TÉCNICO LISBOA

Differential regulatory response of Hodgkin's Lymphoma according to PET result

André Miguel Paralta Patrício

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Rui Miguel Carrasqueiro Henriques
Dr. Rafael Sousa Costa

Examination Committee

Chairperson: Prof. Mário Jorge Costa Gaspar da Silva
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Pedro Tiago Gonçalves Monteiro

July 2021

Acknowledgments

Começo por expressar a mais sincera gratidão pelo apoio dos meus orientadores, Rui Henriques e Rafael Costa. Ajudaram-me a desenvolver a minha paixão pela investigação, estiveram disponíveis sempre que precisei de esclarecimentos ou de discutir ideias, e sem eles não conseguiria ter desenvolvido uma dissertação da qual me orgulho. Obrigado, ajudaram a moldar o profissional que serei daqui para a frente.

A todos os meus colegas de curso que me apoiaram ao longo deste processo através de trocas de ideias, conselhos, ou simples convivências que me ajudaram a chegar ao fim deste percurso, um obrigado.

Agradeço aos meus pais por terem sempre garantido que tinha todas as condições necessárias para ser quem quisesse e chegar onde desejasse.

Por fim, agradeço à Carolina, uma fonte infindável de apoio sem a qual não conseguiria ser quem sou. Obrigado.

Abstract

Hodgkin's Lymphoma (HL) is a type of lymphoma, a class of cancers of the lymphatic system. Despite the advancements in multiagent chemotherapy in the past years, up to 10% of HL cases are refractory to treatment and, after remission, patients experience an elevated risk of death from all causes. These complications are dependent on the prescribed treatment, and therefore an increase in the prognostic accuracy of HL can help improve these outcomes. Interim Fluorodeoxyglucose-Positron Emission Tomography (FDG-PET) is the primary indicator of patient response to treatment, however, it is an intrusive and expensive medical exam. In this context, we present a methodology capable of predicting the result of an interim FDG-PET exam after two courses of Adriamycin, Bleomycin, Vinblastine and Dacarbazine (ABVD) chemotherapy, through the analysis of gene expression profiles using state-of-the-art machine learning algorithms. The presented approach combines dimensionality reduction procedures and hyperparameter optimization of various elected classifiers to study the state-of-the-art predictability of refractory response to ABVD treatment. In addition to this, we propose a data transformation procedure for mapping the original data space into a more discriminative one through the use of biclustering. This methodology produces results superior to the ones previously obtained in this dataset. A further study of gene regulatory relations is performed to obtain novel knowledge about the molecular mechanisms associated with HL and discrimination of treatment response. The approach presented is especially relevant due to the low incidence of this cancer, which results in a lack of works that take advantage of novel machine learning tools and increasingly cheap and fast high-throughput technologies.

Keywords

Hodgkin's Lymphoma; Cancer; Machine Learning; Data Science; Prognostic; Gene expression profile; FDG-PET; Biclustering; Discriminative patterns;

Resumo

O linfoma de Hodgkin é um tipo de linfoma, um cancro do sistema linfático. Apesar dos avanços em quimioterapia multi-agente nos últimos anos, até 10% dos casos de linfoma de Hodgkin são refratários e, depois de entrada em remissão, os pacientes sofrem um risco elevado de morte por várias causas. Estas complicações são dependentes do tratamento prescrito, e por isso mesmo, um aumento na precisão do prognóstico pode ajudar a melhorar estes resultados. Tomografia por Emissão de Positrões usando fluorodeoxyglucose (FDG-PET) intermédia é o principal indicador de resposta ao tratamento, no entanto, é intrusivo e caro. Neste contexto, apresentamos uma metodologia capaz de prever o resultado de uma FDG-PET intermédia após dois ciclos de quimioterapia Adriamycin, Bleomycin, Vinblastine e Dacarbazine (ABVD), através da análise de perfis de expressão de genes usando algoritmos estado da arte de aprendizam automática. A abordagem apresentada combina técnicas de redução de dimensionalidade e optimização de hiperparameterização de vários classificadores de forma a estudar o estado da arte da predictibilidade de resposta refratória ao tratamento. Adicionalmente, propomos uma transformação dos dados para mapear o espaço dos dados original para um mais discriminativo através de biclustering. Esta metodologia produz resultados superiores aos obtidos anteriormente nestes dados. Um estudo adicional das relações regulatórias entre genes é levada a cabo para obter novo conhecimento acerca dos mecanismos moleculares associados com o linfoma de Hodgkin e discriminação da resposta ao tratamento. A abordagem apresentada é especialmente relevante devido à baixa incidência deste cancro, o que resulta numa falta de trabalhos que tirem proveito de ferramentas de aprendizagem automática e de tecnologias de sequenciamento de grande escala cada vez mais económicas e rápidas.

Palavras Chave

Linfoma de Hodgkin; Cancro; Aprendizagem automática; Ciência de Dados; Prognóstico; Perfil de expressão de genes; FDG-PET; Biclustering; Padrões discriminativos;

Contents

List of Figures

List of Tables

Acronyms

ABVD Adriamycin, Bleomycin, Vinblastine and Dacarbazine

AUC Area Under the Curve

BEACOPPesc Bleomycin, Etoposide, Adriamycin, Cyclophosphamide, Oncovin, Pro-carbazine and Prednisone in escalated dose

BicPAMS Biclustering based on PAttern Mining Software

cDNA complementary DNA

CHL Classical Hodgkin Lymphoma

CV Cross-Validation

DLBCL Diffuse Large B-Cell Lymphoma

DNA Deoxyribonucleic Acid

DT Decision Tree

EBV Epstein-Barr Virus

FC Fold Change

FDG Fluorodeoxyglucose

FDG-PET Fluorodeoxyglucose-Positron Emission Tomography

FDR False Discovery Rate

FFPE Formalin Fixed Paraffin Embedded

FIM Frequent Itemset Mining

FN False Negatives

FP	False Positives
FPR	False Positive Rate
GB	Gradient Boosting
GO	Gene Ontology
HIV	Human Immunodeficiency Virus
HL	Hodgkin's Lymphoma
HRS	Hodgkin and Reed-Sternberg
HSC	Hematopoietic Stem Cell
IL-17	Interleukin-17
iPET	interim Fluorodeoxyglucose-Positron Emission Tomography
iPET2	interim PET after two courses of ABVD chemotherapy treatment
KDE	Kernel Density Estimation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbors
LMR	Lymphocyte-to-Monocyte Ratio
mRNA	messenger RNA
MI	Mutual Information
ML	Machine Learning
NB	Naive Bayes
NGS	Next Generation Sequencing
OMIM	Online Mendelian Inheritance in Man
RF	Random Forest
RFE	Recursive Feature Elimination
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

SVM-SMOTE Support Vector Machine Synthetic Minority Oversampling Technique

SVM-RFE Support Vector Machine Recursive Feature Elimination

TN True Negatives

TP True Positives

TPR True Positive Rate

WRST Wilcoxon Rank Sum Test

1

Introduction

Contents

Hodgkin's Lymphoma (HL) is a type of blood cancer that originates in the lymphatic system, more precisely in lymphocytes, a particular type of white blood cells, with patients being commonly diagnosed in their 20s and 30s. Primary symptoms include enlarged lymph nodes in the neck, armpit or groin. In 2018, HL represented 0.4% of all new tumors (79990 new cases) and 0.3% of all cancer deaths (26167 deaths) worldwide [19]. This chapter presents the motivation for this work regarding HL, the approached research problems, the resultant contributions and the outline for the remainder of the document.

1.1 Motivation

Survival of Hodgkin's Lymphoma patients has significantly improved over the past years as a result of the development of multiagent chemotherapy and more effective radiotherapy. Still, about 5–10% of cases are refractory to initial treatment and 10–30% will relapse despite having achieved initial complete remission [6]. Even when this is not the case, after initial remission, patients experience an elevated risk of death from several causes [3]. Most studied causes are cardiotoxicity diseases like myocardial infarction and congestive heart failure [2], and secondary cancers [37]. As reported by Hoppe [68], these following diseases are often treatment-related, and adjustments in these treatments can help reduce the long-term excess risk of death from complications after therapy.

The current prognosis for HL is largely based on the International Prognostic Score (IPS) [61], which predicts for 5-year freedom from progression based on seven risk features: male gender, age $\geq$ 45 years, stage IV by Ann Arbor Classification [23], hemoglobin $<$ 105 g/L, white cell count (WCC) $\geq$ 15000/mm3, lymphocyte count $<$ 600/mm3 or $<$ 8% WCC, and Serum Albumin $<$ 40 g/L. Moccia et al. [101] concluded that although this scoring remains the standard prognostic for patients with advanced stage HL, it does not identify with certainty low or high risk groups, and recommends the use of molecular markers and/or fluorodeoxyglucose Positron Emission Tomography (FDG-PET) scanning for this purpose. Hutchings et al. [71] too, observed that interim FDG-PET (iPET) after two courses of chemotherapy is a strong and independent predictor of progression-free survival in HL, and that in regression analysis, early iPET was stronger than established prognostic factors. Gallamini et al. [51] corroborated these findings by showing the high predictive value of an iPET done after two cycles of chemotherapy. Furthermore, iPET to adapt treatment is recommended by most of the available guidelines [41].

Despite the proven relevance of iPET for HL prognostic, this clinical exam is: i) intrusive, with the need to inject a radioactive tracer; ii) expensive, estimated at €1020 per exam [141]; and iii) impossible to perform in remote locations due to requiring large machinery.

Since HL is a relatively rare cancer (2.86 cases per 100,000 persons annually [70]), it has not been exhaustively studied. In addition, some of the state-of-the-art approaches successfully applied to study more common cancers, such as prognosis and risk prediction using machine learning, have not yet been

comprehensively employed.

The work presented in this dissertation aims to achieve a more precise prediction of patient's response to chemotherapy treatment, and consequently, better adjust the levels of toxicity that the patient undergoes, resulting in reduced risk of treatment-related comorbidities. This is complemented by the study of the molecular mechanisms underlying the development of Hodgkin's Lymphoma and the patient's response to treatment.

1.2 Problem Description

The studied research problem consists in the analysis of gene expression values of HL patients with the objective of better understanding the regulatory mechanisms and predictability of a patient's response to a specific chemotherapy regimen. To achieve this we resort to gene expression profiles obtained from Formalin Fixed Paraffin Embedded (FFPE) diagnostic tumor samples using NanoString's[1] nCounter platform [52]. The level at which a gene is expressed allows us to better understand the state of a given biological system and its reactions to stimuli, making it a fundamental piece of information when trying to understand and make predictions about a certain disease such as Hodgkin's Lymphoma. Even though the expression value of a single gene can be informative, the focus here is placed on their necessary modular interactions for enacting regulatory processes. The large amount of possible co-expression associations challenges the characterization and discovery of discriminative patterns of a certain condition.

Given this, the two main research problems studied in this dissertation are:

- Accurate prediction of Hodgkin's Lymphoma patient's response to ABVD chemotherapy regimen through the analysis of gene expression profiles;
- Identification of the molecular mechanisms that discriminate a patient's refractory response to ABVD chemotherapy treatment.

1.3 Contributions

In order to address the aforementioned problems, we first conducted a thorough optimization and assessment of preprocessing and machine learning techniques to develop a predictor that can, at the moment of diagnosis, classify patients' future interim PET after two courses of ABVD chemotherapy treatment (iPET2) according to treatment response, whether responsive (negative) or non-responsive (positive). State-of-the-art predictors show preliminary satisfactory results and are further analyzed in

[1] https://www.nanostring.com (accessed June 28, 2021)

order to draw conclusions about which genes are important in its decisions. Complementary to the application of state-of-the-art predictors, we propose the use of biclustering techniques to transform the feature space into one consisting of features given by discriminative gene expression patterns and analyze the impact that this transformation has on classification performance. The resulting patterns were studied and compared against established gene sets related to Hodgkin's Lymphoma and multiple predictive models showed an increased classification performance in the transformed feature space. Finally, the gene sets identified by both our predictors and feature selection methods were evaluated resorting to functional enrichment analysis, resulting in the identification of putative gene modules important in how a patient reacts to treatment. These gene modules are involved in various biological mechanisms previously identified as important in the development of multiple cancers, but not in Hodgkin's Lymphoma, leading us to hypothesize that these same mechanisms are discriminative in the response to ABVD chemotherapy in HL patient's.

The contributions provided by this dissertation can therefore be divided into two major groups: i) creation of a methodology capable of anticipating a patient's response to treatment by predicting the result of a PET exam after two cycles of chemotherapy; and ii) identification and analysis of gene interactions at the transcriptomic level relevant to the progression of Hodgkin's Lymphoma and its response to treatment. Along the first group, major contributions are:

1. Identification of important molecular features for performing an accurate classification of a patient's response to treatment through the combination of multiple dimensionality reduction procedures;

2. Assessment of the most adequate classification models for this task, complemented by the optimization of its parameters;

3. Study of the impact of a pattern-based space transformation through biclustering in the classification performance of state-of-the-art predictive models in transcriptomic data;

4. Analysis of the defining characteristics of a patient that lead to worse predictability of reaction to treatment.

The second group encompasses the following contributions:

1. Corroboration of the influence of multiple conditions and mechanisms in the development of Hodgkin's Lymphoma;

2. Identification of putative gene modules discriminative of reaction to treatment;

3. Association of biological mechanisms previously identified as influential in other cancers with Hodgkin's Lymphoma.

Since available technologies for gene expression profiling (such as microarray and RNA-seq) are increasingly cheaper and faster, and biopsies with FFPE conservation are the default procedure to confirm HL diagnosis, our predictive modeling contributions can be easily translated into the medical practice. We hope that this work can help reduce treatment-related mortality by identifying those who need immediate stronger treatment and those who will react well to the standard chemotherapy regimen. In addition to this, the biological analysis of the gene expression profiles carried in this work can contribute to the existing knowledge through the identification of molecular features and associated biological processes that better discriminate the different responses to treatment.

1.4 Document Outline

The remainder structure of this dissertation is organized as follows: Chapter 2 provides background on Hodgkin's Lymphoma, cancer-related transcriptomics and machine learning techniques to be employed. Chapter 3 follows with a compilation of related work. Subsequently, chapter 4 presents an exploratory analysis of the studied dataset. The solution employed to tackle the target research problems is presented in chapter 5, followed by the obtained predictive results in chapter 6. Chapter 7 presents the biological analysis of our results, and chapter 8 concludes this dissertation with our final remarks.

2

Background

Contents

The chapter here presented introduces essential background on Precision Oncology and Machine Learning. Precision Oncology consists in utilizing knowledge about the molecular profile of a tumor to predict disease phenotype, clinical outcome or treatment response, and using this information to tailor treatment to each individual. The introduction of gene expression profiling technologies in this context provides physicians with large amounts of data that need to be analyzed and processed in order to be useful. This is where Machine Learning proves to be a powerful tool, enabling the transformation of seemingly incomprehensible data into actionable results.

We start by offering a more detailed introduction to Hodgkin's Lymphoma and the already mentioned FDG-PET exam, followed by a brief explanation of both cancer-related transcriptomics and Hodgkin's Lymphoma primary molecular features. Following this, we cover the essentials of Classification, Clustering, Dimensionality Reduction and Evaluation Metrics to answer the target problem.

2.1 Precision Oncology

2.1.1 Hodgkin's Lymphoma Disease and Treatment

Hodgkin's Lymphoma (HL) can be divided into two groups: i) Nodular Lymphocyte-Predominant Hodgkin Lymphoma, which accounts for only 5% of cases; and ii) Classical Hodgkin Lymphoma (CHL), the type that will be studied in this work, and therefore, the one referred to when using the denomination Hodgkin's Lymphoma (HL). CHL can be further divided into four subtypes, based on morphology and abundance of its cancer cells and the surrounding micro-environment. These are mentioned in decreasing order of incidence: Nodular Sclerosis, Mixed Cellularity, Lymphocyte-Rich and Lymphocyte-Depleted.

The staging of HL is done using the Lugano classification [30], an adaption of the older Ann Arbor system [23]. This system assigns the stage based on the location of the cancer:

- Stage I: exclusively in a lymph node, a lymphoid organ or a part of a non-lymphoid organ;
- Stage II: in two or more lymph node areas both above or below the diaphragm or in a lymph node and nearby organ;
- Stage III: in lymph node areas on both sides of the diaphragm or above the diaphragm and in the spleen;
- Stage IV: widely spread in one non-lymphoid organ.

In addition to this classification, the letter *E* may be added if a non-lymphoid organ is affected. The patient's stage can be further divided in *A* or *B*, with the letter *B* representing the existence of *B* symptoms: inadvertently loss of more than 10% of body weight, fever or drenching night sweats.

Treatment for HL varies according to the disease's stage but normally consists of multiple cycles of chemotherapy followed by radiotherapy. Limited-stage HL treatment includes two cycles of Adriamycin, Bleomycin, Vinblastine and Dacarbazine (ABVD) chemotherapy and fractionated radiotherapy at 20 Gy, with Gy representing the ionizing radiation dose unit defined by the International System of Units (SI) [106]. For intermediate stage patients, four cycles of ABVD and fractionated radiotherapy at 30 Gy is widely considered. In cases of patients with less than 60 years, a more aggressive treatment can be considered, with two cycles of Bleomycin, Etoposide, Adriamycin, Cyclophosphamide, Oncovin, Procarbazine and Prednisone in escalated dose (BEACOPPesc), two cycles of ABVD and radiotherapy at 30 Gy. Finally, advanced stage can be treated with either six cycles of ABVD or four to six cycles of BEACOPPesc, with localized radiotherapy being optional [137].

2.1.2 Interim Positron Emission Tomography

A Positron Emission Tomography (PET) is an imaging test that uses a radioactive tracer to measure the metabolic activity of cells in body tissues. In the oncology domain specifically, this tracer is created by applying a radioactive atom to glucose, forming the radionuclide Fluorodeoxyglucose (FDG). Since cancer cells have considerably higher metabolic rates than normal cells, the glucose analog FDG will accumulate in regions where cancer is present, showing up as a bright spot on PET scans due to the photons emitted by the radioactive component [8]. This is useful to make a diagnosis, analyze treatment effectiveness or check for cancer recurrences. As already stated, an interim FDG-PET is highly recommended to assess if the current treatment of a HL patient is being effective in treating the disease [1, 71, 76, 101].

2.1.3 Cancer Transcriptomics

Cancer is a class of diseases characterized by the abnormal growth, replication and survivability of cells. The normal life cycle of the cell consists in growing and dividing according to the needs of our body and and dying when it is no longer necessary. Cancer cells will deviate from its normal behaviour, and instead, accelerate the replication process and refuse to die, forming an accumulation of cells known as a tumor. This irregular behavior is induced by either direct alterations to the genome, or to the way the genome is used [144].

It is the aforementioned genome that encodes the information necessary to produce, among other molecules, proteins, the core of all cell function. It is the selective production of proteins that defines the behavior of a cell, its reactions to the changing environment, and, under certain conditions, can also lead to the abnormal behaviour characteristic of cancer. The instructions on how to create these fundamental molecules take the form of Deoxyribonucleic Acid, also known as DNA. The DNA encodes

proteins by using combinations of four nitrogenous bases, namely, Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). A sequence of DNA that encodes a specific protein is termed a gene, and the first step in the creation of an encoded protein is the transcription of the corresponding gene, as is explained ahead. When the transcription of a gene occurs, it is said that the gene was expressed.

Cancer can then have two possible causes: i) alterations in a gene's sequence of nitrogenous bases; and ii) alterations in how much a given gene is transformed into the corresponding protein. The focus of this work will be on the later cause (ii).

There are two main types of genes related to cancer: proto-oncogenes and tumor suppressor genes. The former, encode for molecules responsible for inducing cell growth, proliferation and survivability. The latter, are responsible for DNA reparation and cell growth regulation. It is the over-expression of proto-oncogenes and the under-expression of tumor suppressor genes that often leads to the formation of tumors. To understand how a gene can be over or under-expressed, we must first know in what consists the complete process of converting a gene's encoding into a protein. This process can be, in a general manner, divided into two phases: *transcription* and *translation*.

Transcription consists in the creation of an Ribonucleic Acid (RNA) copy of a specific part of the DNA, a gene. RNA is similar to DNA and is used to transport the encodings present in DNA to the site where they will be translated into proteins. One of the differences between these molecules is the replacing of Thymine (T) by Uracil (U) in RNA. DNA and RNA both have the property of complementarity, and so, each of its bases only binds with a specific base. C only binds with G (and vice-versa), and A only binds with T, or U in the RNA case (and vice-versa). So, when it is necessary to create a transcript of a DNA code sequence for protein creation purposes, an RNA molecule with the inverse bases of this sequence will be produced by transcription of the DNA. This specific type of RNA is called messenger RNA (mRNA). The whole set of RNA molecules produced by transcription forms the transcriptome, and its study is denominated as transcriptomics [36]. The analysis of the transcriptome is used to, among other things, measure gene expression levels and better understand the behaviour of certain cells in normal conditions, or in response to disease progression or drug intake.

The **translation** step happens in the ribosome, the organelle of the cell responsible for receiving a sequence of nitrogenous bases in the form of mRNA and using it to assembly a protein. The mRNA received contains triplets of bases that each encodes an amino acid, the structural units of proteins. By translating each triplet, denominated as codon, to an amino acid, the ribosome will be able to assembly a polypeptide chain that will fold itself originating a protein.

2.1.4 Gene Expression Profiling

In order to understand the transcriptome of a cell or tissue, we must be able to accurately measure the rate at which various genes are being expressed, which can be approximately achieved by measuring the

quantity of mRNA. Even if proteins are the molecules that will ultimately define the function of a cell, when analyzing cell behaviour this approach yields specific properties of interest against the measurement of the actual proteins, namely: i) proteins are molecules that can be interchanged between cells, and so, when analyzing the protein content of a cell, we can not distinguish which proteins were produced by the cell under analysis; and ii) mRNA, contrary to proteins, decays rapidly, and so, a measure of its values gives us an accurate snapshot of how the cell is behaving at that moment.

A variety of technologies have been introduced for the purpose of gene expression profiling, and the following will be addressed: *Microarrays*, *NanoString's nCounter* and *Next Generation Sequencing (NGS)*.

Microarrays [63] use the propensity that unpaired nitrogenous bases have to bind to its complements in order to measure mRNA quantity. First, various *probes*, pieces of synthetic DNA that correspond to a sequence located in a gene of interest, are synthesized. When a probe contacts with a solution of genetic material, it will bind to the correspondent mRNA in a process called *hybridization*. This process is susceptible to errors, and so, repeated probes for the same gene are used and the genetic material to measure is increased by *amplification*. The amplification step consists in the creation of DNA strands complementary to the RNA to be analyzed, called complementary DNA (cDNA), that are then used to create more replicas of the original RNA. This can, however, introduce bias and make the experiment less reproducible. The previously produced probes are then attached to a surface at specific, known locations, forming what is known as a microarray chip. The genetic material to be analysed with this tool is previously fluorescently tagged, and then released in the microarray. The target mRNA binds with the probes, and the fluorescent-intensity values are then measured by an optical scanner. Because the probes are attached at specific locations, it is possible to know which values correspond to each gene.

The **nCounter** procedure [52], created by NanoString[1], introduces a novel digital color-coded barcode technology. For this method is necessary to create two probes that will be bound together. The *capture probe*, similar to the one previously described, and the *reporter probe* that possesses a unique colored barcode that identifies the gene targeted by the capture probe. These probes can be directly deployed in the genetic material, without the need for the amplification step, reading the target mRNA directly. After the mRNA has bound with the probes, loose mRNA and probes are removed, and the remaining are immobilized and aligned on a cartridge. This cartridge is read by a fluorescence microscope that will count the different barcodes, resulting in a count value for each targeted gene. One of the most prominent advantages of nCounter is its ability to perform accurate gene expression on Formalin Fixed Paraffin Embedded (FFPE) tissue, as demonstrated by Reis et al. [114]. FFPE refers to the default form of preservation of cancer biopsy specimens. The process necessary for this type of preservation leads to chemical modifications and degradation of RNA, which together with the limited amounts of sample

[1] https://www.nanostring.com (accessed June 28, 2021)

usually available, make gene expression analysis of this tissues a difficult task.

The introduction of **Next Generation Sequencing** (NGS) methods revolutionized gene expression analysis by substituting the former hybridization-based approach with more sophisticated sequencing techniques [139]. NGS introduced the following advantages: i) it does not require prior knowledge about gene sequences since it does not use probes. This enables whole transcriptome sequencing and analysis of species for which genomes are not yet available; ii) since cDNA sequences generated in NGS can be mapped to targeted regions on the genome, it is easier to remove experimental noise; iii) its results are quantifiable, contrary to microarrays whose values are relative to other signals of the array; and iv) results have lower technical variation and higher levels of reproducibility.

As previously mentioned, the data analyzed in this work was obtained using the NanoString's nCounter platform, an appropriate choice since the sampled tissue is FFPE. Despite the differences in the various referred methods, the results obtained are similar [104, 129, 147], and so, the whole pipeline constructed in this dissertation can be applied to data produced by different gene expression profiling methods.

2.1.5 Molecular Profile of Hodgkin's Lymphoma

A lymphoma is formed by an accumulation of lymphocytes that, due to misregulatory behavior, grow, proliferate and/or survive at higher rates than expected, often starting in the lymph nodes. The two main types of lymphocytes present in our body are *B cells*, responsible for the production of antibodies, and *T cells*, which can differentiate into several distinct types, including: i) *killer cells* (CD8+) responsible for causing immune-mediated cell death in virus-infected cells; ii) *helper cells* (CD4+) that determine how other parts of the immune system (such as regulatory B cells) respond to threats; and iii) *regulatory cells* that as the name implies, regulate immune responses to prevent autoimmune responses. These two types of cells both have an important paper in the immune system. B cells are fundamental for humoral immunity, a type of immunity that relies on macromolecules such as antibodies to respond against extracellular organisms. T cells on the other hand are involved in cell-mediated immunity, activating cells such as phagocytes and killer T cells through cytokines, small proteins used in cell signaling, leading them to destroy intracellular invaders such as viruses and bacteria.

The causes of Hodgkin's Lymphoma are not known, although some risk factors have been identified. The contraction of the Epstein-Barr Virus (EBV), causer of infectious mononucleosis, was shown to increase fourfold the chances of suffering from HL [67], with EBV detected on about 40% of HL patients. Infection by Human Immunodeficiency Virus (HIV) also raises the chances of HL, with Biggar et al. [16] reporting a tenfold increase. Other conditions related to a weaker immune system such as organ transplantation and autoimmune conditions have also been correlated to a higher incidence of this cancer [85].

