 1, 2021.

[106] World Health Organization. The icd-10 classification of mental and behavioural disorders : diagnostic criteria for research, 1993.

[107] Fatih Ozsolak et al. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 2011.

[108] Sunhee Park and David J. Hendry. Reassessing schoenfeld residual tests of proportional hazards in political science event history analyses. *American Journal of Political Science*, 59(4):1072–1087, 2015.

[109] Juan Ramos-González, Daniel López-Sánchez, Jose A. Castellanos-Garzón, Juan F. de Paz, and Juan M. Corchado. A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. *Computers in Biology and Medicine*, 86:98–106, 2017.

[110] Greg Ridgeway. Generalized Boosted Models: A guide to the gbm package. *Compute*, 1(4):1–12, 2007.

[111] Cecil Robinson and Randall Schumacker. Interaction effects: centering, variance inflation factor, and interpretation issues. *Multiple Linear Regression Viewpoints*, 35(1):6–11, 2009.

[112] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009.

[113] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[114] David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA, 1986.

[115] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[116] Kamla Kant Shukla, Praveen Sharma, and Sanjeev Misra. *Molecular Diagnostics in Cancer Patients*. 01 2019.

[117] Tiago C Silva, Antonio Colaprico, Catharina Olsen, Fulvio D'Angelo, Gianluca Bontempi, Michele Ceccarelli, and Houtan Noushmehr. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research*, 5:1542, 2016.

[118] Johannes Smolander, Alexey Stupnikov, Galina Glazko, Matthias Dehmer, and Frank Emmert-Streib. Comparing biological information contained in mRNA and non-coding RNAs for classification of lung cancer patients. *BMC Cancer*, 19(1):1–15, 2019.

[119] Jasper Snoek et al. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, page 2951–2959, 2012.

[120] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks, 2015.

[121] Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1209–1216. Curran Associates, Inc., 2008.

[122] Mats J Stensrud and Miguel A Hernán. Why Test for Proportional Hazards? *JAMA*, 323(14):1401–1402, apr Accessed: February 1, 2021.

[123] Muhammad Aliyu Sulaiman and Jane Labadin. Feature selection based on mutual information. 27(8):1–6, 2015.

[124] Yingshuai Sun, Sitao Zhu, Kailong Ma, Weiqing Liu, Yao Yue, Gang Hu, Huifang Lu, and Wenbin Chen. Identification of 12 cancer types through genome deep learning. *Scientific Reports*, 9(1):1–9, 2019.

[125] Ying Tao, Jianrong Li, Carol Friedman, and Yves A. Lussier. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, 23(13):i529–i538, 2010.

[126] et al. Timo Ojala. Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions . 4(January 1996):2002, 1994.

[127] Jakub M. Tomczak. Prediction of breast cancer recurrence using classification restricted boltzmann machine with dropping, 2013.

[128] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[129] Pascal Vincent and Hugo Larochelle. Extracting and Composing Robust Features with Denoising.pdf. pages 1096–1103, 2008.

[130] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science*, 340(6127):1546–1558, 2013.

[131] Dehua Wang, Yang Zhang, and Yi Zhao. LightGBM: An effective miRNA classification method in breast cancer patients. *ACM International Conference Proceeding Series*, pages 7–11, 2017.

[132] Zhong Wang et al. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[133] Gregory P Way and Casey S Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23:80–91, 2018.

[134] WHO. Cancer Factsheets @ `who.int/mediacentre`, Accessed: February 1, 2021.

[135] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PloS one*, 15(6):e0233678, Accessed: February 1, 2021.

[136] 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.

[137] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using rna-seq data. *Computer Methods and Programs in Biomedicine*, 166:99 – 105, 2018.

[138] Xiucai 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, Accessed: February 1, 2021.

[139] Yuchen Yuan, Yi Shi, Changyang Li, Jinman Kim, Weidong Cai, Zeguang Han, and David Dagan Feng. Deepgene: An advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics*, 17(Suppl 17), 2016.

[140] Huan Zhang, Si Si, and Cho-Jui Hsieh. Gpu-acceleration for large-scale tree boosting, 2017.

[141] X. Zhu, J. Yao, G. Xiao, Y. Xie, J. Rodriguez-Canales, E. R. Parra, C. Behrens, I. I. Wistuba, and Junzhou Huang. Imaging-genetic data mapping for clinical outcome prediction via supervised conditional gaussian graphical model. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 455–459, 2016.

[142] Andréanne N Zizzo, Lauren Erdman, Brian M Feldman, and Anna Goldenberg. Similarity Network Fusion: A Novel Application to Making Clinical Diagnoses. *Rheumatic diseases clinics of North America*, 44(2):285–293, may 2018.