Hodgkin's Lymphoma is characterized by the presence of Hodgkin and Reed-Sternberg (HRS) cells

in the lymph nodes. These cancer cells are derived from germinal center B cells [83] but largely lose their B cell phenotype and morphologically differ from these by its larger size and possible multiple nucleoli. Unlike other cancers, the cancer cells in HL form a minority of the tumor and are surrounded by a reactive inflammatory mixture of non-malignant reactive cells, such as the white blood cells: lymphocytes, macrophages and eosinophils [131]. The HRS cells can show severe deregulated activation of numerous signaling pathways, including the PI3K/AKT, JAK/STAT, MAPK/ERK and NOTCH1 ones [81]. One of the more prominent deregulated pathways is the NF-κB. The HRS cells show a recurrent activity of this pathway, contrary to normal germinal center B cells that only transiently activate it. There are multiple NF-κB transcription factors, proteins that can regulate genes with functions such as cell survival, proliferation, cell adhesion and differentiation. These factors are *REL*, *RELA* (*p65*), *RELB*, *p50* (NF-κB1), and *p52* (NF-κB2), all of them expressed by HRS cells. Inhibition of both canonical and non-canonical NF-κB activity has been shown to cause reduced proliferation and increased apoptosis in HL cell lines [10, 66]. These undesirable proteins can be activated by the aforementioned Epstein-Barr Virus, as explained in detail by Weniger and Küppers [146].

2.2 Machine Learning

2.2.1 Classification

Classification can be described as the task of correctly attributing a class c to a given multivariate data observation $\mathbf{x}_{new}$ from a set of pairs $(\mathbf{x}_i, c_i)$, $i \in 1...n$ where $c_i \in \Sigma$.

A classifier will then be defined as a model M that receives an observation $\mathbf{x}_i$ and returns a prediction of its class, designated as $\hat{z}$. This is, $M(\mathbf{x}_i) = \hat{z}$. The different dimensions of the data observation, $Y = \{y_1, ..., y_m\}$, are designated as the independent variables, and the variable z is the dependent variable/class. The value associated with observation $\mathbf{x}_i$ at variable y_j is represented by a_{ij}.

In order to correctly make predictions, the model M must be trained with a set of data observations, designated as the training set $X = \mathbf{x}_1, ..., \mathbf{x}_n$, to approximate a function to the real distribution of the data. Its prediction capability can then be measured by its performance on a separate set of observations, the test set.

Naive Bayes (NB) [90] is the simplest Bayesian classification approach grounded on the Bayes' Theorem,

$$P(c_k \mid \mathbf{x}_i) = \frac{P(\mathbf{x}_i \mid c_k) \cdot P(c_k)}{P(\mathbf{x}_i)} , \tag{2.1}$$

to calculate the posterior probability of a target $c_k \in z$ given an observation $\mathbf{x}_i$, $P(c_k \mid \mathbf{x}_i)$. The class with

maximum probability given $\mathbf{x}_i$ is chosen as output for the classification,

$$\hat{z} = \underset{k \in \{1,\ldots,K\}}{argmax} \; P(c_k \mid \mathbf{x}_i) \; , \tag{2.2}$$

where K is the number of classes. The posterior probability $P(c_k \mid \mathbf{x}_i)$ depends on the prior probability of c_k $P(c_k)$, the evidence $P(\mathbf{x}_i)$, and the likelihood probability $P(\mathbf{x}_i \mid c_k)$. Since $P(\mathbf{x}_i)$ is constant for all values of c_k, we can ignore it. This leaves us with $P(c_k)$, a simple calculation, and $P(\mathbf{x}_i \mid c_k)$, that can be more computationally expensive or hard to reliably estimate with small amounts of data. To overcome this problem, the Naive Bayes algorithm assumes that all features $y_j \in Y$ are independent, and therefore

$$P(c_k \mid \mathbf{x}_i) = P(c_k) \prod_{j=1}^{m} P(a_{ij} \mid c_k) \; . \tag{2.3}$$

Despite its over-simplified assumptions, the NB algorithm can work well in real-world scenarios and is often used as a baseline to compare with other models.

The **K-Nearest Neighbors** (KNN) algorithm [46] differs from other supervised classification methods that learn functions from the available data. Instead, the "fitting" of the model is just the loading of all the observations of the dataset. Given a new observation, this algorithm will determine its K nearest neighbors, calculating the distance between the new sample and the loaded ones using the different features as dimensions and choosing the K ones that are closer. The class returned is the one that represents the majority of these neighbors. The number K and the distance function are hyperparameters that can be altered according to the domain and desired results.

Associative classifiers originate from the integration of association mining and classification. It represents the process of discovering association rules in a given dataset, which consists in finding an event A that is in some way associated with event B. The generated rule $A \Rightarrow B$ indicates us that event A discriminates B, where A and B can be, for example, a frequent set of values for a subset of features $J \subseteq Y$. By adapting this proceeding to only have the classification target in the right side of the rule, we can generate a set of rules that can be employed by a classifier in order to make predictions.

Decision Trees (DT) [112] represent a simple case of an associative model, with the rules structured as a tree. By utilizing a training dataset, they define the set of rules that will be sequentially applied to a new observation until a class is decided. Each decision represents a node, a split that leads to two (or more) new nodes. This process goes on until a leaf, a terminal node, is reached, corresponding to the class to be returned. The rules for each node are defined as the divisions of the dataset yielding better discriminative ability. The discriminative ability can be measured by various information theory functions, such as entropy or the Gini impurity. Other important hyperparameters can be tuned in order to, for example, define how deep the tree should go before reaching a leaf.

Ensemble algorithms consist of a combination of multiple independent models, that together, seek to obtain a better performance than the individual ones would. This family of algorithms is based on the idea that the combination of single, simple classifiers can result in more robust results since these classifiers reinforce each other's correct decisions and cancel their errors. Ensemble principles can be divided into two groups:

Bootstrap aggregating (bagging): bootstrapping [40] is used on the training set to obtain multiple samples that act as an independent representation of the dataset's true distribution. Each independent classifier will be trained on one of these samples. The output of all classifiers is averaged when it is time to make predictions. Since, according to bootstrapping properties, each sample is approximately representative of the true underlying distribution, the average of the classifiers' results will still be an adequate guess but with a lower variance due to the models being trained on different datasets.

Boosting: boosting, unlike bagging, is an iterative process. Instead of independent models trained in parallel, we have models that try to iteratively overcome the difficulties of each other. The process generally respects the following steps:

1. each training record is given an equal weight;
2. a classifier M_l is trained on the training set to return a c_k for each $\mathbf{x}_i$;
3. a coefficient α_l is assigned to M_l based on its performance;
4. the training record's weights are updated, increasing its value if the classifier M_l could not classify the record correctly and vice-versa;
5. the process from 2. to 4. is repeated until all the desired models have been trained.

The weights associated with each training record are used to indicate to the next model M_{l+1} which are the records that it should give more attention to, this is, the most important records to classify correctly. This process results in a set of models that compensate for each other's weaknesses, reducing the bias of the final model but becoming more prone to overfitting. The coefficients α_l are used to quantify the predictive capacity of each model M_l, giving more or less importance to its outcomes when it is time to make a prediction. The prediction of this ensemble model is calculated as a weighted vote of all models,

$$\hat{z} = argmax_{c_k} \sum_{l=1}^{L} \alpha_l \times M_l(\mathbf{x}_i) \ . \tag{2.4}$$

Random Forest (RF) [20] is an ensemble of Decision Trees (DT). It uses the previously described bagging method but with a slight difference. In addition to the sampling of records for each DT, a

sampling of features is also performed, making every single classifier even more uncorrelated to each other and the final classifier more robust to missing data.

Gradient Boosting (GB) [49] is an ensemble of multiple weak models, usually, decision trees. The difference between this algorithm and Random Forests is that GB uses the boosting method combined with gradient descent. This process consists in the iterative addition of new weak models h_l to form a final model M. The initial model, h_0 consists of constant values derived from the distribution of the z values. From then, the process pursues the following steps:

1. compute the residuals r_i for each $\mathbf{x}_i$'s prediction by M_{l-1}, using a defined loss function L : $r_i = -\frac{\partial L(c_i, M_{l-1}(\mathbf{x}_i))}{\partial M_{l-1}(\mathbf{x}_i))}$;

2. train a weak model h_l to predict the r_i values of each record $\mathbf{x}_i$;

3. using a defined learning rate η update the final model: $M_l = M_{l-1} + \eta\, h_l$.

This is repeated until a predefined number of weak classifiers as been trained and added to the final model M. The probability of a new observation $\mathbf{x}_i$ belonging to the positive class can be obtained by the sum of all weak models h_l predictions multiplied by the learning rate η,

$$\hat{z} = \sum_{l=0}^{L} \eta \times h_l(\mathbf{x}_i) \ . \tag{2.5}$$

It is worth to mention the implementation of this algorithm by the `XGBoost`[2] [27] library as it is purposefully selected in this work. This implementation aims to provide a scalable, portable and distributed Gradient Boosting. It ensures execution speed by parallelization of tree construction and distributed computing while providing increased performance by automatic handling of missing data, performing L1 and L2 regularization, among other features.

Support Vector Machines (SVMs) [18] classify data using a decision boundary $D(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i + b$, with $\mathbf{w}$ and b being the model's weights and $\mathbf{x}_i$ the input vector. This boundary is the hyperplane that best divides the data into its corresponding classes. In SVMs, the best division of the data is quantified by the *maximum margin*, corresponding to the minimum distance between the decision boundary and the closest observations of each class. The greater this distance is the more generalizable the model is, so the objective is to maximize it while having the restriction of correctly separating the classes. The Lagrangian multiplier method is used to solve this problem, resulting in a Lagrange multiplier α associated with each observation, with non-zero values only for the points of each class that are closer to the maximum-margin hyperplane. These are called the *support vectors*, and through a linear combination

[2] `https://xgboost.readthedocs.io` (accessed June 28, 2021)

they will define the weights **w** and b of the decision boundary,

$$\mathbf{w} = \sum_i \alpha_i \, c_i \, \mathbf{x}_i \qquad \text{and} \qquad b = c_i - \mathbf{w} \cdot \mathbf{x}_i \ , \tag{2.6}$$

with c_i taking the values of -1 and 1 for the negative and positive class respectively. Multiclass problems can also be tackled using adequate adaptations. In the case of not linearly separable data, a *kernel trick* can be employed, where the data is projected to a higher dimensional space, allowing it to be correctly separated by a line.

2.2.2 Clustering

Clustering consists in the aggregation of similar observations in groups denominated as clusters. Similarity can be defined in multiple ways according to the input data characteristics and desired results. Depending on the type of clustering, different heuristics can be used to perform the groupings.

In this work, Hierarchical Clustering [145] is considered for an initial exploratory analysis given its inherent properties and normative application in this domain [125, 126]. This class of clustering approaches starts by defining each observation as a cluster and progressively merges the clusters that are closer to each other, according to a defined similarity measure. To do this it is necessary to specify distance between clusters through a hyperparameter known as the linkage method. This process can be repeated until only one cluster remains, or stopped when a certain quantity is achieved. The example given describes an agglomerative hierarchical clustering, but the inverse process can be done, starting with one cluster and progressively splitting them, resulting in a divisive hierarchical clustering.

2.2.3 Biclustering

If it is desired to not be limited to the clustering of samples, we can also simultaneously cluster the samples' features by performing **Biclustering** [60]. This subspace clustering technique looks for subspaces in the input data that are both homogeneous and statistically significant. In gene expression data, for example, this can be beneficial since only a subset of the genes analyzed contribute to significant differentiation between samples. In this case, biclustering will identify a subset of samples where groups of genes show coherent behavior.

Bicluster approaches based on pattern mining use Frequent Itemset Mining (FIM), a method used for finding frequent association rules, to mine homogeneous and significant biclusters. In order to allow for efficient space exploration and guarantee the statistical significance of the found associations, the possible biclusters are restricted by interestingness metrics such as support or lift. Given two events A and B, the support of an association rule $A \Rightarrow B$ corresponds to its overall frequency in the dataset,

while the lift represents the ability of A to predict B as

$$lift = \frac{P(A \cap B)}{P(A)\,P(B)}\ . \tag{2.7}$$

This equation shows us that a lift of 1 means that A and B are independent and that the higher the deviation from 1, the higher is the correlation between event A and event B. By restricting B to the target variable of a given classification problem, we can find rules that discriminate between classes.

The homogeneity of the found biclusters is characterized by their coherence, quality and structure. The coherence defines the correlation between the values of a bicluster, it can be constant if the values are equal, additive if they vary by a fixed set of values, among others. Since the coherence constraints can rarely be fulfilled in real-world data, the biclusters must allow some degree of noise, with the specific type and amount of noise being defined as the quality of the bicluster. The structure will correspond to the number and shape of the found biclusters. Figure 2.1 shows an example illustrating the concepts of statistical significance and discriminative ability in this context.

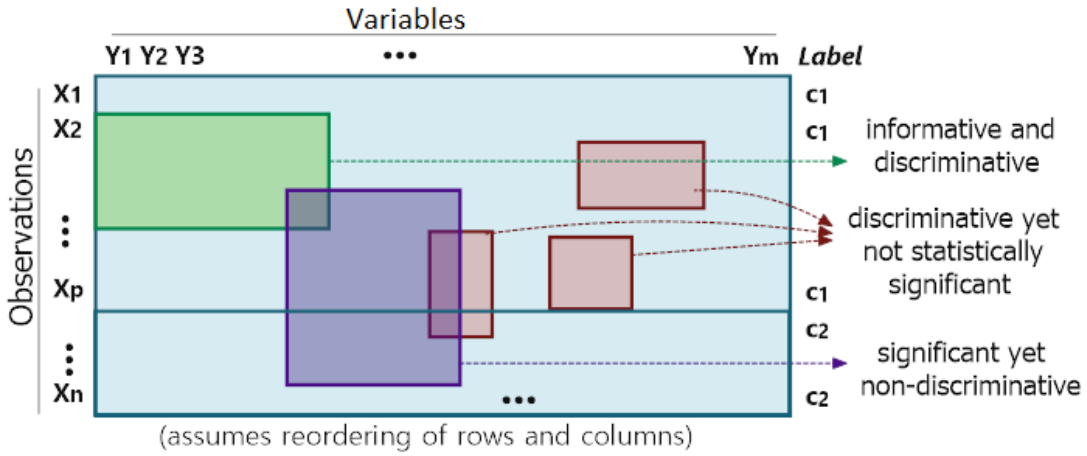

Figure 2.1: Labeled dataset with biclusters characterized according to their statistical significance and discriminative ability (sourced from Henriques and Madeira [65])

2.2.4 Dimensionality Reduction

Feature selection offers a way of reducing dimensionality by identifying the more relevant features of a dataset for a given task. This leads to a reduction in the complexity of the data and the resulting trained model. Unlike feature extraction, this method does not alter the original representation of the variables, which is useful when the goal is not only to improve prediction capability but also to identify and analyze the most important features, as is the case with this work. Feature selection methodologies are divided into three groups: *Filters*, *Wrappers* and *Embedded* [79].

Filters select relevant features by looking at the intrinsic properties of the data, making them the only methods that are model-independent. This property can result in features that are not optimal for classification by a given model, due to a disconnect between the objective function for the filter method and what the model requires [79]. On the other end, filters are able to make an unbiased selection of variables that are related to the target feature. Since it is only necessary to perform the selection once for various models, they tend to be much faster than the alternatives. The taxonomy introduced by Saeys et al. [116] divides Filters in Univariate and Multivariate. The former ignores feature dependencies and the latter does not, making it slower and less scalable but able to identify redundancies. Inside the Univariate techniques, another division can be done, between parametric methods, that assume a given distribution from which the data was sampled, and the non-parametric methods, which do not. Mutual Information and the Wilcoxon Rank Sum Test, both explained ahead, are examples of univariate non-parametric methods.

Wrappers make use of classifiers performance to evaluate if a given subset of features is favorable for prediction. A separate algorithm is used to iteratively feed the classifier with better subsets. The wrapper approach has the benefits of taking into account feature dependencies in the data and the classifier that will be used. This can lead to a greater predictive power but comes at the cost of a greater risk of overfitting and computation time.

Embedded methods, as the name implies, have the feature selection process embedded in the model training process, which removes the need for external algorithms. The features are selected, for example, according to weights assigned to them in regularization models or partitions made in tree-based models. This class of feature selection presents a less computationally expensive approach than Wrappers while maintaining the advantage of directly using a model, ensuring a better predictive power. The SVM-RFE and Random Forest-based algorithms introduced ahead are part of this group of feature selection methods.

It is suggested by Díaz-Uriarte and De Andres [34], for the purpose of gene selection, the use of an *embedded feature selection method based on Random Forests*. This algorithm measures feature importance as the decrease in accuracy of the RF when values of said feature are permuted randomly. In order to select an adequate subset of features, first, the importance of said features is calculated as stated early. Random forests are then iteratively fitted, removing at each iteration a fraction of the least important features still present and evaluating the classifier by its out-of-bag error, a metric specific for bagging algorithms. It is worth clarifying that the feature importance is not recalculated at each step, in order to avoid severe overfitting. This process continues until all features have been removed, with the ideal subset corresponding to the one with the smallest number of genes whose error is within u standard errors of the minimum error obtained between all RFs. The authors of this algorithm test the parameter u with values 0 and 1.

Recursive Feature Elimination (RFE) is a wrapper method for feature selection. It begins by training a classifier in the given data, with the whole set of features, optimizing its own weights **w**. Then, a ranking criterion is calculated using its optimized weights and an external algorithm that evaluates the effects that removing a feature has in the objective function of the classifier. The features with the lowest score are discarded, creating a new subset of features. This process is repeated until a specified number of subsets has been analyzed, and finally outputs a feature subset ranking according to the classifier performance.

The **Support Vector Machine Recursive Feature Elimination** (SVM-RFE) algorithm was introduced by Guyon et al. [55] in the context of gene selection in cancer classification. It constitutes an adaption of the already mentioned RFE feature selection method, that instead of using an external function to compute the ranking of features, uses the magnitude of the weights of a Support Vector Machine. It begins by training an SVM in the whole feature set, computing **w** as shown before, and calculating the ranking r_j for each variable y_j as $(\mathbf{w}_j)^2$. It will then, just like in the original RFE algorithm, use the values r_j to remove the individual or group of features with the lowest ranking. This process repeats itself until no features are left, returning the final ranking of features. The authors further show that this algorithm is not prone to overfitting due to its greedy nature.

The **Mutual Information** (MI) [120] of two random variables can be described as a measure of the mutual dependence between them, or more intuitively, how much information a variable has about the other. To mathematically define MI, the concept of Entropy is needed. Having a variable y_j with values in $\mathcal{Y}_j$ and probability density function f, its entropy will measure the uncertainty of the variable and can be calculated as

$$H(y_j) = -\int_{\mathcal{Y}_j} f(y_j) \log f(y_j)\, dy_j \ . \tag{2.8}$$

If we are looking to compare two distributions, then we can use Relative Entropy. The relative entropy is a measure of the distance between two distributions. The Relative Entropy $D(f||g)$ represents the inefficiency of assuming that the probability density function is g when the true one is f,

$$D(f||g) = \int_{\mathcal{Y}_j} f(y_j) \log \frac{f(y_j)}{g(y_j)}\, dy_j \ . \tag{2.9}$$

Finally, MI is calculated as the relative entropy between the joint probability mass function and the marginal probability mass function,

$$\begin{aligned} MI(y_1; y_2) &= D(p(v_1, v_2)||p(v_1)p(v_2)) \\ &= \int_{\mathcal{Y}_2}\int_{\mathcal{Y}_1} p(y_1, y_2) \log \frac{p(y_1, y_2)}{p(y_1)p(y_2)}\, dy_1\, dy_2 \ . \end{aligned} \tag{2.10}$$

With this, it becomes intuitive that the higher the MI, the higher the difference between the joint probability

and marginal mass functions, which means, the more dependent the variables are. Remember that the definition of independent variables is $p(y_1, y_2) = p(y_1)p(y_2)$, which would result in the MI score being 0.

The **Wilcoxon Rank Sum Test** [97] (WRST) is a nonparametric test, and so, it does not make any assumption about the distribution of the data. It has the purpose of testing if two independent samples are likely derived from the same distribution. This can be seen as testing for the null and research hypotheses, which are, respectively:

- H_0: The two populations are equal;
- H_1: The two populations are not equal.

This test involves the calculation of the statistic U, whose distribution under the null hypothesis is known, enabling us to associate a p-value to the result. Having two populations of size n_1 and n_2, the values of both populations are ranked, with the lowest value receiving a 1, and the highest receiving $n_1 + n_2$. The values R_1 and R_2 are calculated as the sum of the rankings of populations 1 and 2, respectively. Next, for each population i, the statistic U_i is calculated as

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - R_i \ . \tag{2.11}$$

The final statistic U is the smallest value between the previously calculated U_i. In order to correctly perform this test, the following requirements must be fulfilled: i) samples must be independent; and ii) values must be ordinal or continuous.

2.2.5 Evaluation Metrics

In order to correctly evaluate predictive models and fully understand its capabilities for eventual real world use, it is essential to define appropriate performance metrics. Most of the metrics used in binary machine learning problems can be defined as a combination of four values: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The first two correspond to the correct prediction of the class c_i and the last two to the incorrect one. The sum of all these values equals the total number of samples used in the prediction. Accuracy is the most widely used metric, measuring the percentage of correct predictions,

$$accuracy = \frac{TN + TP}{TN + FN + TP + FP} \ . \tag{2.12}$$

Despite its usefulness in understanding, in a general manner, the capability of a predictor, accuracy falls short when it comes to the distinction between positive and negative predictions. If in a concrete problem it is more important to correctly classify positive or negative cases, then, different measures are

needed. The recall (also known as sensitivity or True Positive Rate) and the specificity (also known as inverse recall or True Negative Rate) are appropriate for these cases, respectively defined as

$$recall = \frac{TP}{TP + FN} \quad \text{and} \tag{2.13}$$

$$specificity = \frac{TN}{TN + FP} \, . \tag{2.14}$$

Recall corresponds to the percentage of positive cases correctly predicted among all positive cases, and specificity is the negative counterpart. If we are not as interested in correctly identifying all positive cases but instead want to make sure that all positive predictions made by the predictor are correct, precision can be used,

$$precision = \frac{TP}{TP + FP} \, , \tag{2.15}$$

i.e. the percentage of the correct positive predictions among all the positive predictions made by the predictor. In the case of both recall and precision being important for a given classification task, the F1 score can be used by calculating the harmonic mean of these two metrics,

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \, . \tag{2.16}$$

To better evaluate each class's probabilities approximated by the classifiers we can use Receiver Operating Characteristic (ROC) curves. These correspond to a predictor's True Positive Rate (TPR) and False Positive Rate (FPR) calculated for different thresholds values, with the FPR calculated as

$$FPR = \frac{FP}{FP + TN} \, . \tag{2.17}$$

The threshold value corresponds to the cut-off between classifying a sample as negative or positive. By plotting various points corresponding to different thresholds values and connecting them, we obtain a ROC curve. The closer this curve gets to the top left corner of the graph (high TPR and low FPR) the better is the predictor performance. The Area Under the Curve (AUC) can also represent a good metric, corresponding to the probability that a randomly selected positive sample gets a higher score than a randomly selected negative sample.

ROC curves in unbalanced data can lead to misleading results. As shown by Saito and Rehmsmeier [117], a line close to the top left corner may not indicate a good predictor in this situation, due to the possibility of the predictor favoring the majority class. The Precision-Recall curve is suggested as a better alternative for unbalanced data since it does not suffer from the same problem. This curve is similar to the ROC curve but using instead precision and recall. Figure 2.2 shows an example of both these curves.

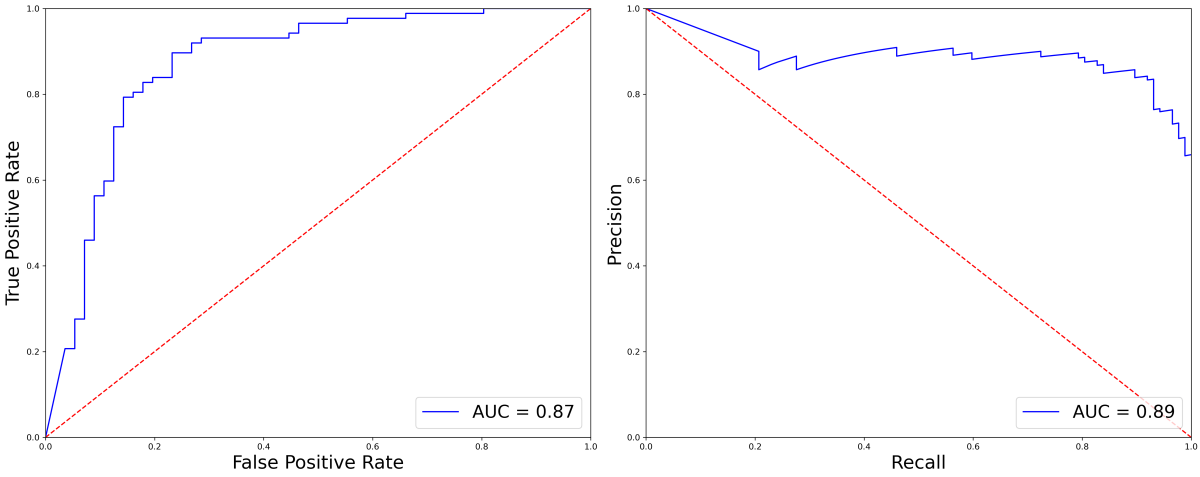

Figure 2.2: Example of Receiver Operating Characteristic (ROC) curve (left) and Precision-Recall curve (right)

3

Related Work

Contents

Advances in Machine Learning (ML) offer the possibility to turn complex data into accurate predictive models used within decision support systems that, together with human specialists, can help make more informed decisions. This is especially true when it comes to the Medicine field. The vasts amount of variables involved in making a diagnosis, prognosis or treatment plan can greatly difficult the work of physicians. Furthermore, the recent advances in high-throughput technologies, such as gene expression profiling, resulted in the creation of the domain of transcriptomic data analysis for diagnostic, prognostic and therapeutic ends, tailored to each individual. This type of analysis is greatly facilitated by the use of ML techniques, that can effectively deal with the high dimensionality of the data. This section describes contributions in the cancer transcriptomics area and its correspondent computational approaches.

3.1 Transcriptomics in Cancer Prognostic

Most of the prominent works in regards to gene expression in lymphomas are focused on Diffuse Large B-cell Lymphoma (DLBCL), the most common type of non-Hodgkin's Lymphoma. Among these works, some focus on the further distinction of subgroups that cannot be morphologically identified [4, 11, 149]. Alizadeh et al. [4] focus solely on the use of hierarchical clustering to make these distinctions, Wright et al. [149] also use clustering but add t-test results for each gene in order to calculate a linear predictor score that is later used to calculate the probability of the classes using Bayes' rule. The last work mentioned, by Bea et al. [11], resorts to t- and χ^2 tests to find abnormal differences between subtypes and uses a Cox proportional hazards approach to predict overall survival.

Works focused on prediction can also be found. Still in the DLBCL domain, Lenz et al. [89] predict survival of patients in two different chemotherapy regimens (CHOP and R-CHOP), using a Cox model to identify genes associated with survival that in turn were used to build a multivariate survival model, with a further analysis of the expression levels revealing three different signatures that were indeed predictive of survival. More in the domain of this dissertation, Shipp et al. [122] utilize supervised machine learning to predict how DLBCL patients will react to treatment (CHOP regimen), more specifically, they use weighted-voting of a combination of informative marker genes, an SVM and a KNN, all with similar results.

When switching to Hodgkin's Lymphoma, the availability of related works decreases. Gene expression profiling is performed by Devilard et al. [33] and Küppers et al. [82], both resorting to a combination of unsupervised hierarchical clustering and supervised clustering to identify important genes. The survivability of HL patients is analyzed in the work of Scott et al. [118], using a penalized Cox regression to separate the patients with high and low risk of death. In Steidl et al. [128], a sparse multinomial logistic regression is used to predict treatment failure.

The work that originated the target dataset used in this dissertation is another example, having a purpose identical to ours – the creation of a gene expression-based model to predict metabolic response after two courses of ABVD in HL patients [94]. The mentioned work performs a more traditional statistical analysis of the dataset in comparison with this dissertation. A multivariate logistic analysis was performed on the clinical variables (*age*, *gender*, *stage* and *Lymphocyte-to-Monocyte Ratio* (*LMR*)$>$*2.1*), showing that only *LMR*$>$*2.1* was significantly associated with the target variable *iPET2*. For the 765 genes, a differential analysis between the two possible target values was performed, identifying 241 significantly deregulated genes, with the majority (71%) being upregulated. This subset was further reduced by means of absolute Fold Change (FC)$>$2 and False Discovery Rate (FDR)$<$0.1, resulting in 13 genes with expression positively correlated with a positive value of *iPET2*. These genes were: chemotactic cytokines *CXCL2*, *CXCL3*, and *CCL18*; myeloid cells receptor *TREM1*; pro-inflammatory gene *SAA1*; the matrix components *PLAU*, *FN1*, and *SPP1*; the membrane matrix interacting proteins *ITG5A*, *CD9*, *LRP1*, and *THBS1*; and the pro-angiogenic factor *VEGFA*. The predictive model was created as follows. First, expression correlation analysis was employed to identify collinearity between the 13 selected genes, resulting in the removal of *CD9* and *FN1*, represented by the *ITGA5* gene. Then, a multivariate logistic regression was utilized to identify the features among the 13 genes and the clinical variable *LMR*$>$*2.1* that were independently associated with the target variable, resulting in the genes *ITGA5*, *SAA1*, *CXCL2*, *SPP1*, *TREM1* and the clinical variable *LMR*$>$*2.1*. These features were used in the final predictive model that, using an independent validation set of size 82 with a positive-negative distribution of 17.1% - 82.9%, obtained an average AUC of 0.68 (0.52-0.84), 69% specificity, 64% sensibility, and 68% accuracy.

It becomes obvious that the application of machine learning algorithms in Hodgkin's Lymphoma gene expression is a yet to be thoroughly explored domain, especially, when the aim is to predict the treatment outcome by means of interim PET.

3.2 Predictive Modeling of High Dimensionality Data

3.2.1 Feature Analysis

The development of Next Generation Sequencing techniques [139] has allowed, over the last years, the cheap and fast creation of datasets characterized by its high dimensionality. The growing interest of the ML community in this domain has led to a considerate increase in the volume of works addressing its inherent challenges [43, 107]. One of the main areas of research is the statistical and ML-based analysis of the dataset's features to detect the most important genes, whether to aid predictive tasks or the exploratory analysis of the role of genes on certain biological processes.

The pursue of better methods for gene identification has resulted in the creation of novel algorithms,

such are the already mentioned SVM-RFE [55] and RF-based [34], and also the Minimum Redundancy - Maximum Relevance feature selection framework, proposed by Ding and Peng [35].

Comparative studies with the goal of identifying the best suited algorithms for the various problems in the domain of gene expression are also recurrent. Some follow a more empirical approach, testing multiple combinations of feature selection methods in order to improve classification performance [17, 91, 92, 110]. Generally, these conclude that feature selection in this domain is of extreme importance. Others, present a more theoretical work, explaining the specifics of the various methods of analysis, and enumerating the ones more used in the literature. Some of the main works belonging to this second group are presented by Saeys et al. [116] and Lazar et al. [86], with the former being well acknowledged and bringing important insights into our work and the latter suggesting the use of the already mentioned Mutual Information and Wilcoxon Rank Sum algorithms. The work developed by Saeys et al. [116] refers to various feature selection techniques in the bioinformatics domain, performing an analytic review of the best methods in the various sub-domains, including, microarrays. They present a brief introduction about the common problems in this type of data, mentioning the characteristic large dimensionality, small sample size and variability and noise introduced by eventual experimental complications, providing references to key studies on the importance of dimension reduction in microarray data. They then proceed to approach what they call “the univariate filter paradigm”, describing the main univariate filter techniques used and why these simple methods are more often used than the complex wrapper and embedded techniques, providing references to works that verify the dominance of univariate filter methods. The review of the microarray domain ends with a suggestion that is followed by us, of performing a pre-reduction of the search space using univariate filter methods, and only then applying more complex methods, such as wrapper and embedded [116].

3.2.2 Predictive Modeling

One of the main goals in the transcriptomics domain, besides the identification of important genes or pathways in given biological systems, is the personalized prediction of an event of interest, be it the survival of a patient, the response to a given treatment, or disease predisposition. The classification of the functionality of a given gene is also a common objective. For these tasks, classification and regression models are generally learned to provide accurate predictions. An analysis of this domain’s literature shows a relative predominance of the SVM, with consistently good performance [21, 50, 88], and KNN also presenting good results [38, 109, 151] and an inherent simplicity for explaining clinical decisions in accordance with the observed outcome for individuals with most similar biological and clinical profiles. The gene expression process, as already explained, is prone to introduce noise in the data, which can lead to the assumption that ensemble algorithms will perform well due to their ability to not overfit to this noise. Indeed, some works show good results with ensemble methods, especially Random Forests

[53, 75, 113]. Still, there is a lack of studies about the efficacy of ensemble algorithms in this domain. The work by Tan and colleagues [133] shows the superior performance of ensemble learning in cancer classification but only when compared to single decision trees and not other algorithms. They state that a small number of training samples can lead a single individual classifier to approximate a different function to the data each time it is trained while maintaining the same performance. An ensemble algorithm, on the other hand, can get a more precise approximation of the true distribution of the data by averaging multiple functions. A distinction is made between the bagging and boosting methods, with bagging having a better performance, explained by the fact that these algorithms are less prone to overfit the noise present in the dataset.

Biclustering techniques have been largely employed in the discovery of putative gene sets in transcriptomic data [14, 29, 148], with pattern-based biclustering showing relevant performance indicators in diverse biological data contexts. Biclustering based on PAttern Mining Software (BicPAMS) [64] is an example of such, presenting an integration of dispersed state-of-the-art contributions on pattern-based biclustering. The software presented by this work allows for a high level of parametrization while performing efficient searches with guarantees of optimality. Among other aspects, BicPAMS allows for the customization of the returned bicluster's: i) coherency, through the selection of its assumption, strength and orientation; ii) structure, by adjustments in the minimum number of biclusters and support or in the pattern representations; and iii) quality, ensured by the parametrizable post-processing procedures.

4

Exploratory Analysis

Contents

This chapter presents a preliminary data analysis conducted with the goal of better understanding the data used throughout this dissertation as well as identifying major challenges to the descriptive and predictive ends. It starts with an overall description of the dataset followed by an in-depth exploratory analysis of its various aspects, including: class-conditioned variable distribution analysis; correlation analysis; and hierarchical clustering.

4.1 Data Description

This dissertation uses the dataset freely available at the National Center for Biotechnology Information Gene Expression Omnibus, denominated as Series GSE132348[1]. It consists of 106 samples of patients diagnosed with Classical Hodgkin Lymphoma. Each individual has associated the normalized expression levels of 765 different genes, obtained using the already mentioned nCounter platform over the RNA extracted from FFPE diagnostic tumor samples (section 2.1.4). These 765 genes correspond to the PanCancer Immune Profiling Panel of NanoString Techonologies [25], a gene expression panel tailored to better profile the immune response in cancer.

The following clinical variables are also included: gender, age, stage of disease according to the Lugano classification, and Lymphocyte-to-Monocyte Ratio (LMR)$>$ 2.1. Finally, each record contains the result of an interim PET realized after two courses of ABVD chemotherapy (*iPET2*), which was transformed to "positive" and "negative" values according to its classification on the Deauville 5-point scale [98], with PET defined as positive when its ordinal value is greater or equal than 4. For clarification, a positive PET is one that shows an indication of bad prognosis under the current treatment, while a negative PET outcome is associated with cancer remission. More information about the process of data collection can be found in the original work [94]. The data is relatively unbalanced, with 84 (80%) *iPET2* negatives, and 21 (20%) *iPET2* positives.

4.2 Data Profiling

In order to better understand the data at hand, an exploratory analysis is presented. We start with the study of the individual clinical variables (*gender*, *age*, *stage* and *LMR*$>$*2.1*) and the relations amongst some of them. The samples are evenly distributed among both genders, with a distribution of 54 (52.4%) females and 49 (47.6%) males. There is no evidence in the data that the distribution of *iPET2* results varies with *gender*. The patients' ages range from 15 to 75 years, with a mean of 40 years. In figure 4.1 we can see the Kernel Density Estimation (KDE) of the ages distribution according to *iPET2* result (left) and disease stage (right). *iPET2* positive cases tend to be more focused around 32 years, while

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132348 (accessed June 28, 2021)

negative ones are more spread out, as seen in figure 4.1 on the left. We can also notice, on the right, a tendency of younger patients in stage II of the disease, and older ones in stage IV, with stage III more spread out. Since the data only contains one patient of stage I, its distribution was not plotted.

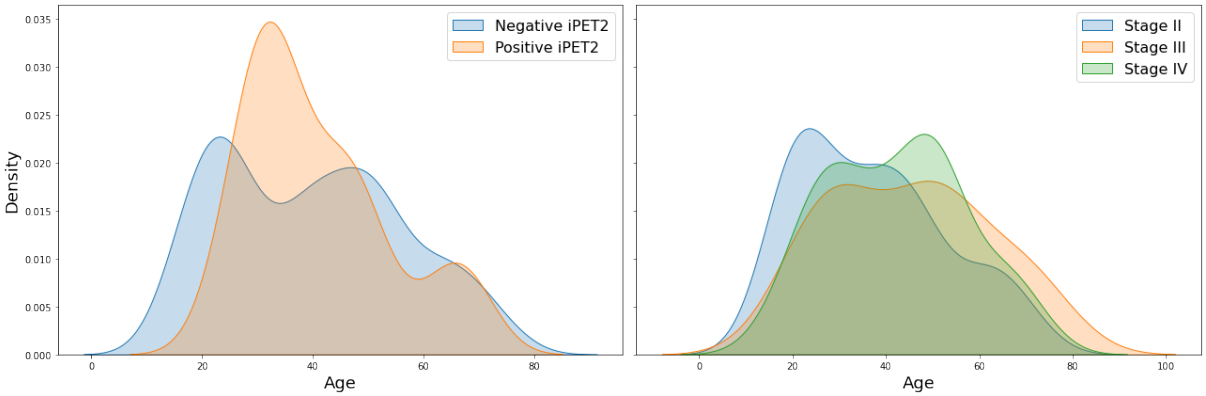

Figure 4.1: Age distribution by *iPET2* result (left) and disease stage (right)

Next, in figure 4.2, we can see the distribution of the various stages (left), and the percentage of negative and positive *iPET2* cases inside each stage (right). The only stage with more positive than negative cases is III A, and some stages possess only negative cases, probably due to the small number of samples they encompass, 1 for I A and 5 for III B. It is worth noting that our dataset does not possess any patients in stage I B.

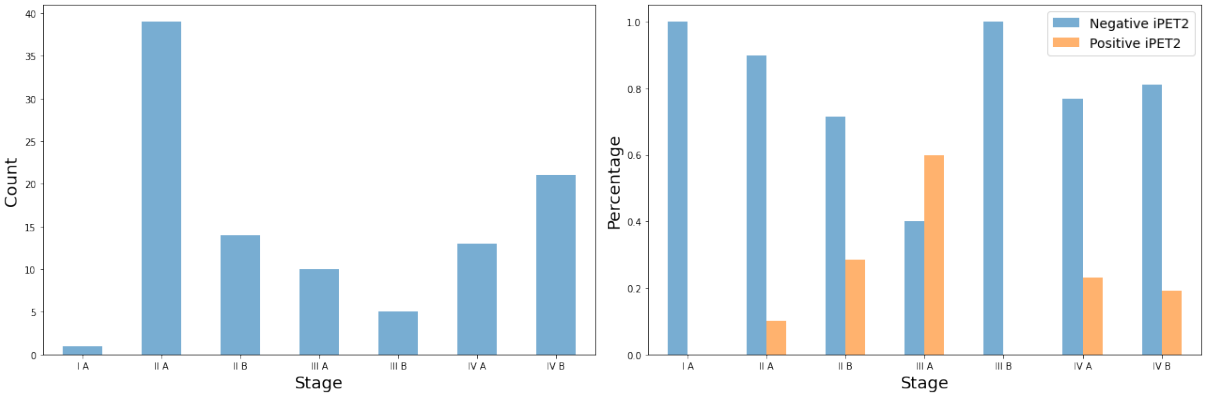

Figure 4.2: Distribution of stages (left) and *iPET2* results distribution by stage (right)

In respect to the binary variable Lymphocyte-to-Monocyte Ratio (*LMR>2.1*), although it has an even distribution with 49 (47.6%) positives and 54 (52.4%) negatives, a tendency is noted on the conditioned distribution of *iPET2* results. As seen in figure 4.3, negative cases of *LMR>2.1* have an increased percentage of positive *iPET2* results, possibly indicating a correlation between the two variables. This possibility will be further explored.

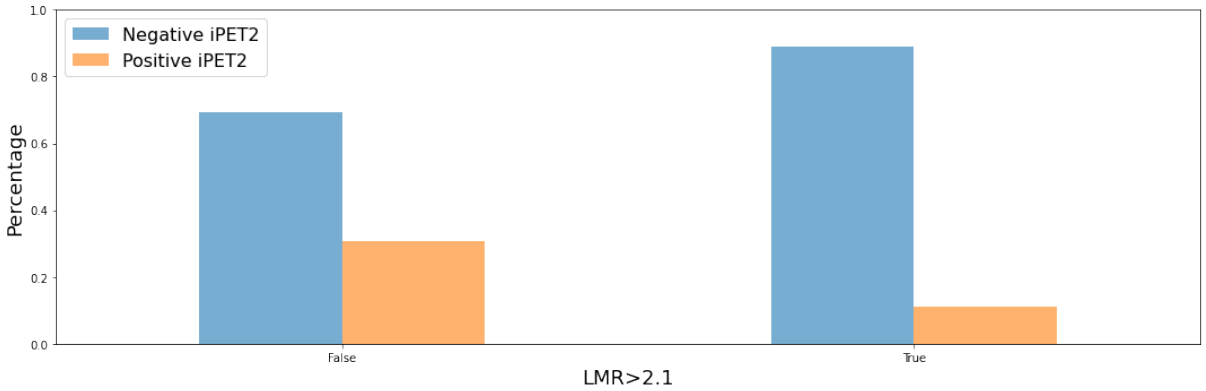

Figure 4.3: *iPET2* results distribution by *LMR>2.1* value

Correlations amongst the mentioned clinical variables were measured, using the χ^2 statistic in relations amongst categorical variables, and the one-way ANOVA test for the correlation between *age* and the remaining variables. The values presented in figure 4.4 correspond to the p-value of these statistical tests, which means, that in order to conclude that a correlation is present and significant, it should be lower than a defined threshold, in this case, p-value < 0.05.

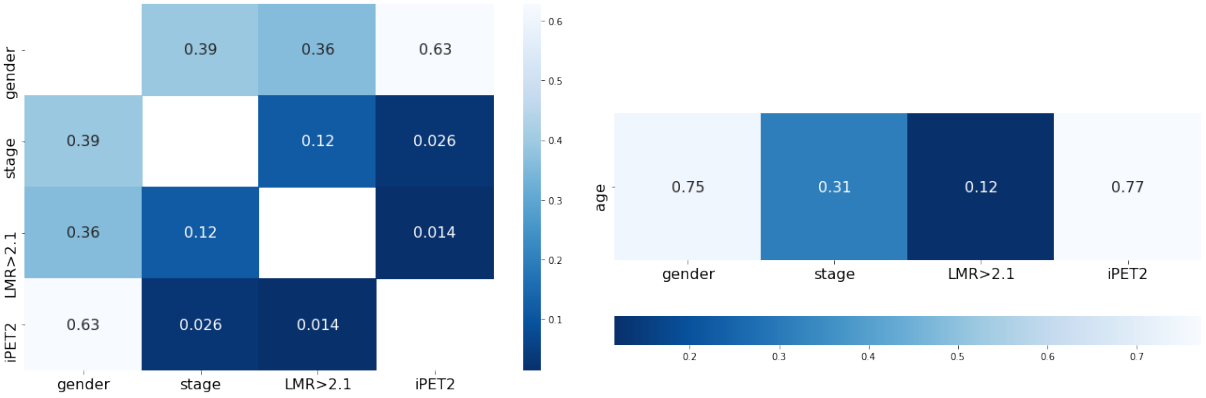

Figure 4.4: Correlations expressed as p-values between categorical variables (left) and between *age* and remaining variables (right)

By observing figure 4.4 we can conclude that the only significant relations are the ones between *stage* and *iPET2*, and between *LMR>2.1* and *iPET2*. It is worth noticing that only the latter correlation was identified by the original article [94].

Following this, we advance to the analysis of the gene expression data. It is important to mention that Luminari et al. [94] identified a 13-gene signature associated with the variable *iPET2*. This set of genes will be compared with our results. A Shapiro-Wilk test for normality [121] was performed on the 765 genes, and 83% of these presented a p-value below 0.05, indicating that they do not follow a normal distribution.

The first analysis of gene differentiation between classes was performed using Fold Change (FC). This corresponds to the proportion of difference in expression of a gene from one class to another, in this case, from negative to positive cases. The proportion values presented are $\log_2$ transformed in order to have an equal representation of positive and negative proportions. A volcano plot, as seen in figure 4.5, was used to analyze the FC of various genes. This plot presents each gene as a point, with the x-axis corresponding to the $\log_2$(Fold Change) and the y-axis corresponding to the $-\log_{10}(p\text{-value})$. This enables an analysis of the FC value while also confirming the significance of said value. As customary in this plots, we further define two thresholds, one for the $\log_2$(Fold Change) at 1 (or-1), corresponding to doubling or halving of the expression values, and another for the $-\log_{10}(p\text{-value})$, at 1.3, corresponding to a p-value of 0.05. With these in place, the genes that are situated in the top right and left corners (colored differently), correspond to the genes with a statistically significant value of Fold Change. In this group we find 6 down regulated and 47 up regulated genes, indicating that the majority of deregulated genes are oncogenes.

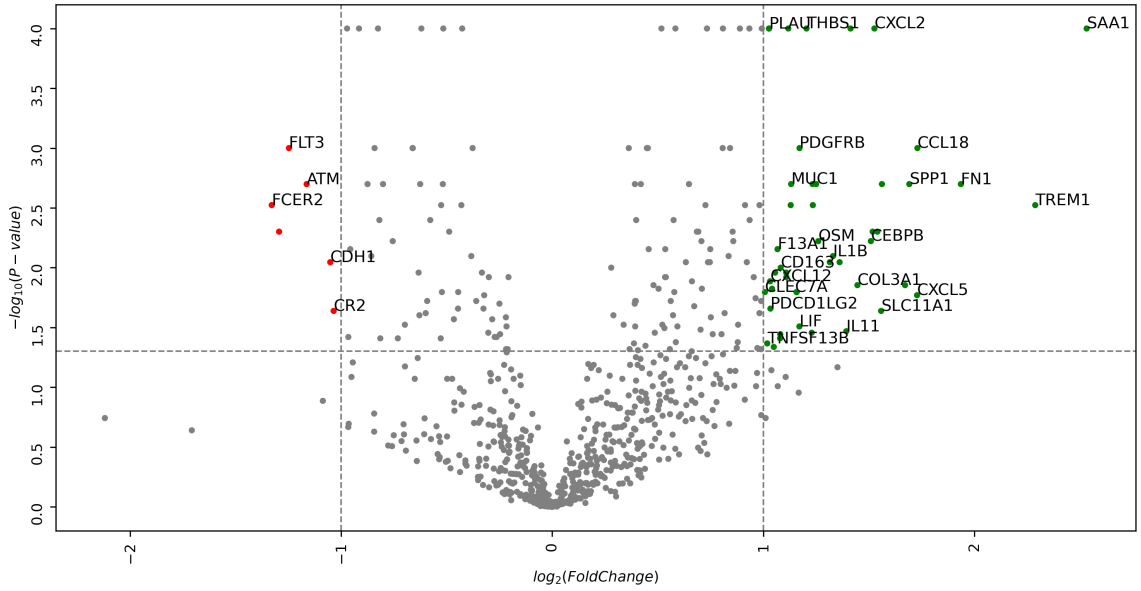

Figure 4.5: Volcano plot of all genes, with significant FC values as red (down regulated) and green (up regulated)

Next, we tried another approach to identify relevant genes, namely the feature selection approach described in 2.2.4, using a combination of Mutual Information (MI) and Wilcoxon Rank Sum Test (WRST). With a maximum threshold of 0.05 p-value, we obtained 95 and 99 genes with MI and WRST, respectively. The union of the two sets contains 172 genes and the intersection 21. By analyzing the 21 genes found in the intersection of the methods, it is noted that all these genes present a down regulation from negative to positive cases, a trend contrary to the one found in FC analysis. This indicates that the genes identified with these methods are all cancer suppressing genes. This relation can be seen in figure 4.6.

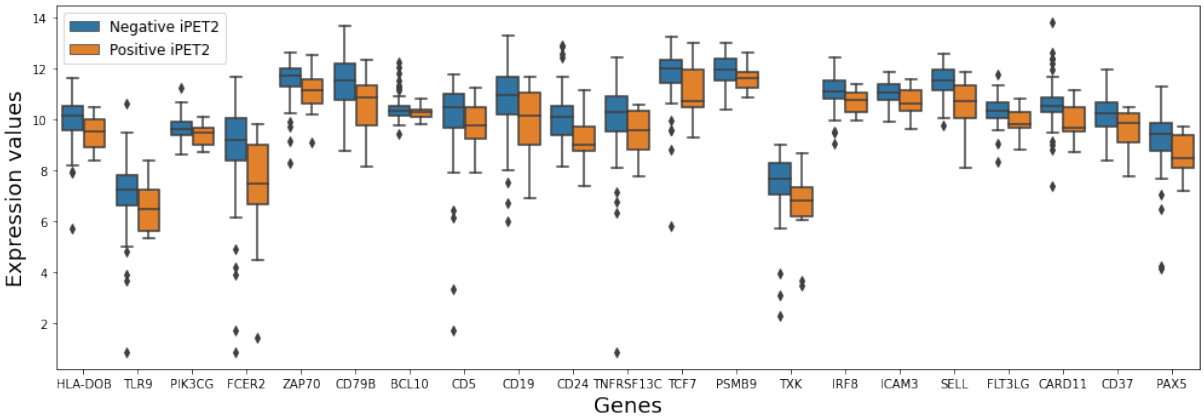

Figure 4.6: Boxplot representing the distributions of the genes selected by the intersection of Mutual Information and Wilcoxon Rank Sum Test

Through the Venn diagram in figure 4.7, we can observe the intersections of genes between the three previously mentioned methods for gene selection, namely Mutual Information ∪ Wilcoxon Rank Sum Test, Fold Change and the previously selected by Luminari et al. [94]. The low quantity of genes found in the intersections of the methods indicates a discordance between them. These results further corroborate our previous notion that no single method can accurately identify all the important genes.

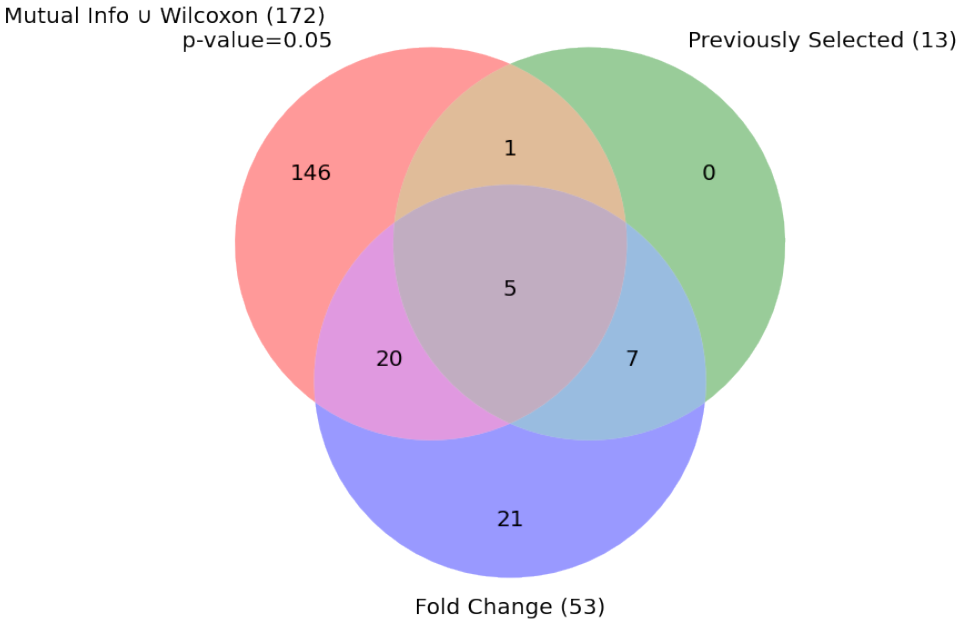

Figure 4.7: Intersections between the three different gene sets

We next resorted to hierarchical clustering to evaluate if a good separation of patients according to its class could be performed, while also trying to form clusters of related genes. Utilizing the set of genes selected from the intersection of the methods Mutual Information and Wilcoxon Rank Sum Test,

and trying different linkage methods, the best differentiation obtained is displayed in figure 4.8, showing the genes as rows and the samples as columns. Positive and negative *iPET2* samples are represented in the top row, respectively, with blue and green. The results of the hierarchical clustering suggest a significant difficulty in grouping together the patients by class, confirming that this is indeed a difficult problem that should be carefully approached.

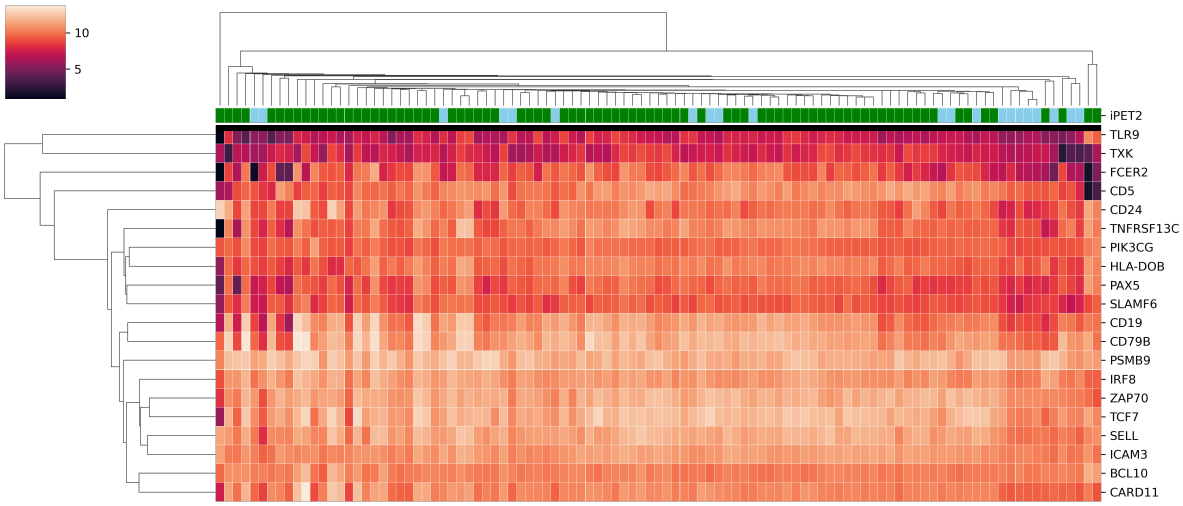

Figure 4.8: Hierarchical clustering of genes (rows) and samples (columns), with the sample's class represented as blue (positive *iPET2*) or green (negative *iPET2*) on the top row

Finally, in order to best assess the intrinsic difficulty of the classification task at hand, the top nine discriminative features were identified through the calculation of its Mutual Information. These features are plotted in figure 4.9 where we can observe the high overlap between classes, indicating that even the most discriminative features fail to individually separate the two classes. This confirms once again that the tackled classification problem is indeed a difficult one.

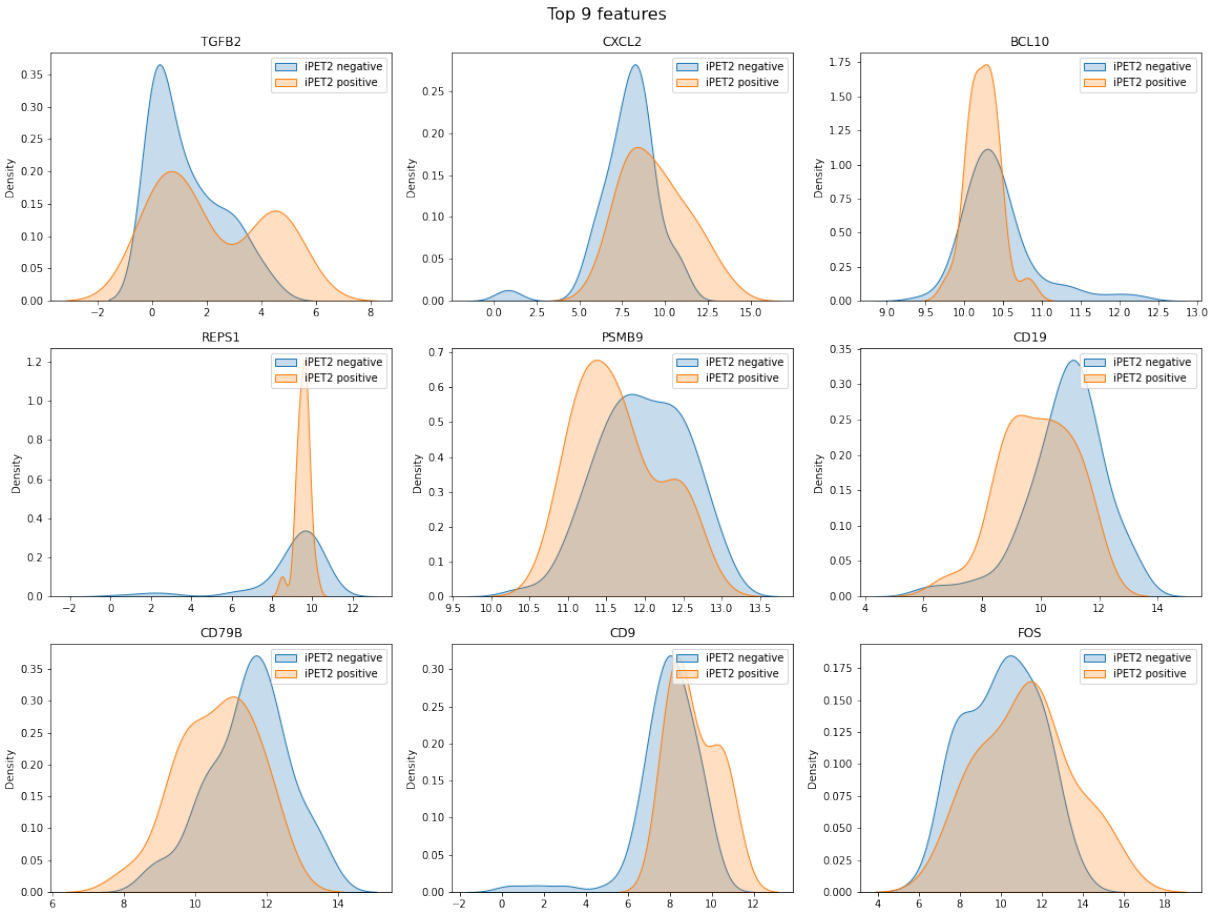

Figure 4.9: Distribution of top discriminative genes according to Mutual Information

5

Solution

Contents

This chapter describes the methodology to answer the target research problems: i) accurate prediction of response to treatment by anticipating *iPET2* result; and ii) identification of molecular mechanisms responsible for different responses to the treatment through the analysis of gene expression profiles and trained predictive models.

As depicted in figure 5.1, this methodology has two major objectives, corresponding to the previously mentioned predictive and descriptive problems. In order to achieve the objective denominated as "Prognostic", we start with the preprocessing of the data, followed by a feature analysis divided into two phases, which will result in an optimally transformed dataset where we perform a predictive analysis using state-of-the-art ML models. These results will then be compared with the ones obtained with the same classifiers when the data suffers a bicluster-based space transformation. The goal of obtaining "New Knowledge" is achieved through a parallel analysis of the results obtained throughout the already mentioned steps. Functional enrichment analysis [130] will be employed on various gene sets, namely, on the ones originated through the two feature analyses and on the patterns identified by the biclustering algorithm. The gene sets selected as potentially important in discriminating treatment response will then suffer a more detailed biological analysis together with the gene interactions implied by the trained ML models.

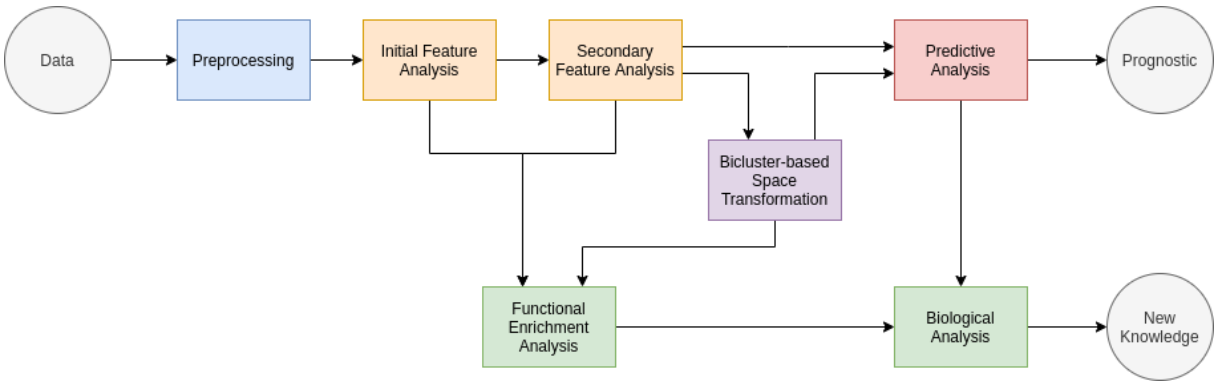

Figure 5.1: Methodology of solution

5.1 Preprocessing

One sample not containing the value for the target variable *iPET2* was removed from the dataset. Further anomalies were encountered, with one patient not having a value for the variable *LMR>2.1* and another with *stage* defined as "I", without specifying the subcategory, *A* or *B*. Both samples were removed, which resulted in a new distribution of 82 (79.6%) *iPET2* negatives and 21 (20.4%) *iPET2* positives.

The variable *stage* was encoded as an integer according to its severity, starting in 1 at "I A", and ending in 8 at "IV B", indicating that both an higher stage and the *B* variant are more problematic.

The values of gene expression are already properly transformed, as explained in detail by Luminari et al. [94], being worthy of mention that after analysis of each gene's distributions these were $\log_2$ transformed in order to better represent variations. With these values all belonging to the same domain, further normalization would only eliminate the possibility of comparing if one gene is more or less expressed than the other, although making possible the comparison of variance among different patients. In our specific case, it is desired to be able to compare levels of expression between genes, and so, the gene expression values are not further normalized. The only numerical clinical variable, *age*, is normalized only when this transformation makes a difference, namely, on non-associative classifiers.

Although the dataset is considerably unbalanced, with a target's distribution of around 80/20, the usage of balancing techniques is carefully considered. Due to the clinical nature of the data, oversampling methods, that either duplicate observations or create synthetic samples, can introduce bias to the data. Subsampling methods, on the other hand, if performed in our already small dataset, would result in a loss of observations that would further impact the ability to offer generalization guarantees. Having these facts in consideration, the utilization of balancing methods was decided through a comparative analysis of its effects on classifier performance, with the combination of both over and subsampling techniques being considered as a possible solution. The most promising method was oversampling using Support Vector Machine Synthetic Minority Oversampling Technique (SVM-SMOTE) [105], and so, it is the method used in the presented results unless stated otherwise. The Synthetic Minority Oversampling Technique (SMOTE) [26] is an oversampling technique that instead of simply duplicating samples of the minority class, creates new synthetic samples by finding two close samples from the minority class and randomly selecting a point between them. SVM-SMOTE [105] is an alternative of SMOTE that uses an SVM to decide which minority samples should be used to create the new ones, focusing on the minority class instances residing along the decision boundary of this model.

5.2 Feature Analysis

The data at hand is characterized by a high dimensionality, 770 features, and a low number of samples, corresponding to 103 patients. In this context, the posterior descriptive and predictive analysis can benefit from dimensionality reduction. This is recurrent in the analysis of gene expression data and its utility has been demonstrated [110, 124]. Feature selection was chosen specifically instead of feature extraction methods, such as Principal Component Analysis [69], in order to maintain the original representation of the variables and study the resultant more significant genes and clinical variables. We follow the practice suggested by Saeys et al. [116] and do a pre-reducing of the feature space using univariate filter methods, and then, apply the more complex embedded methods. This decision is further supported by frequent results in the literature where univariate filter methods have similar or better results

than more complex embedded/wrapper methods [53, 58, 108, 142]. This will also enable a more robust analysis of the genes selected as important in defining response to treatment. The initial selection by a filter approach is independent of algorithms and so its result is not influenced by classifier biases.

5.2.1 Initial Feature Selection

By performing a Shapiro-Wilk test [121], it was noted that most features do not follow a normal distribution. Due to this and the small number of samples, parametric tests, that assume a specific distribution of the data, were not considered. The initial selection was then performed by the non-parametric Wilcoxon Rank Sum Test and Mutual Information, already mentioned in section 2.2.4. The former returns a p-value on the probability that, given an independent variable y_j and the dependent binary variable z, the distributions $y_j|z = 0$ and $y_j|z = 1$ are equal. The latter statistic does not test a hypothesis, but a p-value can be subsequently generated using a one-sided permutation test [39]. The features with a p-value below a defined threshold and present in the results of any of the algorithms were selected. Since the approaches used are univariate, and therefore do not take into account interactions between variables, a less strict than usual threshold of 0.1 was used for the p-value, in order to be confident that important genes are not removed.

5.2.2 Secondary Feature Selection

A second feature selection stage by embedded methods is performed by the algorithms already explained in sub-section 2.2.4, using Support Vector Machines and Random Forests. The SVM-RFE [55] stands out for its good results, especially when in combination with SVM classifications models [110], while the Random Forest based feature selection [34] provides a smaller set of genes than most alternatives while maintaining predictive power. This additional phase of feature selection aims to further diminish the feature space, leaving only the truly important features for the task of predicting treatment response. The resultant genes are known to be discriminative of our target variable, and so, will be further analyzed to identify relevant molecular mechanisms influential in how a patient reacts to treatment.

5.3 Predictive Analysis

5.3.1 Machine Learning Models

With basis on the literature referenced along section 3.2.2, SVM, KNN and RF are selected as adequate classifier candidates to integrate our pipeline. It is hypothesized that the XGBoost classifier can also have a good performance due to its ensemble nature and consistent performance in other domains, and so, it is also included. In addition to these, the Decision Tree is covered with the expectation that an

analysis of its nodes brings important insights about the genetic component of the disease. Finally, both Naive Bayes and a random classifier are utilized as baselines predictors.

In order to obtain the best classification possible, all these algorithms are subjected to parameter optimization through the Tree Parzen Estimator algorithm [15] implemented by the `Hyperopt`[1] library, in addition to an optimized application of the aforementioned preprocessing techniques. The specific parameters subjected to optimization per classifier are available in table A.1 (appendix A).

5.3.2 Evaluation Methodology

Since we are working with a small dataset, with only 103 samples, the use of Cross-Validation (CV) for evaluation of the trained models is pursued. It has been shown that CV is a good way to avoid biased estimates that would come from performing a single evaluation on a reduced number of samples, as would be the case if we used the hold-out method [74]. The number of folds is fixed at 10. It was shown by Varma and Simon [140] that using CV for model selection can result in a highly biased estimate if the same CV procedure is used for parameter tuning, and so, we employ the use of nested CV to avoid this.

Nested Cross-Validation is used to separate the data used for hyperparameterization from the data used for testing the model. An intersection between these two sets would result in the model learning the best parameters for the data in which it will be tested, and consequently, perform better than its true predictive capability. In a nested CV, inside each CV loop used for model evaluation, another CV must be performed in the training data for parameter tuning and preprocessing. An example of a nested CV with 5 folds is shown in figure 5.2.

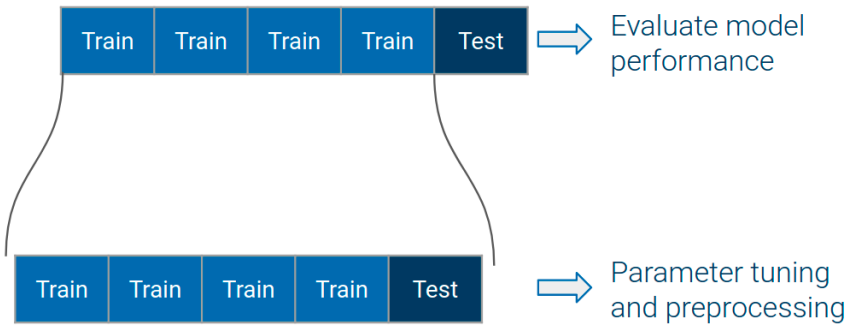

Figure 5.2: Nested cross-validation example

5.3.3 Performance Metrics

In order to properly define the adequate evaluation metrics, it is first necessary to have a clear understanding of the goal of the predictor. Our objective is to create a reliable decision-support system that

[1] `http://hyperopt.github.io/hyperopt/` (accessed June 28, 2021)

can help decide the intensity of the treatment a patient must undergo, with a positive prediction indicating that a more aggressive regime is necessary. If the predictor returns a positive result for someone that should have been negative, then we are subjecting the patient to more toxicity than necessary. If the contrary happens, then the weaker treatment can end up not being enough, and the patient must be treated with yet another, stronger regimen, resulting again in more toxicity than necessary.

Taking all this into consideration, the following metrics were chosen: i) AUC for an overall indicator of how the predictor performs even if the decision threshold is not optimized; ii) recall and precision to ensure that the predictor does not skew towards the majority class, especially important due to the data being unbalanced; and iii) specificity to guarantee the identification of patients who will react well to the standard treatment, avoiding the prescription of an unnecessarily stronger chemotherapy regimen.

Receiver Operating Characteristic (ROC) curves are further plotted to evaluate the best threshold for classification in addition to Precision-Recall curves to provide unbiased insights about the overall performance, both offering a more comprehensive comparison of the behavior of predictors.

The metric optimized in the hyperparameterization step is the F1 score, already explained in section 2.2.5, so that the classifiers can attain a balanced performance in both precision and recall.

5.4 Bicluster-based Space Transformation

Although the feature selection process is heavily explored in order to obtain the best possible feature space for classification, this approach has its limitations. The first one corresponds to the specific limitations of the algorithms used, with the first phase using univariate methods that do not take into consideration feature interactions and the second phase using embedded methods that are biased towards a specific predictor. Another limitation lays in the fact that this approach only removes features, not exploring the possibility of combining multiple genes in a single feature, an approach that can be useful when analysing gene expression data since most biological mechanisms are dependent on the interactions of many genes.

An alternative approach to this is the creation of new variables corresponding to relevant and discriminative patterns of gene expression. It is expected that this type of feature space can better represent the complex interactions between multiple genes. An efficient and effective way of mining these gene expression patterns is through biclustering. As previously explained, biclustering finds subspaces (biclusters) associated with homogeneous and statistically significant patterns, with the rows in our case corresponding to patients, the columns to genes and the values to the expression of these genes.

5.4.1 Hyperparameterization

The biclustering algorithm used for this transformation is the already mentioned BicPAMS [64]. As previously stated, this algorithm is characterized by allowing a high level of parametrization, an advantage that is fully explored in this work. With the goal of better understanding how the search parameters provided by BicPAMS affect the biclustering task, we perform an in-depth analysis of the impact its variations have in classification performance according to multiple metrics. A Naive Bayes predictor will be trained and tested using the already mentioned evaluation methodology on multiple datasets, each generated with varying values for multiple biclustering parameters. The parameters under analysis are: i) number of iterations, indicating how much times the search for discriminative biclusters is performed; ii) minimum lift, corresponding to the lift value a bicluster must achieve to be considered discriminative; iii) number of labels, defining in how many intervals the gene expression values of each column are discretized; and iv) maximum number of biclusters, corresponding to the number of top biclusters to be considered for the space-transformation, where the priority criterion is given by the discriminative power of a bicluster as given by its lift.

5.4.2 Space Transformation

After the biclustering task is concluded, the dataset is transformed so that each feature corresponds to one of these patterns encompassing multiple genes. Each value of the transformed dataset will then represent the similarity between the gene expression expectations associated with a given pattern and the actual levels of gene expression observed for an individual. The steps followed to perform this transformation are:

1. Mine discriminative patterns, represented by a set of features and its corresponding values;
2. Create new features, each corresponding to a found pattern;
3. Calculate the values of the new dataset by determining the distance between each individual's gene expression levels and the found patterns. This distance can be established in two ways:
    - Calculate the Euclidean distance between the individual's and pattern's values;
    - Determine if an individual presents a given pattern or not, resulting in a binary dataset. A tolerance threshold can be defined to accommodate for some noise.

Figure 5.3 presents an example of the described process, showing a found bicluster and corresponding the pattern on the left, and the resultant values on the transformed dataset using Euclidean distance on the right.

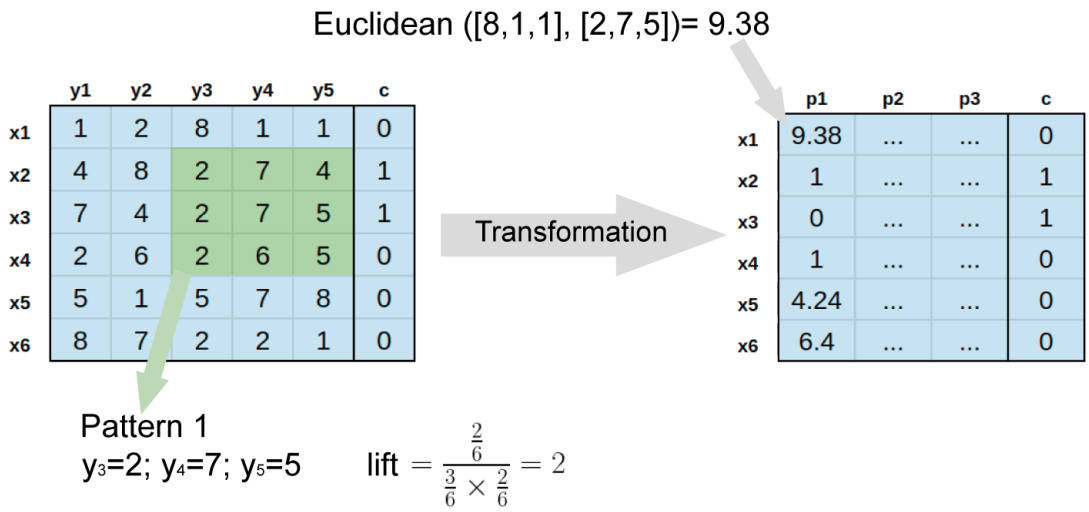

Figure 5.3: Example of space transformation using biclustering

5.5 Biological Analysis

Following the feature and predictive analysis, a biological analysis will be conducted to obtain insight about important genes and their interactions in the context of Hodgkin's Lymphoma.

The multiple gene sets obtained throughout the described methodology will be submitted to functional enrichment analysis using the Enrichr tool [80]. This type of analysis receives a gene set and checks for significant overlaps with annotated gene sets representing prior biological knowledge, thus allowing inference of new knowledge. Enrichr receives a set of genes and returns various enriched terms against multiple knowledge bases. In the present work these terms are ranked according to the c-score, the metric recommended by the platform,

$$c\text{-}score = \ln p \times z\,, \tag{5.1}$$

where p represents the p-value and z represents the z-score computed to assess the deviation from an expected rank precomputed using Fisher's exact test [45]. On the top of each term's bar is also represented the p-value adjusted using the Benjamini-Hochberg method for correction for multiple hypotheses testing [12]. The chosen knowledge bases against which the gene sets will be enriched are: i) Kyoto Encyclopedia of Genes and Genomes (KEGG) [77] for the analysis of enriched pathways, the set of the molecular interactions, reactions and relations networks between these genes; ii) Gene Ontology (GO) [31] for the analysis of enriched biological processes in which the genes are involved; and iii) Online Mendelian Inheritance in Man (OMIM) [59] for the analysis of enriched diseases, to verify if the gene sets are characteristic of Hodgkin's Lymphoma or other related disorders/diseases.

The first step of feature analysis, described in section 5.2, will result in a subset of genes believed to be influential in a patient's response to treatment. These genes will be analyzed by means of functional

enrichment analysis and compared with other related works, namely, the already mentioned work by Luminari et al. [94]. This is done to confirm our findings about the function of certain genes in the development of the disease, and hopefully, identify novel regulatory interactions.

After the second step of feature analysis using embedded methods and the learning of predictive models, the aforementioned process of gene analysis will be repeated. This time, it is believed that it will be possible to identify relations between the various genes by analyzing the structure of trained models, such as tree-based classifiers. These relations, besides indicating the genes' discriminative power, can also reveal putative regulatory modules of genes involved in relevant biological functions.

The evaluation of the findings from our biological analysis will be divided into three major components: i) test the non-triviality and actionability of the gene interactions found by confirming that they are indeed specific and representative of Hodgkin's Lymphoma; ii) examine the extent to which these patterns enable accurate discrimination of a patient's response to treatment; and iii) assess the biological novelty of these findings.

6

Predictive Results

Contents

In order to answer the target problem of performing an accurate prediction of response to treatment by anticipating *iPET2* result, we use a combination of Feature Analysis, Bicluster-based Space Transformation and Predictive Analysis. These steps are combined to allow for the best possible classification performance, with both the Feature Analysis and Bicluster-based Space Transformation aiming to create the optimal feature space for the classifiers used in this work, and the Predictive Analysis making use of state-of-the-art machine learning models and techniques to increase predictive power.

In this chapter, we present an in-depth analysis of the predictive results obtained by this methodology, including the results of both our feature selection stages, the impact of hyperparameterization on the creation of a pattern-based feature space and the classification results using both of these approaches.

6.1 Feature Selection

The initial phase of feature selection already detailed in chapter 5 returned a total of 250 features out of the initial 770. These features correspond to 248 genes and the clinical variables *stage* and *LMR>2.1*, filtering out the variables *age* and *gender*. The already mentioned work by Luminari et al. [94] found, during a first analysis, a 13-gene signature positively correlated with *iPET2* in addition to the variable *LMR>2.1*. Comparing our results with these ones we find that our feature set contains nine out of the thirteen genes and the *LMR>2.1* variable. The selection of the variable *stage* by our algorithm points to a relation with the target *iPET2* not identified through the multivariate logistic analysis performed by Luminari et al. [94].

Performing a similar analysis as the one in chapter 4 we can study the 36 genes found in the intersection of both filters, MI and WRST, in figure 6.1. It is visible that the previously noted trend of cancer suppressing genes is maintained.

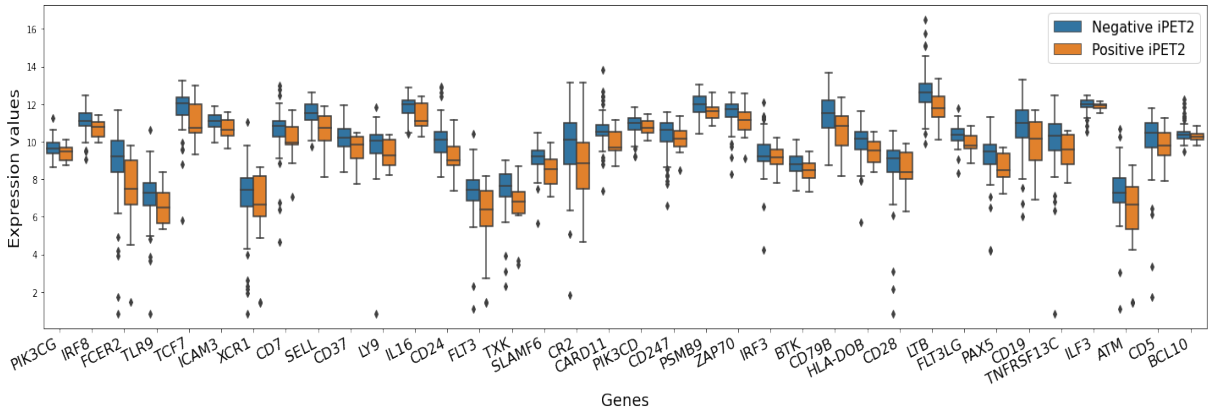

Figure 6.1: Boxplot representing the distributions of the genes selected by the intersection of Mutual Information and Wilcoxon Rank Sum Test

The second and final phase of feature selection is divided in two algorithms, SVM-RFE and RF-based feature selection, both explained in chapter 2. The SVM-RFE algorithm, as previously stated, only orders the features by their importance, and so, cross-validation with a varying number of features was performed in order to select the optimal number. Figure 6.2 shows various scores obtained for multiple numbers of features, with the final criteria for decision being the highest value of F1-score. The selected number of features was 14 in accordance.

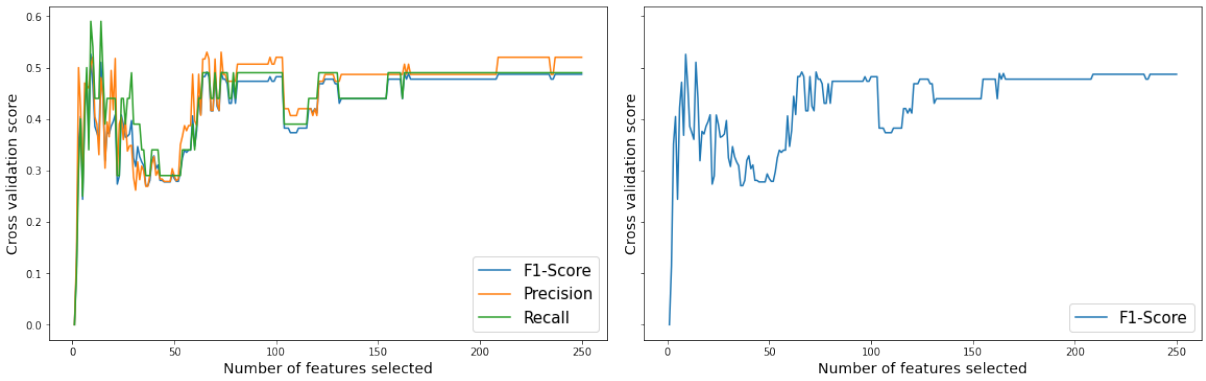

Figure 6.2: Cross validation score with varying number of features in SVM-RFE

The top 14 features returned by the feature selection method SVM-RFE corresponded to the genes *VEGFA*, *MFGE8*, *TLR5*, *CD80*, *SH2D1A*, *S100B*, *CXCL2*, *IL12RB1*, *HLA-C*, *CD8B*, *FCER2*, *CCL8*, *IL2* and *ENTPD1*. Considering these genes, only *CXCL2* and *VEGFA* are also found in the 13-gene signature previously mentioned.

The RF-based feature selection consistently returned only one feature, the gene *PLAUR*. Even if this is a highly discriminative feature, one single gene does not possess enough information to accurately make predictions, and so, this feature selection method was not further pursued.

6.2 Classification Performance

In order to provide an adequate baseline of classification results, we performed twenty iterations of training and validation of a random classifier in our dataset. Figure 6.3 shows the convergence of the random classifier's performance according to various evaluation metrics as more iterations are performed. It is the convergence value of these metrics that should be interpreted as the baseline value for this classification task, namely, a precision of 0.21, a recall of 0.49 and a specificity of 0.51. The AUC metric is omitted since by definition a random classifier achieves an AUC of 0.5.

Considering these reference values, we present the results obtained following the methodology presented in chapter 5 in two different configurations of feature selection and evaluation of the data. The first case purposefully corresponds to an approach similar to the followed in Luminari et al. [94], so that

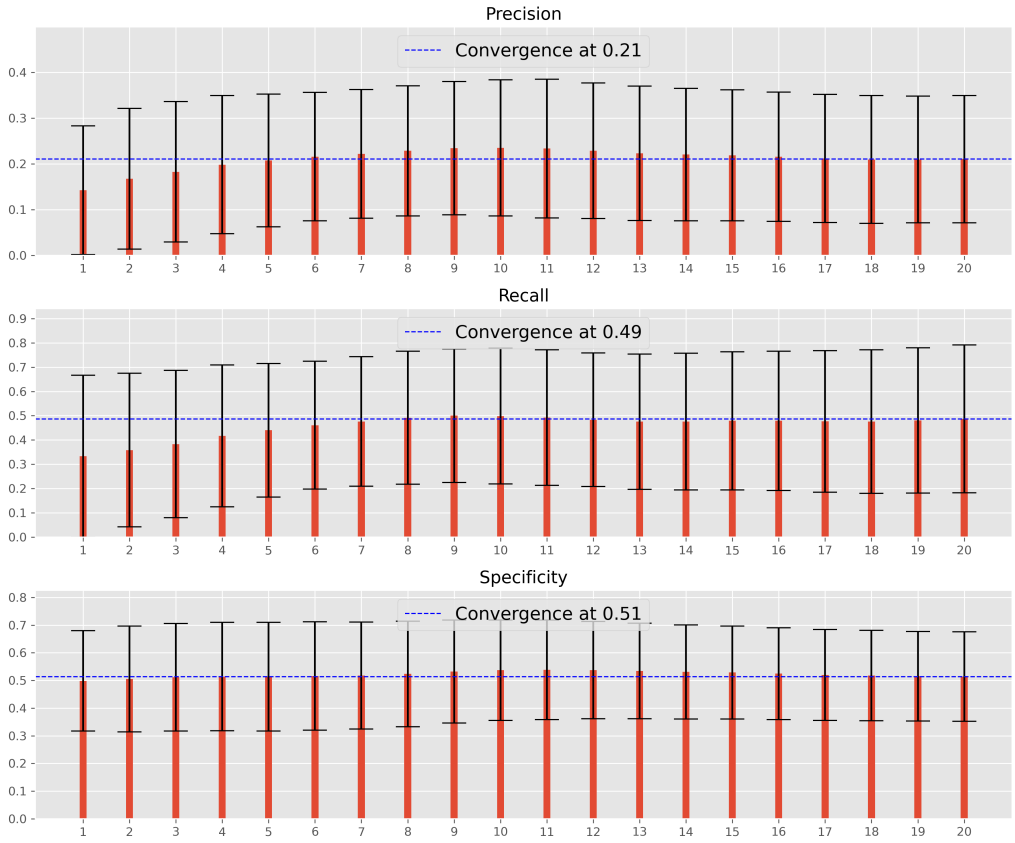

Figure 6.3: Random classifier convergence

our results can be compared to the ones obtained in this work. We first apply the two phases of feature selection which results in a dataset with the fourteen features mentioned in the previous section. This dataset then suffers the defined preprocessing steps and is used to train and evaluate our classifiers using internal bootstrapping. Our results are presented in figure 6.4, where each color encodes a given performance metric and the horizontal lines correspond to the random classifier's results in said metric.

Since we do not have access to the validation set used in Luminari et al. [94], we can only compare our results to the ones obtained in the training set, corresponding to an AUC of 0.84. We can then state that the combination of our two feature selections steps, data balancing using SVM-SMOTE and the classifier SVM, can achieve a superior mean result of 0.97 AUC. The high predictive power of an SVM in this setting is to be expected due to its recurrent good performance in this type of data [21, 50, 88]. In addition to this, this classifier is paired with the data balancing technique SVM-SMOTE and the feature selection algorithm SVM-RFE, both using an SVM as the base of its decisions.

The second configuration presented is more careful in the evaluation of our classifiers and only performs the feature selection step inside each fold of a nested cross-validation, as previously described. The results obtained in this setting are presented in figure 6.5. In this case, even though we are still

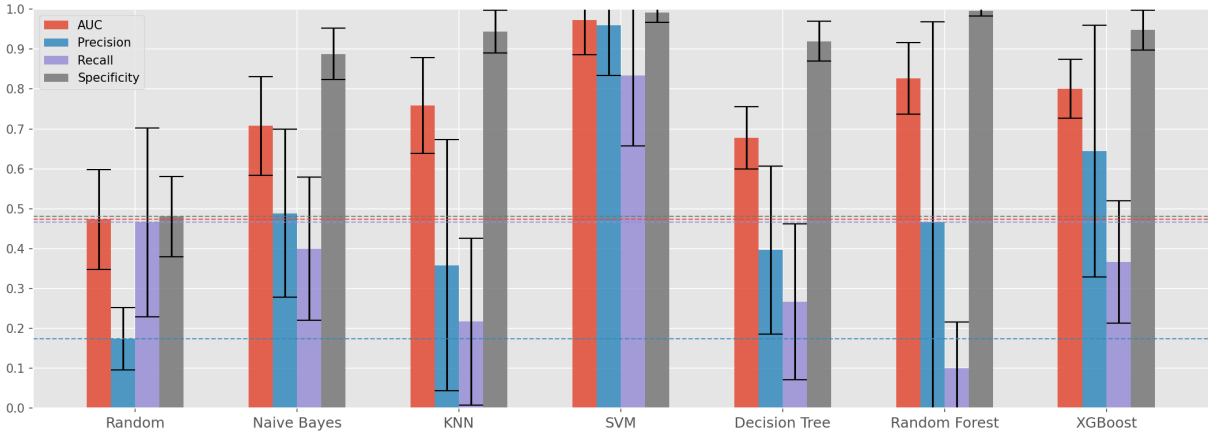

Figure 6.4: Optimized classifier's performance in the prediction of *iPET2* (configuration one)

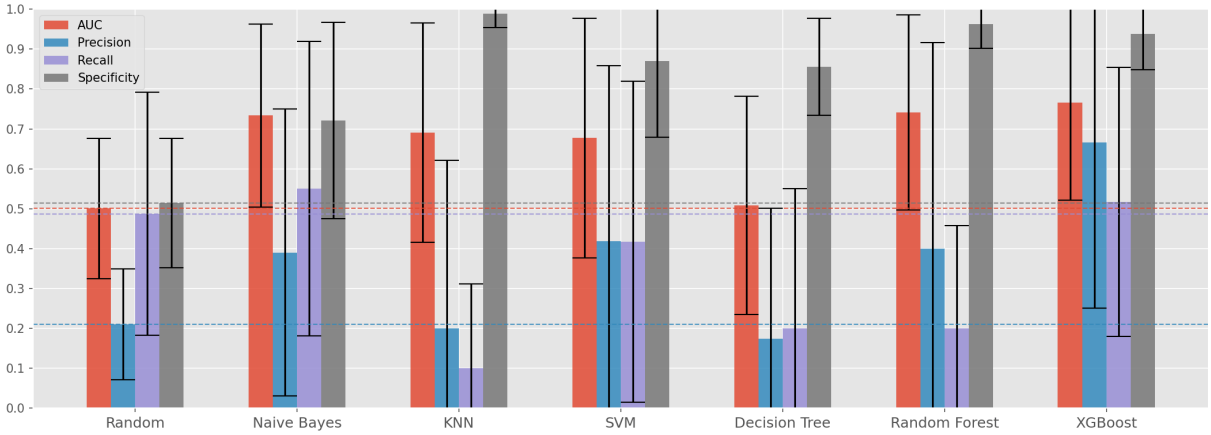

Figure 6.5: Optimized classifier's performance in the prediction of *iPET2* (configuration two)

using the algorithms SVM-SMOTE and SVM-RFE, the predictor XGBoost achieves better results than the previous best performer SVM, with an AUC of 0.77, a precision of 0.67, a recall of 0.52 and a specificity of 0.94.

KNN on the other hand stands out by presenting a consistently bad performance, worse than the defined baseline on the second configuration. This can be possibly explained by two reasons: i) a high percentage of outliers in our data; and ii) high expression variability of most of the selected genes, leading to inflated differences between individuals that belong to the same class.

Two out of the three tree-based classifiers, Decision Tree and Random Forest, also present comparable performance to the random classifier in one of the configurations. The contrast of their bad performance with the overall good performance of the XGBoost algorithm leads us to conclude:

1. A single Decision Tree is unable to fully take advantage of all the discriminative gene interactions present in the dataset. Its nature is to analyze features individually, and so, it fails to encompass all the available information;

2. Since the Random Forest and XGBoost classifiers are both ensembles of Decision Trees, the discrepancy between their results must originate due to the type of ensemble used, namely, bagging and boosting. Since bagging uses multiple independently trained Decision Trees to make predictions, the fact that a Decision Tree cannot assimilate the knowledge in the data will lead to an overall lack in performance for the Random Forest. Boosting on the other hand trains the trees iteratively, allowing for each consequent model to improve where the previous one failed, resulting in a classifier that can better model more complex interactions.

Overall, it is observable that all the classifiers can easily attain a high specificity, but at the cost of mediocre precision and recall. In other words, classifiers have an easier time correctly classifying patients with negative *iPET2*, guaranteeing that the patients that will react well to the ABVD regimen are correctly identified. The main difficulty with this predictive problem is in the correct classification of positive patients, possibly due to the low number of samples of this class leading to the inability of the classifiers to correctly learn how to identify them.

The predictive results can be further analysed through the study of the classifiers' Precision-Recall curves presented in figure 6.6, on the left for the first configuration and on the right for the second configuration. KNN is omitted from these graphics due to its dependence on an ineffective method to calculate the appropriate thresholds. ROC curves are also provided in figure A.1 (appendix A).

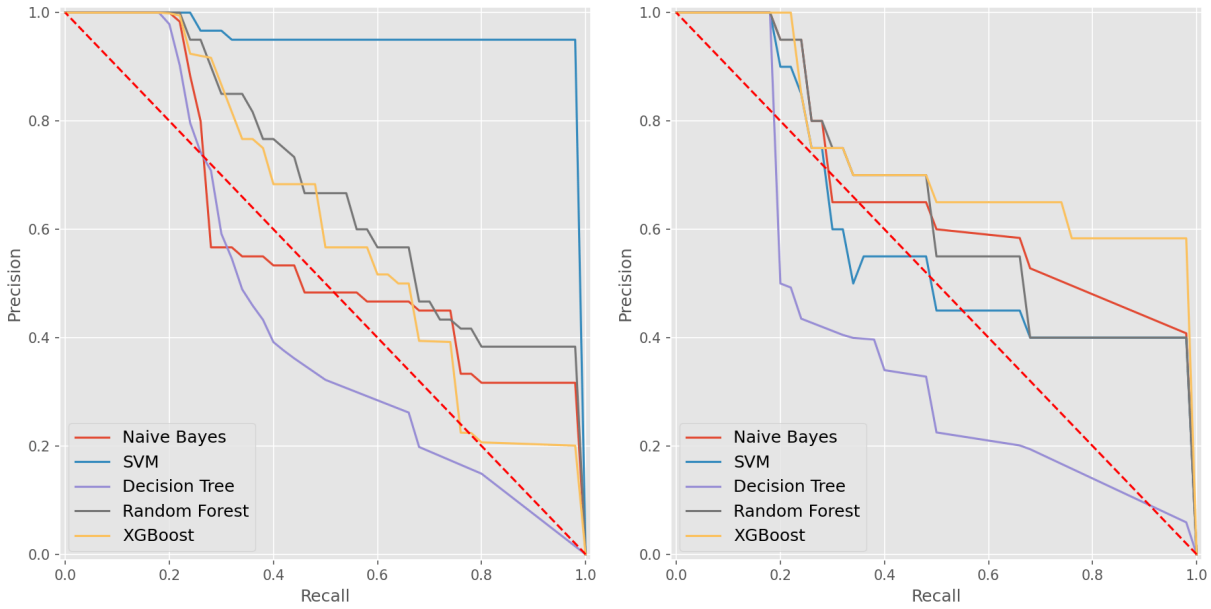

Figure 6.6: Precision-Recall curves for configuration one (left) and two (right)

6.2.1 Advanced Aspects

It is in our interest to better understand what leads a certain patient to be misclassified by our predictors. To do so, we plot some of the characteristics of the correctly classified individuals against wrongly classified ones by the best predictor in the second configuration, XGBoost. This analysis is provided for the four clinical variables, *age*, *gender*, *stage* and *LMR>2.1* in figure 6.7. Starting with the variable *gender*, no significant trend is noted, only a slight tendency for incorrectly classifying positive cases in female patients. *LMR>2.1* on the other hand shows a more clear inclination for correctly classifying positive cases when this variable is "False", with the percentage of True Positives (TP) being higher than the False Positives (FP) percentage. Regarding the *stage* variable, the values "I A" and "III B" are omitted due to the low number of samples corresponding to each one. In the observation of the remaining values only "III A" shows a significant deviation from the others, with all the positive cases correctly predicted but at the cost of a worse performance in the negative cases. The final plot is dedicated to the variable *age* and is presented in a stacked view, where the bins encompassing a 10-year period from each class (TN, FN, TP or FP) are stacked to facilitate a comparative analysis between them. We can then recognize that the majority of False Negatives (FN) occur in patients between 30 and 40 years, and the False Positives (FP) are more evenly distributed with a slightly higher concentration in patients between 20 to 30 years.

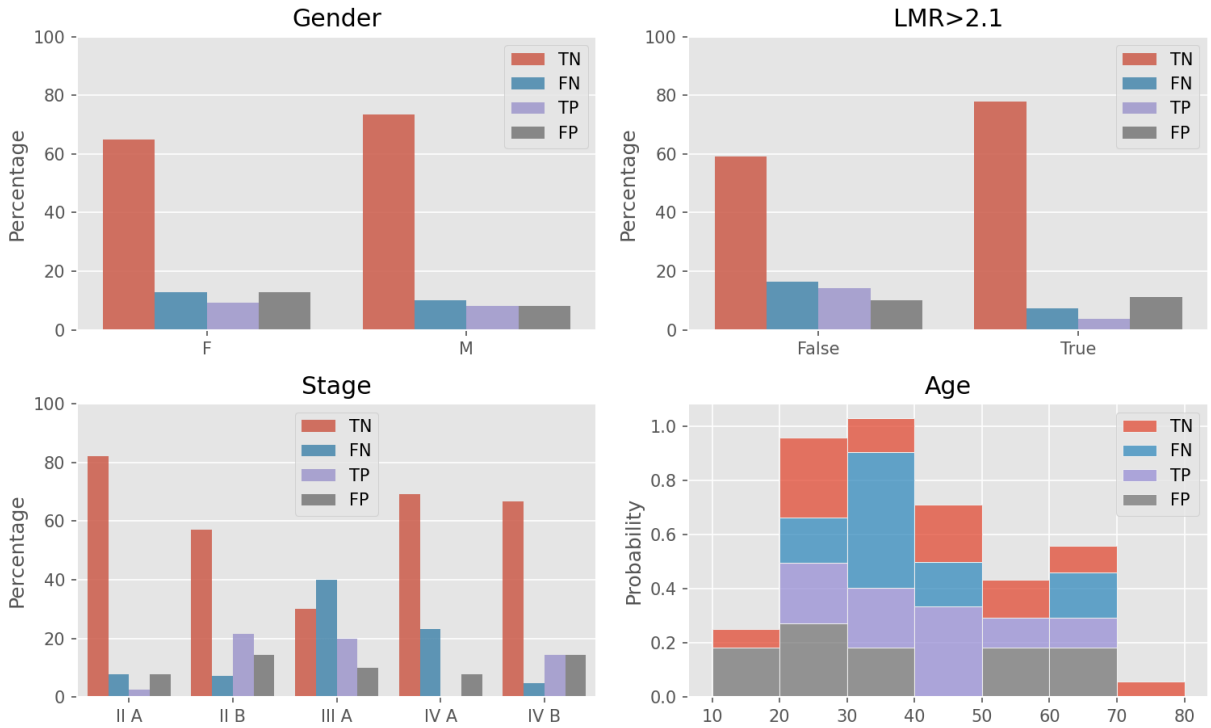

Figure 6.7: Distribution of predictions for the clinical variables

Although this analysis is useful to understand the factors that lead to a worse performance of our top classifier, it leaves out useful information obtainable through the other classifiers tested, and so, in figure 6.8 the intersection between the sets of wrongly predicted individuals by each classifier is plotted.

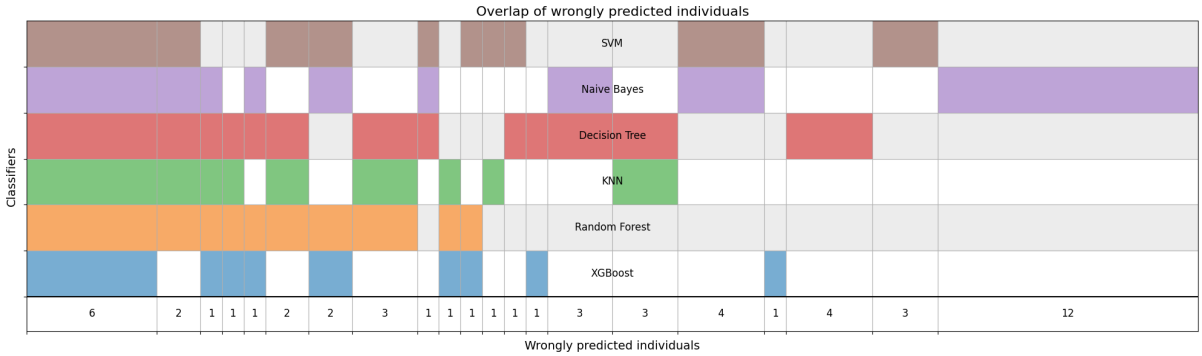

Figure 6.8: Overlap of wrong predictions by the studied classifiers

We can observe that there are six individuals that all predictive models fail to classify and not a single one that is correctly classified by all of them. These six individuals are further studied in order to understand which characteristics make them so hard to classify and are also compared with seven other individuals that are only misclassified by our two worst predictors, KNN and DT. In figure 6.9 we can see multiple pie charts representing the distribution of the four clinical variables in three different groups: the set of six individuals all wrongly classified, the set of all the individuals of our dataset and the set of seven individuals only misclassified by KNN, DT or both.

We can see in figure 6.9 that the *gender* variable does not have a visible impact on the predictive capacity of classifiers. *LMR>2.1* on the other hand clearly has some influence, with the all wrong set having a majority of "True" *LMR>2.1* cases while the barely wrong set presents the inverse tendency, with more "False" cases, reinforcing the trend previously noted in figure 6.7. The distributions of the variable *stage* indicate that patients in stage "III A" are harder to predict, while patients in stage "II B" are easier, with the remaining stages not presenting any significant differences. Finally, the *age* variable, here divided in four twenty-year intervals, shows a tendency for older patients, between 60 and 80 years, to be misclassified, and younger ones, between 0 and 20 years, to be correctly classified.

As our previous results have indicated, the majority of useful information about how the patient will react to treatment is present on the gene expression values, and so, it is imperative that the analysis of the factors inducing wrong classifications be extended to these features. In figure 6.10 we can see again the distributions of the nine more discriminative genes previously plotted in figure 4.9 but this time with the indication of the values corresponding to each of the six wrongly classified individuals. We can then observe that as expected, the majority of the values are found in the intersection of both distributions, where the classification is harder to perform.

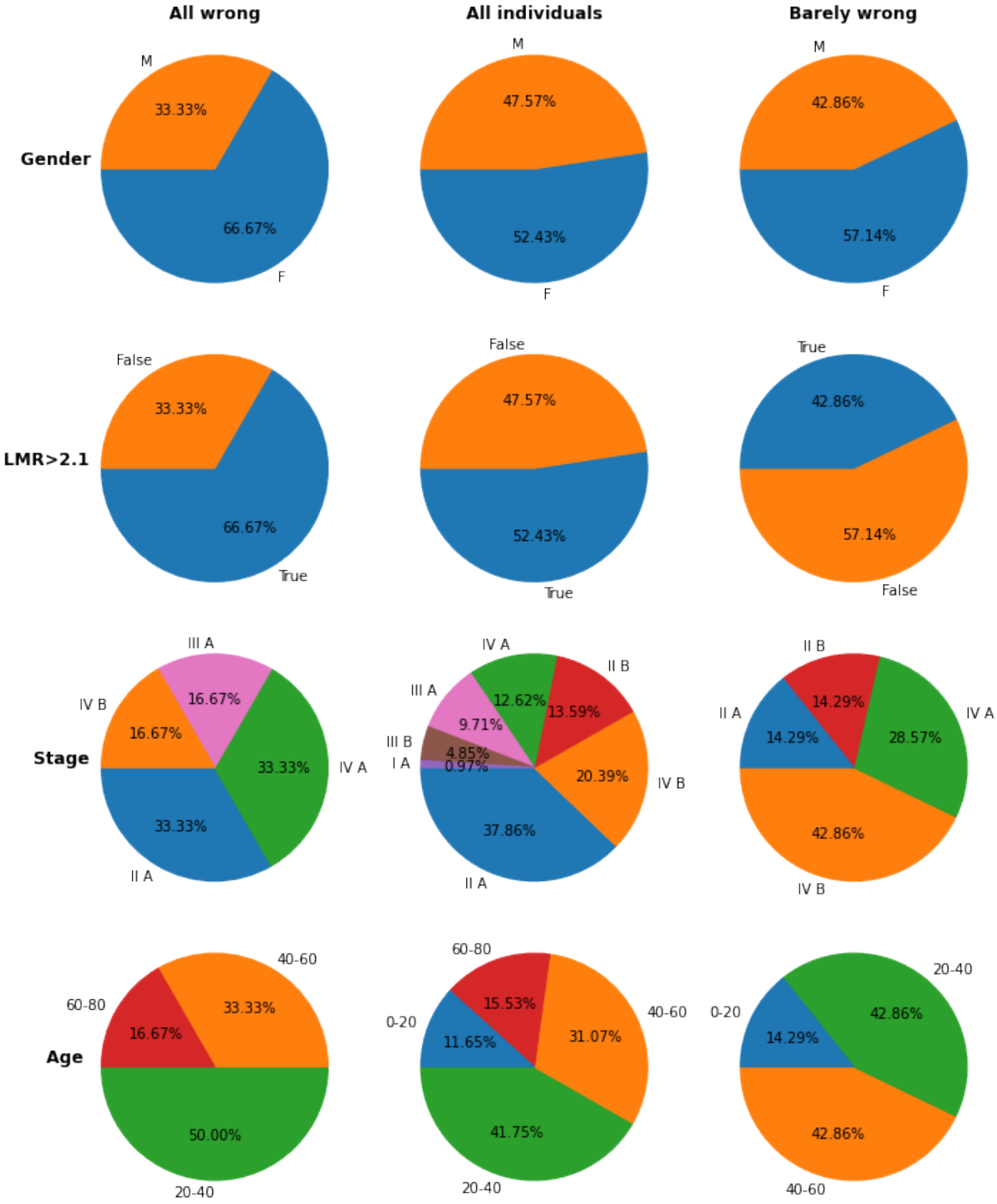

Figure 6.9: Distribution of clinical variables across various sets of individuals

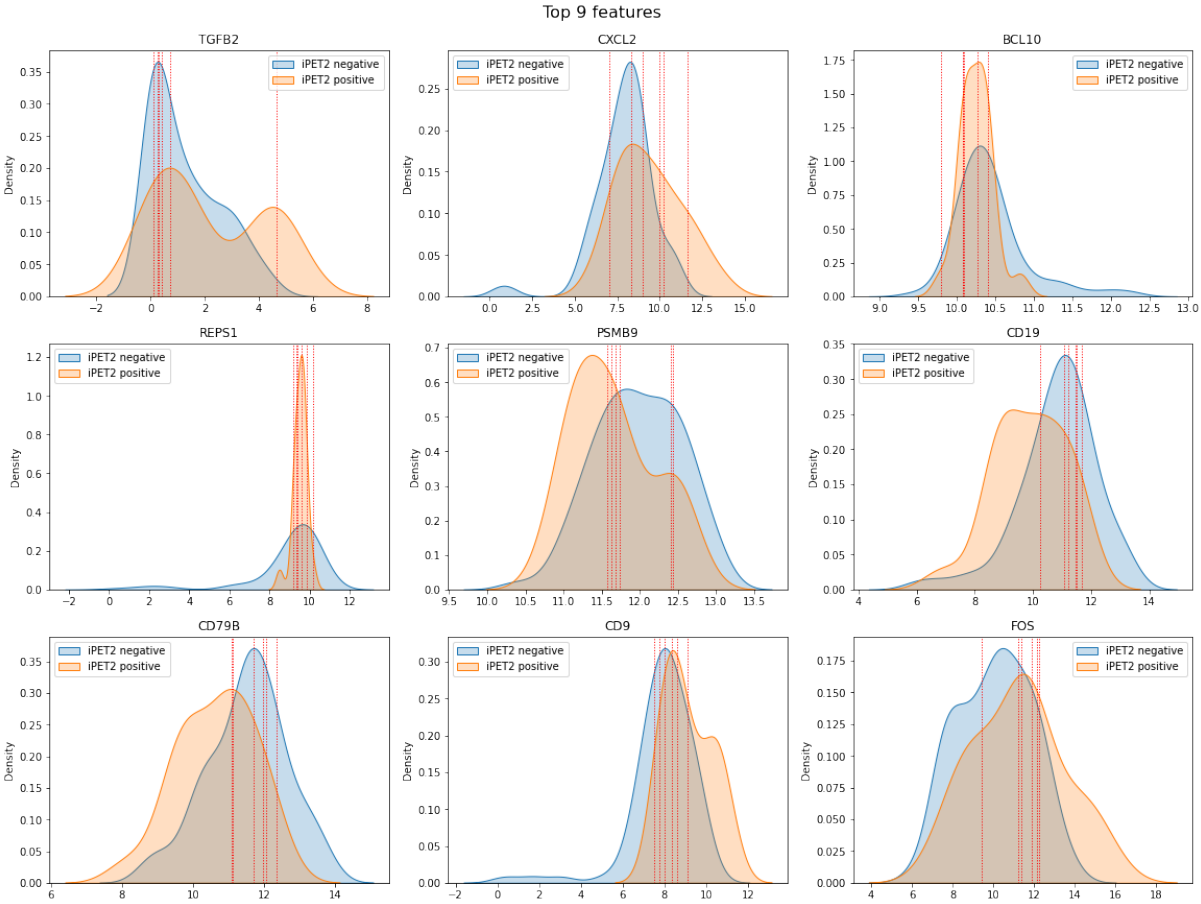

Figure 6.10: Distribution of top discriminative genes according to Mutual Information with highlighted values (red vertical lines)

6.3 Bicluster-based Space Transformation

The results presented until here correspond to the classification task in a feature space reduced by multiple feature selections. This approach is well suited to filter the most discriminative genes for the task at hand but lacks the ability to effectively represent the complex gene interactions responsible for the outcome of the patient. As explained in section 5.4, these interactions will be captured by the bicluster mining algorithm BicPAMS [64] and used to create new features representing discriminative and significant gene expression patterns. The remainder of this section presents an hyperparameterization analysis and the classification results obtained in the transformed feature space.

6.3.1 Hyperparameterization

With the goal of better understanding the impact that BicPAMS' multiple parameters have on classification performance, we present a comparative analysis of how the different parameterizations affect a Naive Bayes results according to multiple classification metrics. The graphics here plotted represent the variations of one parameter while the other ones are fixed. The default parameters in BicPAMS are: number of iterations = 3, minimum lift = 1.25, number of labels = 4, maximum number of biclusters = 100 and distance calculation by Euclidean distance. These parameters are explained in detail ahead.

The **number of iterations** indicates how many times the mining process is repeated. In each new iteration, the already discovered biclusters are masked in order to force the mining process to find other less trivial biclusters and, as a consequence, offer a more comprehensive coverage of the associations present in the dataset. If the value of this parameter is too high, it can result in biclusters that do not contain any useful information together with a steep computational cost. Figure 6.11 shows just this, with the increase of the number of iterations resulting in a better performance until a certain value is reached, from where there is no further advantage in increasing the number of iterations.

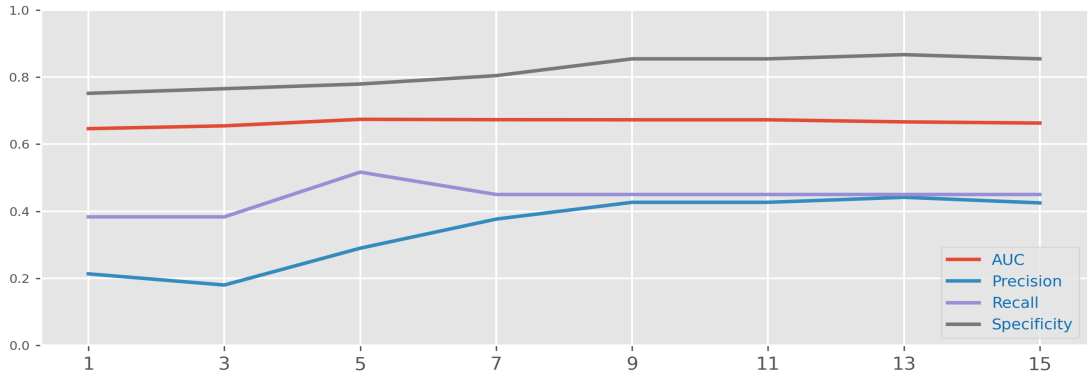

Figure 6.11: Variation of performance according to the number of iterations

The **minimum lift** is a placed threshold to determine whether a given bicluster is sufficiently discriminative. By increasing this value we force the found biclusters to be more discriminative, and consequently, more useful for the classification task, but just as with the number of iterations, it requires great computational power in order to mine enough biclusters that fulfil this condition. In addition, increasing the minimum lift results in an overall lower number of found patterns, possibly leaving out patterns that although not as discriminative, can still help in the learning process of a model. In figure 6.12 we can see that in this specific problem the increase of the minimum lift has almost no effect. These results point towards a solution with a higher number of iterations and low minimum lift, achieving an acceptable trade-off for computational power. There is also evidence pointing to the importance of less discriminative patterns in the learning process.

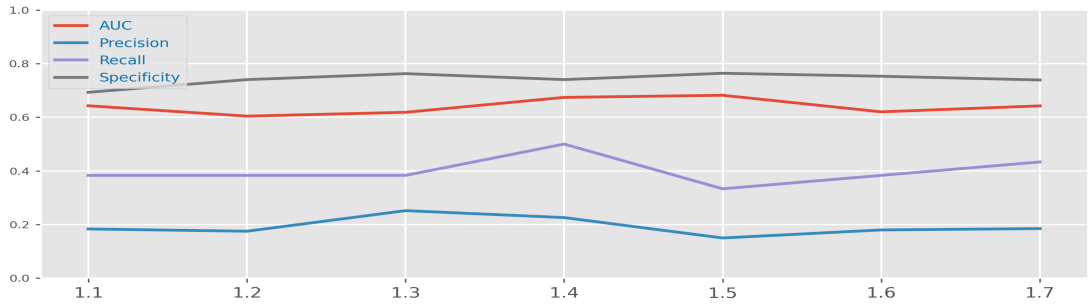

Figure 6.12: Variation of performance according to the minimum lift

Pattern-based biclustering mining generally requires data to be discretized. The **number of labels** corresponds to the number of intervals in which to discretize the values of each column. The higher this value, the closer two gene expression values need to be in order to belong to the same bicluster. Until a certain point, this increases the detail of the biclusters, but too high of a value will undoubtedly complicate the mining of discriminative biclusters. In figure 6.13, we can observe that an initial increase of this parameter results in better precision and recall since a low value does not allow for the correct discrimination between different gene expression values. It is also visible that there is no increase in performance after the eleven labels mark.

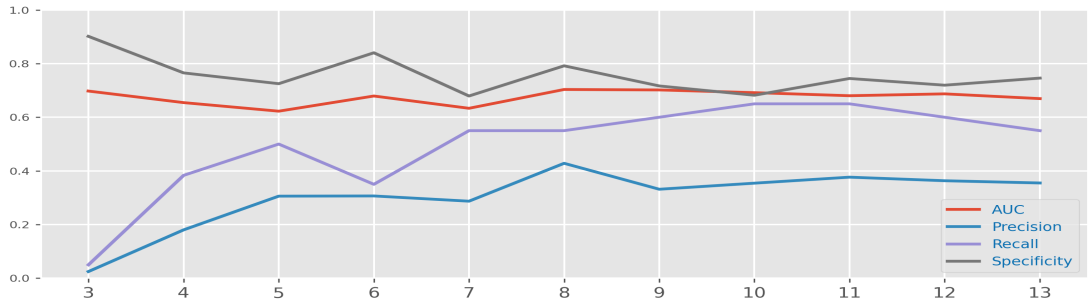

Figure 6.13: Variation of performance according to the number of labels

One of the conditions to stop the mining process is reaching a determined minimum number of biclusters, but with a high number of iterations the final number of biclusters found will be excessive. The **maximum number of biclusters** defines the number of mapped features in the transformed data space by the postprocessing filtering of the bottom discriminative biclusters according to their lift. Figure 6.14 presents the performance obtained with various values of this parameter, showing an increase until two hundred and fifty biclusters from where on are no further improvements.

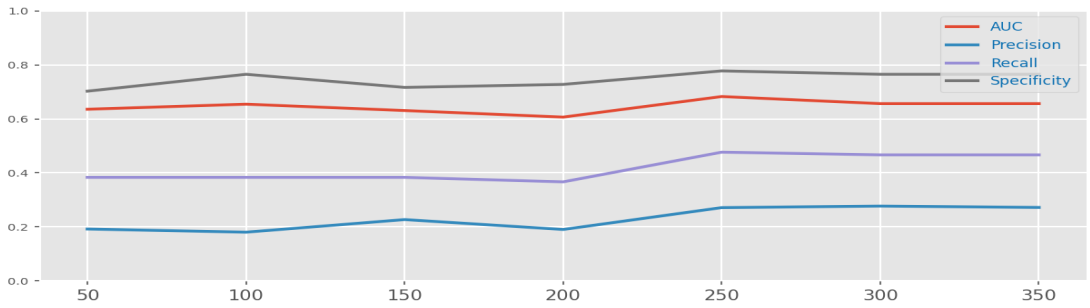

Figure 6.14: Variation of performance according to the maximum number of biclusters

Finally, a parameter exterior to the BicPAMS algorithm was also studied. This parameter corresponds to the **distance criterion** between a patient and a pattern's values, determining how the values present in the transformed dataset are computed. As explained in section 5.4, this distance can be calculated as the Euclidean distance or reduced to a binary value indicating if a patient possesses a given pattern or not. In the second case, a tolerance threshold can be included to accommodate for noise. Figure 6.15 shows the performance of a Naive Bayes classifier in three different settings, the binary transformation with thresholds of 0.5 and 1, and the transformation using Euclidean distance. As expected, by using a numeric representation instead of a binary one there is less loss of information, and consequently, we can achieve better results in the most difficult metrics for this classification task, precision and recall.

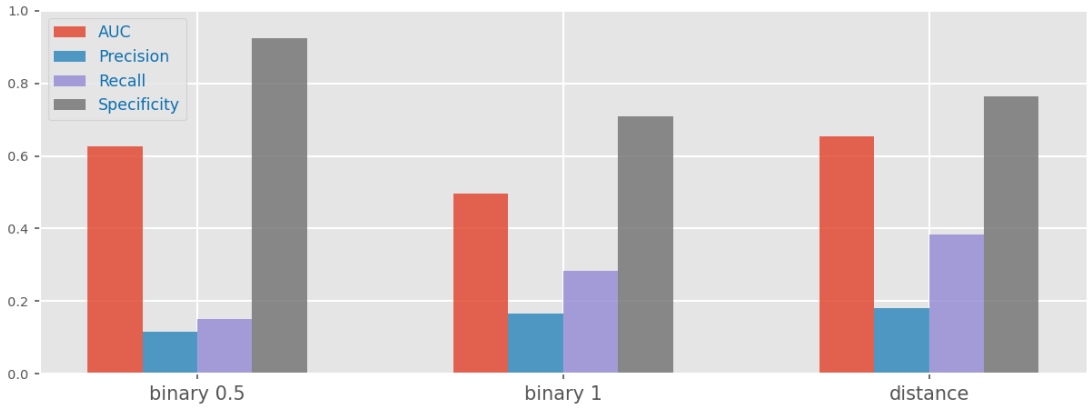

Figure 6.15: Variation of performance according to the distance criterion

6.3.2 Comparative Results

In order to execute a comparative analysis of the effects of pattern-based feature space transformation, we perform a transformation by means of biclustering using the following parameters decided according to empirical evidence: number of iterations = 9; minimum lift = 1.3; number of labels = 10; maximum number of biclusters = 250; and distance criterion = Euclidean distance. With this configuration, it is possible to carry out an effective exploration of the available biclusters by performing multiple iterations while maintaining a relatively low minimum lift so that the computational requirements do not get too high. The elevated number of labels guarantees that the found patterns discriminate fine levels of expression while the relatively high number of biclusters ensures that the most discriminative and significant biclusters are maintained. In figure 6.16 we can see the classification results obtained in this transformed space. Figure 6.17 provides a direct comparison for each individual metric between these results and the ones previously obtained by our second configuration of feature selection and evaluation.

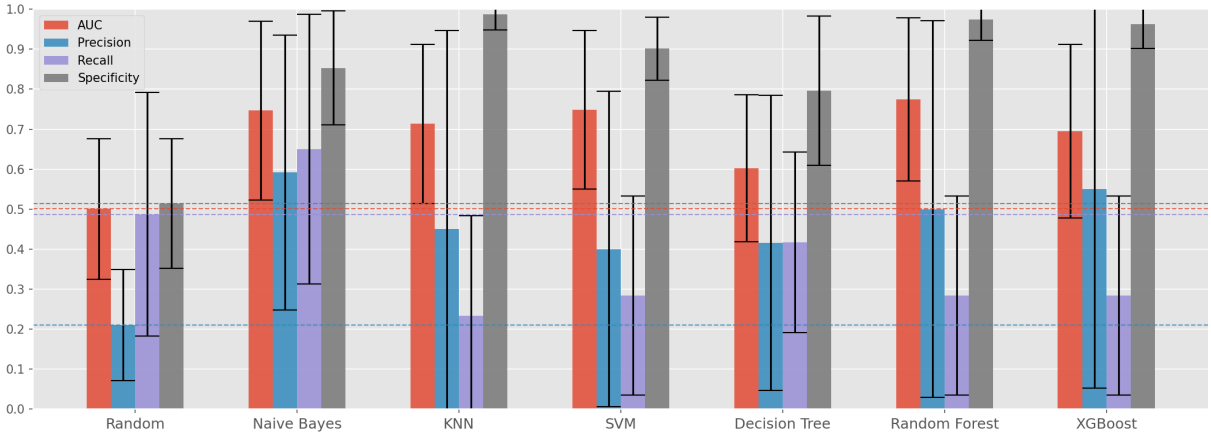

Figure 6.16: Optimized classifier's performance in the prediction of *iPET2* using pattern-based feature space

We can see that the transformed feature space consisting of the mined biclusters increases the classification results of the majority of the models (figure 6.17). The SVM and XGBoost classifiers that already presented good results are not as significantly affected by this transformation, but all the others benefit from it and present an increase in performance in all the studied metrics.

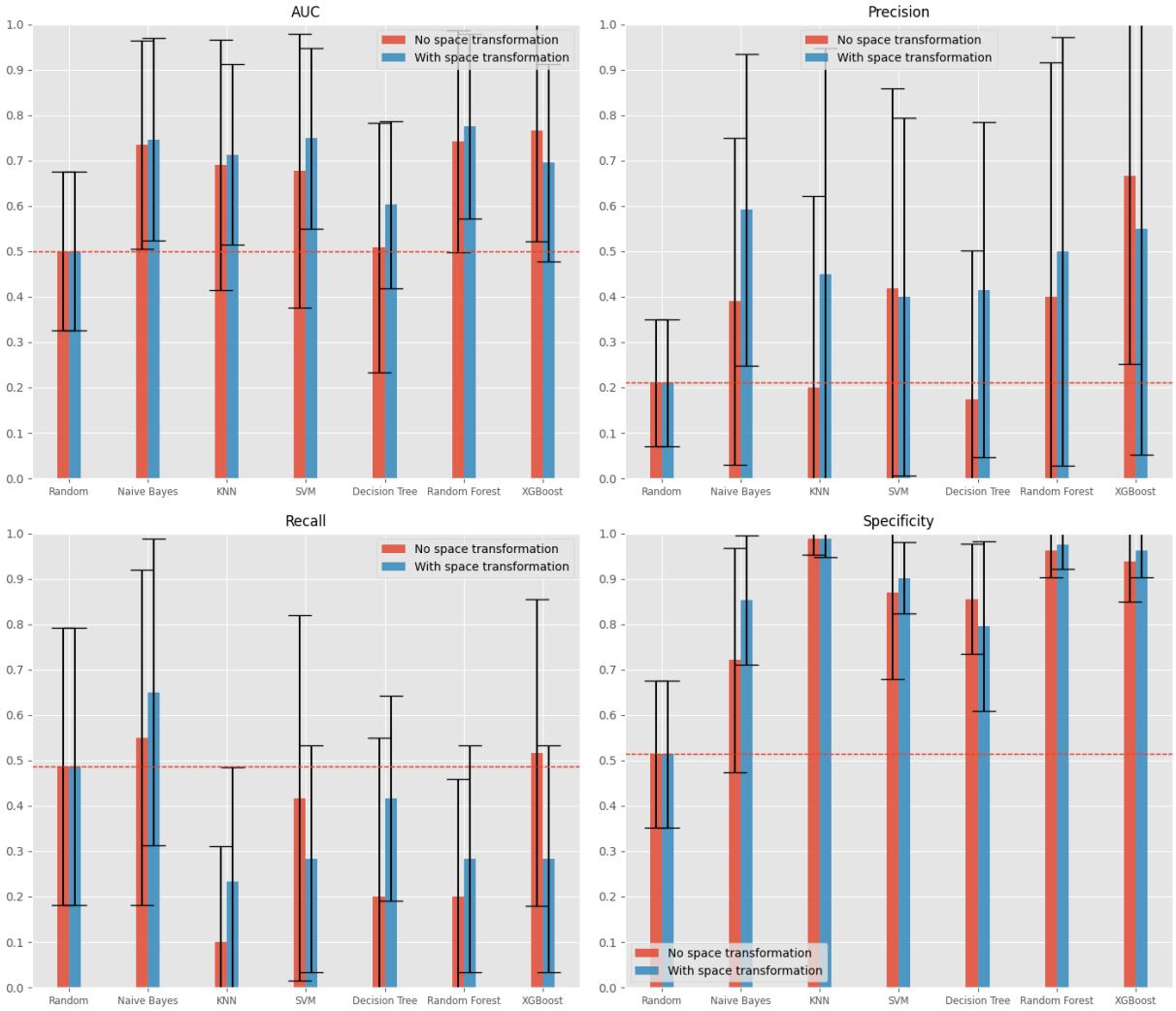

Figure 6.17: Direct comparison of results with and without bicluster-based space transformation

7

Biological Analysis

Contents

The second and final research problem to be tackled is the retrieval of potentially novel knowledge associated with the patient's response to ABVD chemotherapy treatment. To do this, the present chapter covers a biological-oriented analysis of the multiple gene sets discovered in past chapters and of the trained classification models. We start with a functional enrichment analysis of the gene sets returned by our feature selection algorithms, followed by the same procedure for some of the gene patterns identified by the bicluster-based space transformation and finish with the study of some of our classification models and the features defined as more important by them. Table A.2 (appendix A) provides a short description and corresponding Fold Change (FC) value for the genes mentioned in this chapter.

7.1 Feature Selection

After the first feature selection stage, we are left with 250 features, 248 of them genes, which translates in filtering out 517 genes. The second stage of feature selection resulted in 14 genes, a further reduction of 234 genes. Both these dimensionality reductions corroborate to produce a set of genes increasingly discriminative of the target variable *iPET2*. Therefore, we perform a functional enrichment analysis on both gene sets in order to understand how exactly do these genes affect the patient's response to treatment. This analysis will be performed using the Enrichr tool [80] as explained in section 5.5.

We start our analysis with the 248-gene set from our first feature selection. In figure 7.1 we can see the top ten enriched pathways in this set for the Kyoto Encyclopedia of Genes and Genomes (KEGG) knowledge base.

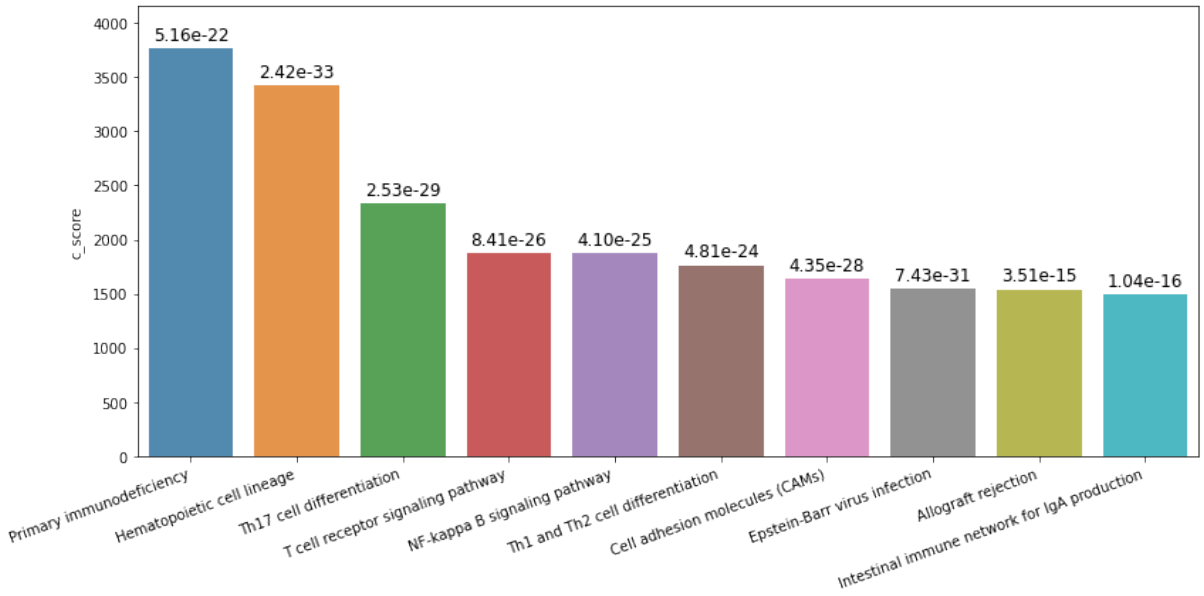

Figure 7.1: Enriched terms in the KEGG knowledge base over gene set obtained using the first phase of feature selection

It is worth noticing the presence of the already mentioned *NF-κB signalling* pathway as well as the *Epstein-Barr virus infection* pathway, confirming the previously stated influences of both in Hodgkin's Lymphoma.

The term with the highest score is *Primary immunodeficiency*, which comprises a set of disorders characterized by the deficient function of the immune system. It is important to make the distinction between this condition and secondary immunodeficiency, which originates from external factors such as malnutrition or infections. As previously stated, a weaker immune system has been correlated to a higher incidence of HL [85] and the gathered results suggest its influence on how a patient responds to treatment.

In figure 7.2 we can see the distributions of the genes overlapping with the *Primary immunodeficiency* pathway ordered by Fold-Change (FC) value. It is worth remembering that a high FC value indicates that a gene is more expressed in positive cases than in negatives ones, this is, patients with a positive *iPET2* have higher levels of expression of this gene in relation to patients with negative *iPET2*. A gene with high FC will then be linked to a worse reaction to treatment and a low FC value with a better reaction.

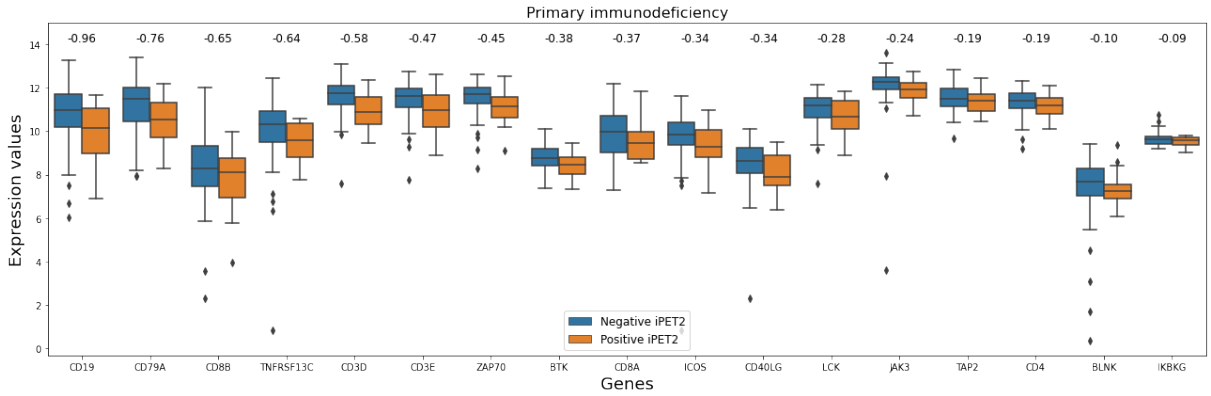

Figure 7.2: Boxplot representing the distributions of the genes enriched in the term Primary Immunodeficiency with corresponding fold change values on top

We can see through this boxplot that all the genes belonging to this pathway present a negative FC value, indicating a tendency for this pathway (response to primary immunodeficiency) to be elicited in patients with better treatment responses. The exact cause of this is not clear, it could be related to the fact that the majority of immunocompromised patients diagnosed with HL correspond to the Mixed-Cellularity subtype, but this subtype is often classified as the second worst prognosis [22, 48, 84] or just as good as other subtypes [13, 138]. We did not found any literature relating immunodeficiency and prognosis in HL, indicating that the enrichment of this term is probably novel and worthy of future research.

The second highest score term is related to haematopoiesis, the process from which Hematopoietic Stem Cells (HSCs) originate blood cells. These cells are multipotent and can replenish all blood type

cells, including the myeloid and the lymphoid cell lines. The first one corresponds to the differentiation of monocytes, macrophages, netrophils, basophils, eosinhophils, erythrocytes and platelets. The latter corresponds to T cells, B cells, natural killer cells and innate lymphoid cells. Understandably, the process from the initial multipotent to cell to the specialized blood cell is described by the *Hematopoietic cell lineage*.

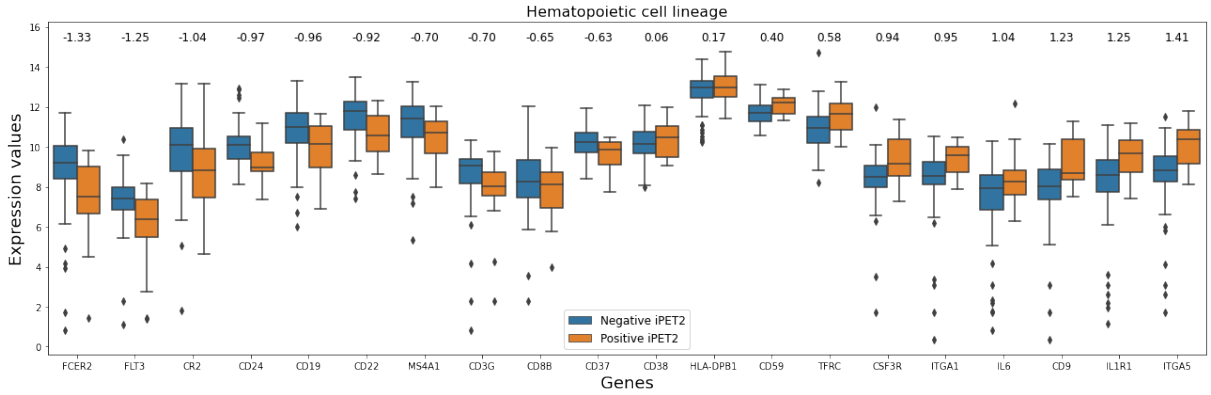

Figure 7.3: Boxplot representing the distributions of the genes enriched in the term Hematopoietic cell lineage with corresponding fold change values on top

In figure 7.3 we can see the distributions of the genes overlapping with the *Hematopoietic cell lineage* pathway ordered by FC value. On the left, we have the genes more expressed in negative cases and on the right the ones more expressed in positive cases. By analysis of this pathway, we can note an interesting trend, where the genes with the highest FC value (*CSF3R*, *ITGA1*, *IL6*, *CD9*, *IL1R1* and *ITGA5*) are all involved in the myeloid cell line, with the *IL1R1* and *CSF3R* genes participating in the development of neutrophils and the rest in the development of platelets. The lowest FC value genes (*FCER2*, *FLT3*, *CR2*, *CD24*, *CD19* and *CD22*) on the other end are all, with exception of the gene *FLT3*, present in the process of development of B cells. The results lead to conclude that the process of differentiation of B cells is in some way related to a good response to ABVD chemotherapy treatment while the differentiation of platelets has the inverse relation, indicating a bad response.

The contribution of platelets to the progression of cancer is a well studied subject. Platelets have a role in cancer metastases, angiogenesis and in protection from immune responses, as reviewed by Bambace and Holmes [9]. Platelets are also a considerable source of *VEGFA*, a major contributor to angiogenesis, the formation of new blood vessels which is fundamental in cancer growth. Our data shows that this gene has an FC value of 1, meaning that it is twice as expressed in bad reactions to treatment than in the other cases. Thrombocytosis, the overproduction of platelets, is widely observed in patients of various cancers [123], and a relation between elevated platelet count and bad prognosis has been found in a variety of cancers, such as gastric [72], colorectal [103], renal [132], uterine [54], ovarian [100] and cervical [93]. We have no knowledge of works that have found this relation in Hodgkin's

Lymphoma. Based on the results here presented we can hypothesize that this same relation is present in HL.

B cells play an important role in humoral immunity through the production of antibodies. Even though HL is characterized by Hodgkin and Reed-Sternberg (HRS) cells which are derived from B cells, it is possible that increased production of B cells can translate in a better response to treatment since these cells have been shown to be involved in various anti-tumor activities [24, 73, 78, 135]. The whole pathway as represented by the KEGG knowledge base is available in figure A.2 (appendix A).

The third term corresponds to the pathway adjacent to Interleukin (IL)-17-producing helper T (Th17) cell differentiation. The Th17 cell originates from CD4+ T cells and is characterized by the production of interleukin-17 (IL-17), a cytokine believed to be generally favorable to the growth of tumors [62, 143], normally being produced by adjacent tissue, stromal, and/or inflammatory cells, as explained by Tesmer et al. [136].

Changing to the Gene Ontology (GO) knowledge base we can study the biological processes in which these genes are involved. The corresponding enriched terms are displayed in figure 7.4 where we can see that our top 3 terms correspond respectively to the positive regulation of lymphocytes, T cells and cytokines proliferation. The HL microenvironment is rich in T helper cells [127] which release cytokines, and the disease is characterized by an accumulation of lymphocytes, explaining this way the presence of these enriched terms. We can also see the *T-helper 17 cell lineage commitment* term, which together with the previous results, points to an important role of this cell in discriminating the way a patient responds to treatment.

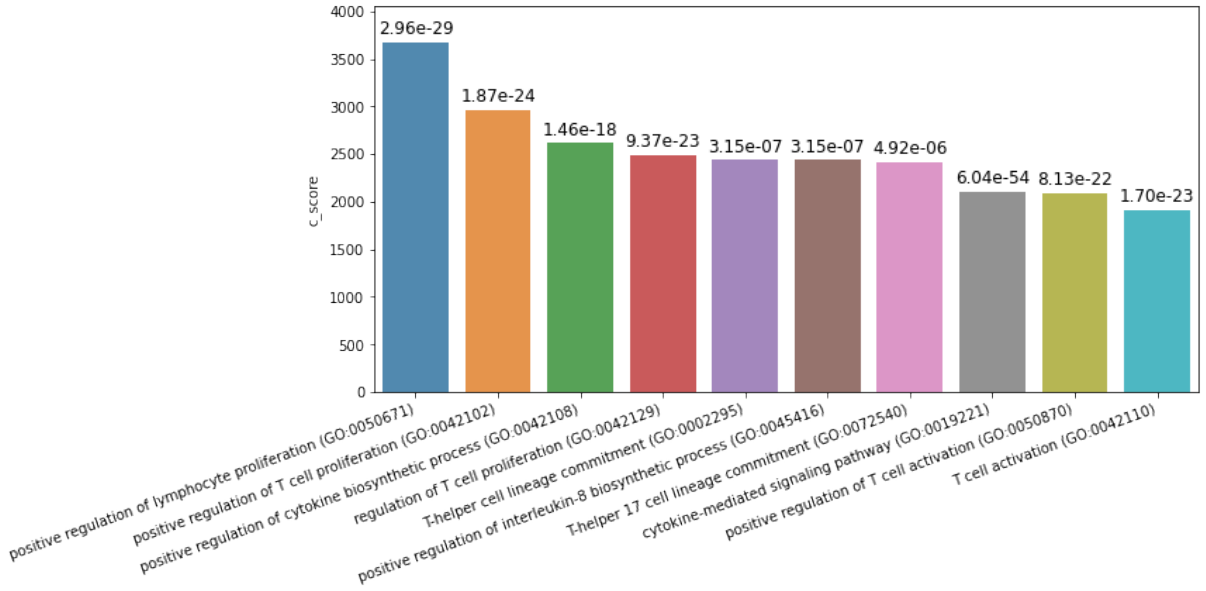

Figure 7.4: Enriched terms in the Gene Ontology Biological Process knowledge base over gene set obtained using the first phase of feature selection

We conclude the analysis of this gene set with the Online Mendelian Inheritance in Man (OMIM) Disease knowledge base, where the enriched terms will correspond to diseases in which these genes play important roles. These terms are displayed in figure 7.5, where we can observe the presence of the term *lymphoma* and the predominance of *protein C deficiency* and *immunodeficiency*. The reappearance of immunodeficiency indicates that this condition indeed has some relation with the development of HL. With respect to protein C deficiency, a disorder that increases a person's risk to develop abnormal blood clots, we were not able to find any direct relation between it and HL.

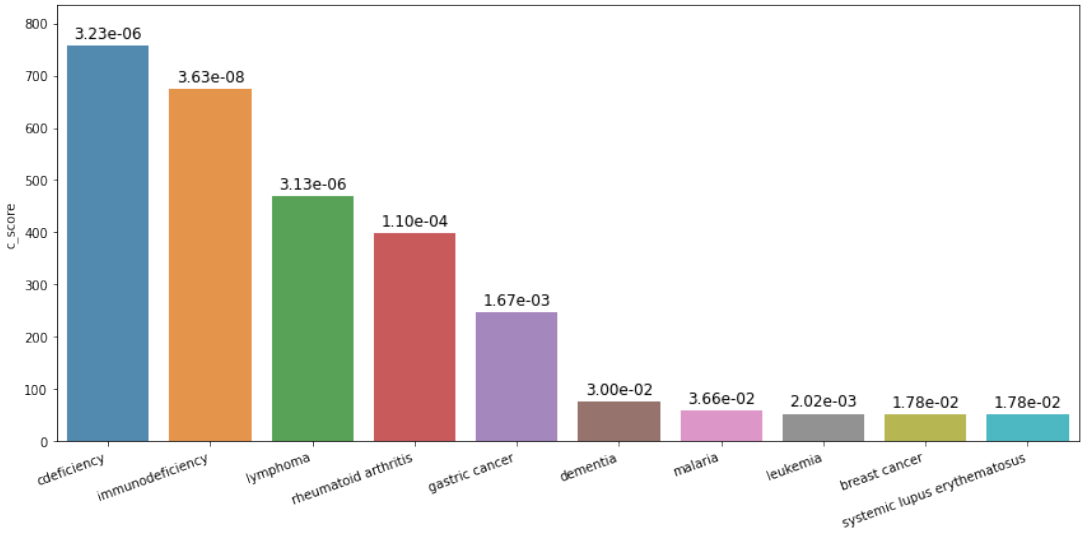

Figure 7.5: Enriched terms in the OMIM knowledge base over gene set obtained using the first phase of feature selection

Next, we submit the 14-gene set obtained through the second feature selection to a similar analysis. Starting once again with the KEGG Pathway knowledge base, the enriched terms are presented in figure 7.6. Only the terms *Allograft rejection* and *Cell adhesion molecules* coincide with the terms of the previous set, indicating that the second feature selection focuses on a specific set of genes with mostly different functions. An interesting trend is present in the top five enriched terms, with all of them corresponding to immune system related complications. More specifically, the top four all encompass some type of autoimmune response or graft rejection, and accordingly, all have the same three enriched genes, *IL2*, *HLA-C* and *CD80*.

Finally, 7.7 shows the enriched terms in the GO Biological Process knowledge base, revealing some overlap of the enriched terms with the previous gene set and the predominance of the *positive regulation of T-helper 1 type immune response* term. T-helper 1 type cells are responsible for the cell-mediated immune response, they produce Th1-type cytokines that produce proinflammatory responses responsible for killing intracellular parasites and for perpetuating autoimmune responses. T-helper 2 type cells on the other hand are responsible for the humoral immune response and produce Th2-type cytokines

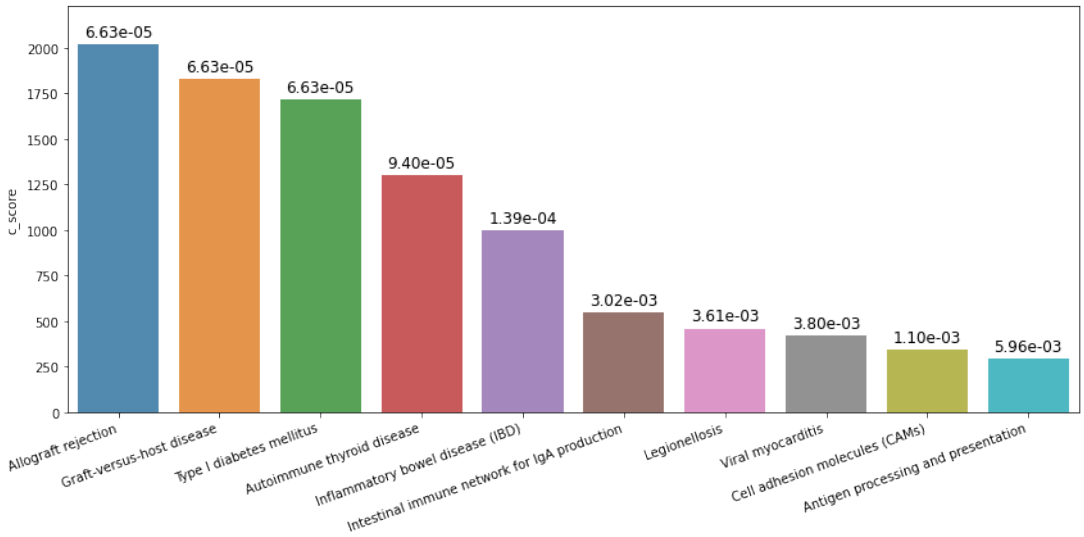

Figure 7.6: Enriched terms in the KEGG knowledge base over gene set obtained using the second phase of feature selection

that have an anti-inflammatory effect, counteracting eventual excessive autoimmune responses caused by Th1-type cytokines. Ideally, a balance of these two cytokines should exist in order for the immune system to correctly work. Multiple studies have identified an imbalance towards the predominance of Th2 response in multiple cancer patients [7, 28, 42, 115, 150], namely in non-Hodgkin's Lymphoma [32].

By reviewing our results, while considering the importance of these mechanisms, an important trend is noted. The condition of immunodeficiency, enriched in the first analyzed set, is characterized as the malfunction of the immune system, which can be originated through a Th1/Th2 imbalance. All the top four terms enriched in figure 7.6 correspond to conditions originated due to autoimmune responses or graft rejection, mechanisms mediated through the Th1-type response. The enriched biological processes are focused on the differentiation and proliferation of T cells, a trend normal in HL, which is activated and regulated again through Th1-type cytokines.

Serrano et al. [119] found that the Th1 and Th2 patterns on HL cancer patients and non-cancer reactive lymph nodes were similar with no distinctive bias towards one of them. Even if this is the case, differences in the Th1/Th2 responses of cancer patients seem to have a relation to how the patient will react to chemotherapy treatment. More specifically, since our feature selection methods have selected mainly genes involved in the Th1 response, it is possible that the balance between Th1 and Th2 responses is not as important as the absolute values of Th1 related genes. Even though a Th1 dominant response has been related to a good prognosis in lymphoma murine models [87], and its consequent inflammation to protection against B-cell lymphomas in mouses [57], the same conclusions about the good effects of Th1 responses cannot be drawn from our results, since there is no clear trend in the FC values of the genes involved in the related enriched terms.

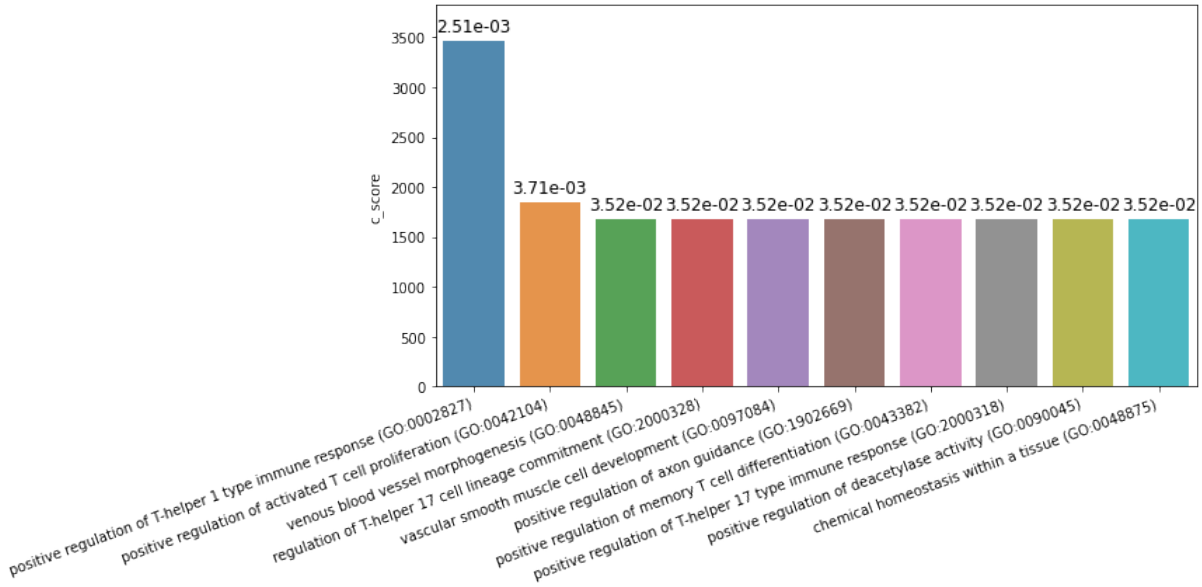

Figure 7.7: Enriched terms in the Gene Ontology Biological Process knowledge base over gene set obtained using the second phase of feature selection

7.2 Enrichment of Discriminative Gene Patterns

Since most biological mechanisms originate from multiple complex gene interactions, the analysis of gene expression patterns across multiple genes will certainly provide useful information. In this section, we provide an overview of some of the mined patterns using the previously mentioned parameterization for the BicPAMS algorithm. The analysis of the found patterns reveals a high variability of the resultant enriched terms. Next, we plot the enriched pathways for some of the found patterns with higher lift.

Some of the patterns present a compact number of distinctively enriched terms, such as the ones in figure 7.8 where the terms *B cell receptor signaling pathway* (left) and *Primary immunodeficiency* (right) present much higher values than the remaining ones. In these cases, the biclustering algorithm found patterns that are coherent with the biological knowledge available.

Other patterns are more well distributed while still presenting some consistency, with the majority of enriched terms being related in some way. This can be seen in figure 7.9 where the condition myeloid leukemia is the clear focus.

Finally, some patterns possess genes that do not present an obvious function but present a high lift value, and so, have high discriminative power, such as in figure 7.10. These sets of genes probably encompass more complex mechanisms, or groups of mechanisms, that are not as obvious when compared with the available knowledge bases.

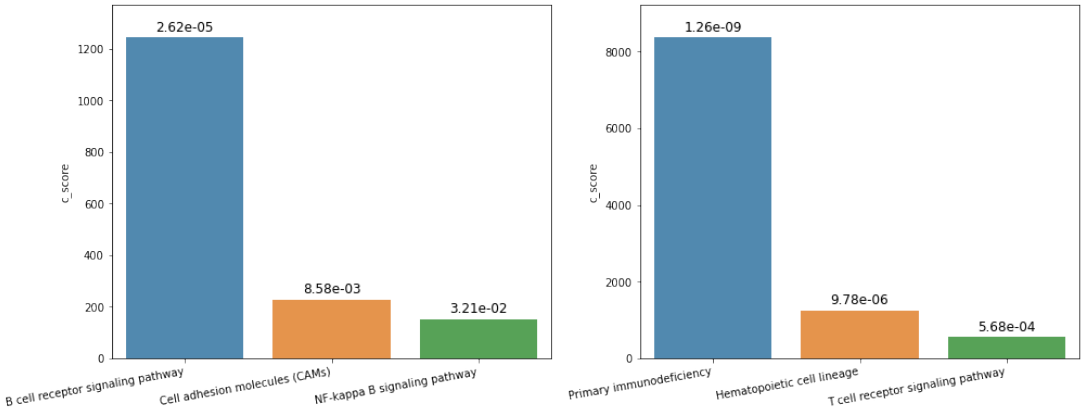

Figure 7.8: Enriched terms in patterns with identifiers 18 and 37 composed by the following genes and corresponding discretized expression values: (*SYK*; 7), (*IL1RN*; 8), (*MAGEB2*; 4), (*BTLA*; 1), (*CD22*; 0), (*CXCL2*; 9), (*LILRB3*; 9), (*ROPN1*; 7), (*JAM3*; 2), (*EIF2B4*; 3), (*ERCC3*; 3), (*CD19*; 1), (*TCF7*; 1), (*DNAJC14*; 1), (*CXCR5*; 0), (*PDGFRB*; 2), (*SELL*; 0) on the left, and (*ANP32B*; 2), (*C8G*; 1), (*CD3E*; 2), (*ITK*; 2), (*CD8A*; 2), (*POU2F2*; 2), (*C1R*; 9), (*BLNK*; 3), (*TNFRSF13C*; 1), (*KLRF1*; 3), (*CD19*; 0), (*CD180*; 5), (*ITGA5*; 9) on the right

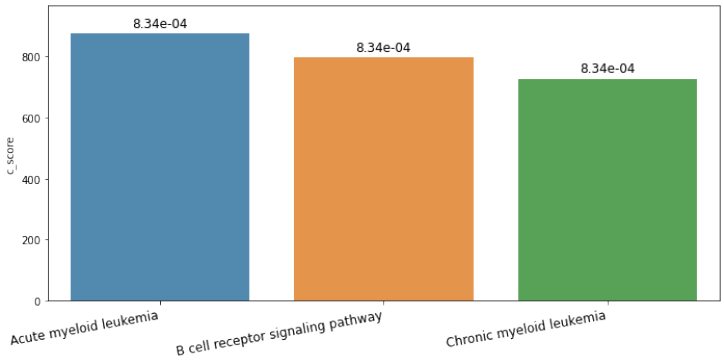

Figure 7.9: Enriched terms in pattern with identifier 2 composed by the following genes and corresponding discretized expression values: (*CD79B*; 3), (*G6PD*; 0), (*TXK*; 0), (*CASP1*; 2), (*CD28*; 2), (*ITGA1*; 7), (*CDH1*; 0), (*KLRK1*; 1), (*RUNX1*; 9), (*CDK1*; 1), (*TCF7*; 1), (*TGFB2*; 5), (*CXCR6*; 1), (*CHUK*; 7), (*CD79A*; 3)

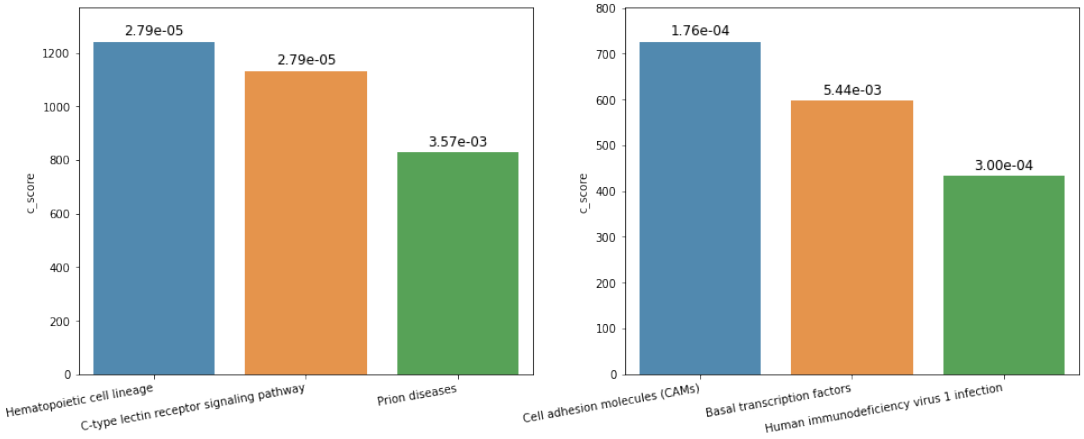

Figure 7.10: Enriched terms in patterns with identifiers 30 and 52 composed by the following genes and corresponding discretized expression values: (*C7*; 0), (*MAGEB2*; 9), (*CLEC7A*; 8), (*CD96*; 2), (*PECAM1*; 7), (*MERTK*; 9), (*IL1R1*; 9), (*CLEC6A*; 2), (*FLT3*; 0), (*EIF2B4*; 8), (*IL6*; 9), (*ITGA5*; 9), (*GNLY*; 1), (*IKBKB*; 8) on the left, and (*IRAK4*; 9), (*TLR8*; 6), (*HLA*-C, 3), (*FLT3*; 3), (*CD4*; 3), (*CNOT10*; 3), (*VEGFA*; 8), (*ERCC3*; 4), (*DHX16*; 3), (*ICAM3*; 6), (*CD99*; 7), (*TBP*; 4), (*MFGE8*; 5), (*NFATC2*; 8) on the right

7.3 Exploration of Predictive Models

In the analysis of the trained predictive models, we start with the best classifier in configuration two (section 6.2), XGBoost. In figure 7.11 we can see the top ten features ordered by their normalized impurity-based importance in the prediction by XGBoost.

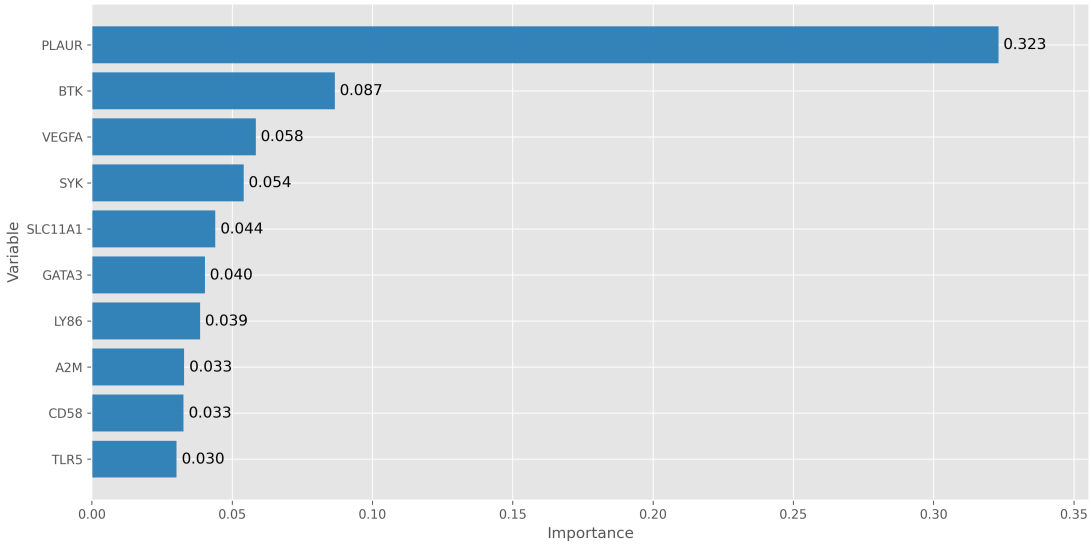

Figure 7.11: Feature importance in XGBoost's prediction

We can see that the gene *PLAUR* is considered focal, as it was with the Random Forest-based feature selection which only returned this gene. It is also highlighted as one of the top four most important features in the prediction by the Random Forest algorithm. To further confirm the importance of this gene in the decision by tree-based classifiers we plot a decision tree trained on our dataset, with each node presenting the distribution of the gene used for the split and the leafs displaying the distribution of the cases that ended up in it. This can be seen in figure 7.14, where we can confirm that the first feature chosen to split the data is exactly the gene *PLAUR*.

PLAUR stands for Plasminogen Activator Urokinase Receptor. It encodes the receptor for urokinase, an enzyme that among other functions is involved in extracellular matrix degradation through the activation of a proteolytic cascade which has been involved in cancer progression [134]. Urokinase has been further related to the facilitation of cancer metastasis [95] and identified as a possible prognostic biomarker in breast cancer [96]. The interaction between this enzyme and its receptor also plays an essential role in tumor invasion and metastasis [5]. *PLAUR* shows increased expression in malignant tumors of multiple cancers and is correlated with tumor recurrence and poor prognosis [44, 47, 56, 99, 102]. This increased expression was not found in Hodgkin's Lymphoma by Plesner et al. [111]. The importance given to this gene by our predictive models, especially by our best predictor XGBoost, points to it being fundamental in the discrimination of how a patient will respond to treatment.

More precisely, this gene appears to be more expressed in cases that do not react well to treatment, presenting an FC value of 0.89. The distribution of its values is presented in figure 7.12.

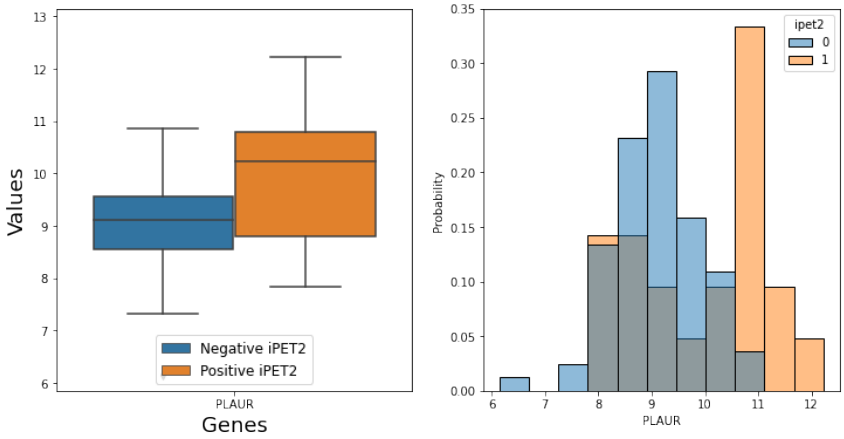

Figure 7.12: *PLAUR* distribution

The second most important gene according to the XGBoost classifier is *BTK*, a gene that is again found on the top ten important features of the Random Forest and used for the second split in the tree drawn in figure 7.14. *BTK* stands for Bruton Tyrosine Kinase, an enzyme fundamental in B-cell development, differentiation and signaling, with one of its related pathways being the already seen *B Cell Receptor Signaling Pathway*. The distribution of expression values for this gene is present in figure 7.13.

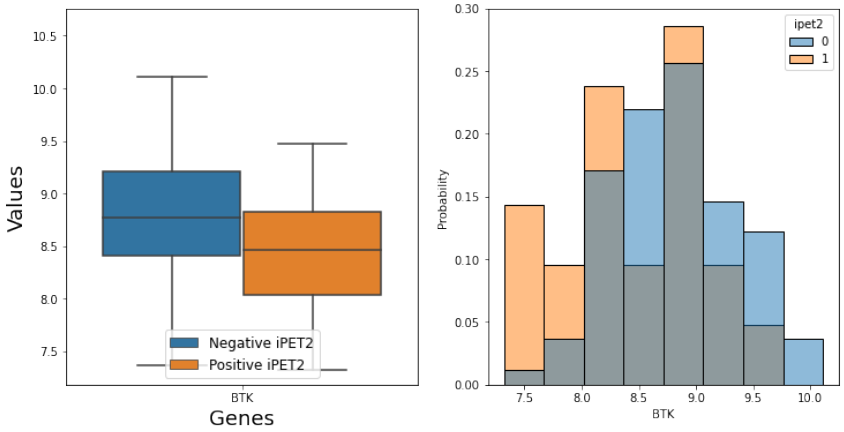

Figure 7.13: *BTK* distribution

An overall analysis of the enriched terms using associations extracted from the drawn tree (figure 7.14) indicates that there are no obvious relations between genes in a path from the root to a leaf (class-conditional pattern). The results here presented support the initial thesis that this is indeed a complex problem that cannot be answered by simple relations between small sets of genes.

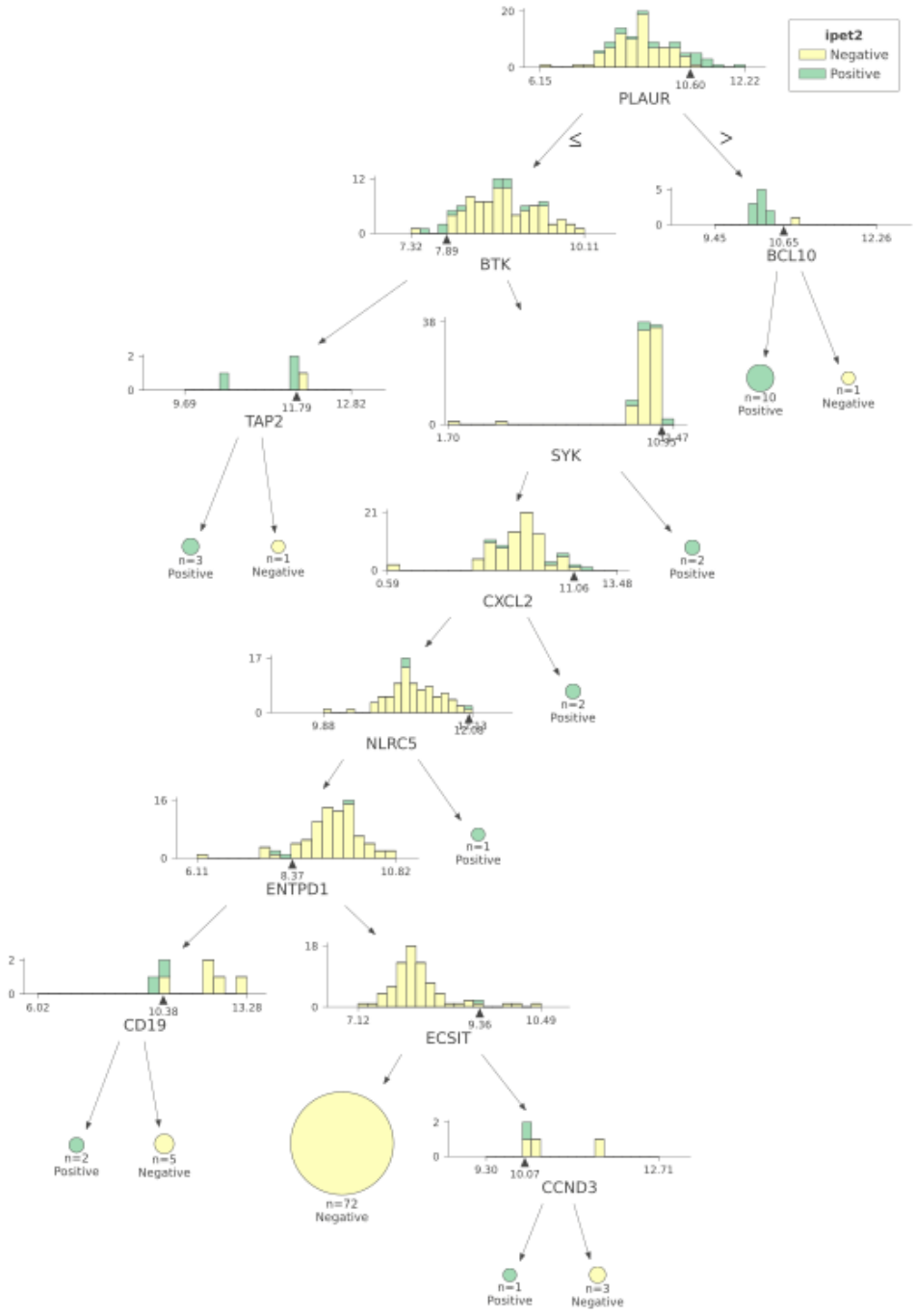

Figure 7.14: Trained decision tree

8

Conclusion

Contents

8.1 Concluding Remarks

The present dissertation proposes a methodology for the prediction of treatment response in Hodgkin's Lymphoma patients and presents a complementary analysis of the gene modules involved in the process. To achieve this, the developed work utilizes gene expression profiles obtained from diagnostic tumor samples in addition to clinical variables and an FDG-PET performed after two courses of ABVD chemotherapy. By predicting the result of an interim FDG-PET, it is possible to anticipate if an ABVD regimen will be sufficient, or if it is necessary to prescribe a stronger alternative. An accurate prediction can then be translated into less undergone toxicity by preventing stronger treatments when they are not necessary and advise them as a first option when alternatives would not be enough.

The methodology used for the prediction task encompasses multiple steps, combining dimensionality reduction procedures with state-of-the-art machine learning techniques in order to enable the best possible classification performance. The utilized data is cleaned, normalized and balanced using the SVM-SMOTE technique, creating optimal conditions for the learning process. To deal with the sizable feature space (770 features), we perform two separate stages of dimensionality reduction, the first relying on classic filter techniques, Mutual Information and Wilcoxon Rank Sum Test, and the second using the SVM-RFE algorithm to account for non-linear variable dependencies. Having the data optimally transformed, we proceed to use a nested Cross-Validation setting to both optimize the parameterization and test the predictive power of multiple elected machine learning models. This pipeline produces results superior to the previously obtained in the same data, with the optimal combination corresponding to the SVM-SMOTE and SVM-RFE algorithms and an optimized Support Vector Machine, resulting in an AUC of 0.93, a precision of 0.96, a recall of 0.83 and a specificity of 0.99. Further analysis of the classification results is also provided, leading to the conclusion that individuals of older age and with lymphocyte-to-monocyte ratio greater than 2.1 are more prone to be misclassified by our predictors.

A biclustering-based space transformation was further proposed and applied to the data in order to investigate how different mappings using the pattern-based biclustering algorithm BicPAMS impact classification performance. In essence, this transformation helps us move from the initial gene-centric space to a more discriminative space where the features represent discriminative modules defined by co-expressed genes. An in-depth analysis of multiple parameters resulted in an optimal configuration used to transform the data and then repeat the evaluation of the models' performance. The transformation from gene-based to pattern-based feature space resulted in increased performance according to all evaluated metrics in the majority of the models.

The gene sets obtained along the steps of the proposed methodology were analyzed resorting to functional enrichment analysis and the gene associations implied by the trained classifiers were compared against existing biological knowledge. The analysis of the enriched terms present in our various gene sets provided some insights into molecular mechanisms that are present in other cancers and can

aid in discriminating response to treatment in Hodgkin's Lymphoma. The role of Th1-type immunological responses and expression levels of the *PLAUR* gene are highlighted.

8.2 Future Work

The following directions are suggested as future work:

- Evaluate the potential increase in classification performance by:
  - using different, more specialized classifiers, i.e. predictive models better able to handle the inherent high-dimensionality and overlapping class-conditional distributions of expression per gene;
  - increasing the quantity of available data, in individuals to allow greater generalization capability, and in genes to include possibly important molecular mechanisms, namely regulatory ones performed by non-coding genes;
- Further test the impact of a pattern-based feature space on classification performance, namely by resorting to solutions requiring higher computational power and ensembles of biclustering algorithms with different characteristics;
- Provide subsequent research and validation of the molecular mechanisms identified as discriminative of refractory response to treatment;
- Expand the approach here presented to different treatment regimens;
- Strengthen this methodology by including complementary omics data, such as the genome, epigenome or proteome.

8.3 Scientific Communication

The principal contributions presented by this dissertation are currently submitted to bioinformatic journals of excellence, including BMC Bioinformatics, Cancers and Bioinformatics Oxford.

Complementarily, part of the methodological contributions here presented were applied to study the predictability of medical needs and survivability in COVID-19 infected patients. This contribution was submitted and accepted in the journal JMIR (Q1).

Patrício, A., Costa, R. S., Henriques, R. (2021). Predictability of COVID-19 Hospitalizations, Intensive Care Unit Admissions, and Respiratory Assistance in Portugal: Longitudinal Cohort Study. Journal of Medical Internet Research, 23(4), e26075.

Bibliography

[1] Aldin, A., Umlauff, L., Estcourt, L. J., Collins, G., Moons, K. G., Engert, A., Kobe, C., von Tresckow, B., Haque, M., Foroutan, F., et al. (2019). Interim pet-results for prognosis in adults with hodgkin lymphoma: a systematic review and meta-analysis of prognostic factor studies. *Cochrane Database of Systematic Reviews*, (9).

[2] Aleman, B. M., van den Belt-Dusebout, A. W., De Bruin, M. L., van't Veer, M. B., Baaijens, M. H., Boer, J. P. d., Hart, A. A., Klokman, W. J., Kuenen, M. A., Ouwens, G. M., et al. (2007a). Late cardiotoxicity after treatment for hodgkin lymphoma. *Blood*, 109(5):1878–1886.

[3] Aleman, B. M., van den Belt-Dusebout, A. W., Klokman, W. J., van't Veer, M. B., Bartelink, H., and van Leeuwen, F. E. (2007b). Long-term cause-specific mortality of patients treated for hodgkin's lymphoma. *Optimizing treatment of patients with Hodgkin's lymphoma*, 21:95.

[4] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

[5] Andreasen, P. A., Kjøller, L., Christensen, L., and Duffy, M. J. (1997). The urokinase-type plasminogen activator system in cancer metastasis: a review. *International journal of cancer*, 72(1):1–22.

[6] Ansell, S. M. (2014). Hodgkin lymphoma: 2014 update on diagnosis, risk-stratification, and management. *American Journal of Hematology*, 89(7):771–779.

[7] Ashkenazi, E., Deutsch, M., Tirosh, R., Weinreb, A., Tsukerman, A., and Brodie, C. (1997). A selective impairment of the il-2 system in lymphocytes of patients with glioblastomas: increased level of soluble il-2r and reduced protein tyrosine phosphorylation. *Neuroimmunomodulation*, 4(1):49–56.

[8] Bailey, D. L., Maisey, M. N., Townsend, D. W., and Valk, P. E. (2005). *Positron emission tomography, Chapter 16*, volume 2. Springer.

[9] Bambace, N. and Holmes, C. (2011). The platelet contribution to cancer progression. *Journal of thrombosis and haemostasis*, 9(2):237–249.

[10] Bargou, R. C., Emmerich, F., Krappmann, D., Bommert, K., Mapara, M. Y., Arnold, W., Royer, H. D., Grinstein, E., Greiner, A., Scheidereit, C., et al. (1997). Constitutive nuclear factor-kappab-rela activation is required for proliferation and survival of hodgkin's disease tumor cells. *The Journal of clinical investigation*, 100(12):2961–2969.

[11] Bea, S., Zettl, A., Wright, G., Salaverria, I., Jehn, P., Moreno, V., Burek, C., Ott, G., Puig, X., Yang, L., et al. (2005). Diffuse large b-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood*, 106(9):3183–3190.

[12] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

[13] Bennett, M. H., MacLennan, K. A., Hudson, B. V., and Hudson, G. V. (1989). The clinical and prognostic relevance of histopathologic classification in hodgkin's disease. In *Progress in Surgical Pathology*, pages 127–151. Springer.

[14] Bentham, R. B., Bryson, K., and Szabadkai, G. (2017). Mcbiclust: a novel algorithm to discover large-scale functionally related gene sets from massive transcriptomics data collections. *Nucleic acids research*, 45(15):8712–8730.

[15] Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *25th annual conference on neural information processing systems (NIPS 2011)*, volume 24. Neural Information Processing Systems Foundation.

[16] Biggar, R. J., Jaffe, E. S., Goedert, J. J., Chaturvedi, A., Pfeiffer, R., Engels, E. A., and Study, H. C. M. (2006). Hodgkin lymphoma and immunodeficiency in persons with hiv/aids. *Blood*, 108(12):3786–3791.

[17] Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135.

[18] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

[19] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

[20] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[21] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267.

[22] Butler, J. J. (1971). Relationship of histological findings to survival in hodgkin's disease. *Cancer research*, 31(11):1770–1775.

[23] Carbone, P. P., Kaplan, H. S., Musshoff, K., Smithers, D. W., and Tubiana, M. (1971). Report of the committee on hodgkin's disease staging classification. *Cancer research*, 31(11):1860–1861.

[24] Carmi, Y., Spitzer, M. H., Linde, I. L., Burt, B. M., Prestwood, T. R., Perlman, N., Davidson, M. G., Kenkel, J. A., Segal, E., Pusapati, G. V., et al. (2015). Allogeneic igg combined with dendritic cell stimuli induce antitumour t-cell immunity. *Nature*, 521(7550):99–104.

[25] Cesano, A. (2015). ncounter® pancancer immune profiling panel (nanostring technologies, inc., seattle, wa). *Journal for immunotherapy of cancer*, 3(1):1–3.

[26] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[27] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[28] Chen, Y.-M., Yang, W.-K., Ting, C.-C., Tsai, W.-Y., Yang, D.-M., Whang-Peng, J., and Perng, R.-P. (1997). Cross regulation by il-10 and il-2/il-12 of the helper t cells and the cytolytic activity of lymphocytes from malignant effusions of lung cancer patients. *Chest*, 112(4):960–966.

[29] Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Ismb*, volume 8, pages 93–103.

[30] Cheson, B. D., Fisher, R. I., Barrington, S. F., Cavalli, F., Schwartz, L. H., Zucca, E., and Lister, T. A. (2014). Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification. *Journal of clinical oncology*, 32(27):3059.

[31] Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261.

[32] Cortes, J. and Kurzrock, R. (1997). Interleukin-10 in non-hodgkin's lymphoma. *Leukemia & lymphoma*, 26(3-4):251–259.

[33] Devilard, E., Bertucci, F., Trempat, P., Bouabdallah, R., Loriod, B., Giaconia, A., Brousset, P., Granjeaud, S., Nguyen, C., Birnbaum, D., et al. (2002). Gene expression profiling defines molecular subtypes of classical hodgkin's disease. *Oncogene*, 21(19):3095–3102.

[34] Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1–13.

[35] Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.

[36] Dong, Z. and Chen, Y. (2013). Transcriptomics: advances and approaches. *Science China Life Sciences*, 56(10):960–967.

[37] Dores, G. M., Metayer, C., Curtis, R. E., Lynch, C. F., Clarke, E. A., Glimelius, B., Storm, H., Pukkala, E., Van Leeuwen, F. E., Holowaty, E. J., et al. (2002). Second malignant neoplasms among long-term survivors of hodgkin's disease: a population-based evaluation over 25 years. *Journal of clinical oncology*, 20(16):3484–3494.

[38] Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.

[39] Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pages 181–187.

[40] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[41] Eichenauer, D., Aleman, B., André, M., Federico, M., Hutchings, M., Illidge, T., Engert, A., and Ladetto, M. (2018). Hodgkin lymphoma: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 29:iv19–iv29.

[42] Elsässer-Beile, U., Kölble, N., Grussenmeyer, T., Schultze-Seemann, W., Wetterauer, U., Gallati, H., Mönting, J. S., and Von Kleist, S. (1998). Th1 and th2 cytokine response patterns in leukocyte cultures of patients with urinary bladder, renal cell and prostate carcinomas. *Tumor biology*, 19(6):470–476.

[43] Finotello, F. and Di Camillo, B. (2015). Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in functional genomics*, 14(2):130–142.

[44] Fisher, J., Field, C., Zhou, H., Harris, T., Henderson, M., and Choong, P. (2000). Urokinase plasminogen activator system gene expression is increased in human breast carcinoma and its bone metastases—a comparison of normal breast tissue, non-invasive and invasive carcinoma and osseous metastases. *Breast cancer research and treatment*, 61(1):1–12.

[45] Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

[46] Fix, E. (1951). *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine.

[47] Foekens, J. A., Peters, H. A., Look, M. P., Portengen, H., Schmitt, M., Kramer, M. D., Brünner, N., Jänicke, F., Meijer-van Gelder, M. E., Henzen-Logmans, S. C., et al. (2000). The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research*, 60(3):636–643.

[48] Franssila, K. O., Kalima, T. V., and Voutilainen, A. (1967). Histologic classification of hodgkin's disease. *Cancer*, 20(10):1594–1601.

[49] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[50] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914.

[51] Gallamini, A., Hutchings, M., Rigacci, L., Specht, L., Merli, F., Hansen, M., Patti, C., Loft, A., Di Raimondo, F., D'Amore, F., et al. (2007). Early interim 2-[18f] fluoro-2-deoxy-d-glucose positron emission tomography is prognostically superior to international prognostic score in advanced-stage hodgkin's lymphoma: a report from a joint italian-danish study. *Journal of clinical oncology*, 25(24):3746–3752.

[52] Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., et al. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3):317–325.

[53] Glaab, E., Bacardit, J., Garibaldi, J. M., and Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS one*, 7(7):e39932.

[54] Gücer, F., Moser, F., Tamussino, K., Reich, O., Haas, J., Arikan, G., Petru, E., and Winter, R. (1998). Thrombocytosis as a prognostic factor in endometrial carcinoma. *Gynecologic oncology*, 70(2):210–214.

[55] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

[56] Guyton, D. P., Evans, D. M., and Sloan-Stakleff, K. D. (2000). Urokinase plasminogen activator receptor (upar): a potential indicator of invasion for in situ breast cancer. *The breast journal*, 6(2):130–136.

[57] Haabeth, O. A. W., Lorvik, K. B., Hammarström, C., Donaldson, I. M., Haraldsen, G., Bogen, B., and Corthay, A. (2011). Inflammation driven by tumour-specific th1 cells protects against b-cell cancer. *Nature communications*, 2(1):1–12.

[58] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[59] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517.

[60] Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129.

[61] Hasenclever, D., Diehl, V., Armitage, J. O., Assouline, D., Björkholm, M., Brusamolino, E., Canellos, G. P., Carde, P., Crowther, D., Cunningham, D., et al. (1998). A prognostic score for advanced hodgkin's disease. *New England Journal of Medicine*, 339(21):1506–1514.

[62] He, D., Li, H., Yusuf, N., Elmets, C. A., Athar, M., Katiyar, S. K., and Xu, H. (2012). Il-17 mediated inflammation promotes tumor growth and progression in the skin. *PloS one*, 7(2):e32126.

[63] Heller, M. J. (2002). Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153.

[64] Henriques, R., Ferreira, F. L., and Madeira, S. C. (2017). Bicpams: software for biological data analysis with pattern-based biclustering. *BMC bioinformatics*, 18(1):1–16.

[65] Henriques, R. and Madeira, S. C. (2021). Flebic: Learning classifiers from high-dimensional biomedical data using discriminative biclusters with non-constant patterns. *Pattern Recognition*, 115:107900.

[66] Hinz, M., Lo¨ser, P., Mathas, S., Krappmann, D., Do¨rken, B., and Scheidereit, C. (2001). Constitutive nf-κb maintains high expression of a characteristic gene network, including cd40, cd86, and a set of antiapoptotic genes in hodgkin/reed-sternberg cells. *Blood*, 97(9):2798–2807.

[67] Hjalgrim, H., Askling, J., Rostgaard, K., Hamilton-Dutoit, S., Frisch, M., Zhang, J.-S., Madsen, M., Rosdahl, N., Konradsen, H. B., Storm, H. H., et al. (2003). Characteristics of hodgkin's lymphoma after infectious mononucleosis. *New England Journal of Medicine*, 349(14):1324–1332.

[68] Hoppe, R. (1997). Hodgkin's disease: complications of therapy and excess mortality. *Annals of oncology*, 8:S115–S118.

[69] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

[70] Howlader, N., Noone, A., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., et al. (2020). Seer cancer statistics review, 1975–2017. *National Cancer Institute*.

[71] Hutchings, M., Loft, A., Hansen, M., Pedersen, L. M., Buhl, T., Jurlander, J., Buus, S., Keiding, S., D'Amore, F., Boesen, A.-M., et al. (2006). Fdg-pet after two cycles of chemotherapy predicts treatment failure and progression-free survival in hodgkin lymphoma. *Blood*, 107(1):52–59.

[72] Ikeda, M., Furukawa, H., Imamura, H., Shimizu, J., Ishida, H., Masutani, S., Tatsuta, M., and Satomi, T. (2002). Poor prognosis associated with thrombocytosis in patients with gastric cancer. *Annals of surgical oncology*, 9(3):287–291.

[73] Jahrsdo¨rfer, B., Blackwell, S. E., Wooldridge, J. E., Huang, J., Andreski, M. W., Jacobus, L. S., Taylor, C. M., and Weiner, G. J. (2006). B-chronic lymphocytic leukemia cells and other b cells can produce granzyme b and gain cytotoxic potential after interleukin-21-based activation. *Blood*, 108(8):2712–2719.

[74] Jiao, Y. and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, 4(4):320–330.

[75] Johnson, N. T., Dhroso, A., Hughes, K. J., and Korkin, D. (2018). Biological classification with rna-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? *Rna*, 24(9):1119–1132.

[76] Johnson, P., Federico, M., Kirkwood, A., Fosså, A., Berkahn, L., Carella, A., d'Amore, F., Enblad, G., Franceschetto, A., Fulham, M., et al. (2016). Adapted treatment guided by interim pet-ct scan in advanced hodgkin's lymphoma. *N Engl J Med*, 374:2419–2429.

[77] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

[78] Kemp, T. J., Moore, J. M., and Griffith, T. S. (2004). Human b cells express functional trail/apo-2 ligand after cpg-containing oligodeoxynucleotide stimulation. *The Journal of Immunology*, 173(2):892–899.

[79] Kuhn, M. and Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

[80] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97.

[81] Küppers, R. (2009). The biology of hodgkin's lymphoma. *Nature Reviews Cancer*, 9(1):15–27.

[82] Küppers, R., Klein, U., Schwering, I., Distler, V., Bräuninger, A., Cattoretti, G., Tu, Y., Stolovitzky, G. A., Califano, A., Hansmann, M.-L., et al. (2003). Identification of hodgkin and reed-sternberg cell-specific genes by gene expression profiling. *The Journal of clinical investigation*, 111(4):529–537.

[83] Küppers, R. and Rajewsky, K. (1998). The origin of hodgkin and reed/sternberg cells in hodgkin's disease. *Annual review of immunology*, 16(1):471–493.

[84] Landberg, T. and Larsson, L.-E. (1969). Hodgkin's disease: Retrospective clinico-pathologic study in 149 patients. *Acta radiologica: therapy, physics, biology*, 8(5):390–414.

[85] Landgren, O., Engels, E. A., Pfeiffer, R. M., Gridley, G., Mellemkjaer, L., Olsen, J. H., Kerstann, K. F., Wheeler, W., Hemminki, K., Linet, M. S., et al. (2006). Autoimmunity and susceptibility to hodgkin lymphoma: a population-based case–control study in scandinavia. *Journal of the National Cancer Institute*, 98(18):1321–1330.

[86] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., and Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1106–1119.

[87] Lee, P. P., Zeng, D., McCaulay, A. E., Chen, Y.-F., Geiler, C., Umetsu, D. T., and Chao, N. J. (1997). T helper 2-dominant antilymphoma immune response is associated with fatal outcome. *Blood, The Journal of the American Society of Hematology*, 90(4):1611–1617.

[88] Lee, Y. and Lee, C.-K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9):1132–1139.

[89] Lenz, G., Wright, G., Dave, S., Xiao, W., Powell, J., Zhao, H., Xu, W., Tan, B., Goldschmidt, N., Iqbal, J., et al. (2008). Stromal gene signatures in large-b-cell lymphomas. *New England Journal of Medicine*, 359(22):2313–2323.

[90] Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer.

[91] Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437.

[92] Liu, H., Li, J., and Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics*, 13:51–60.

[93] Lopes, A., Daras, V., Cross, P. A., Robertson, G., Beynon, G., and Monaghan, J. M. (1994). Thrombocytosis as a prognostic factor in women with cervical cancer. *Cancer*, 74(1):90–92.

[94] Luminari, S., Donati, B., Casali, M., Valli, R., Santi, R., Puccini, B., Kovalchuk, S., Ruffini, A., Fama, A., Berti, V., et al. (2020). A gene expression–based model to predict metabolic response after two courses of abvd in hodgkin lymphoma patients. *Clinical Cancer Research*, 26(2):373–383.

[95] Madunić, J. (2018). The urokinase plasminogen activator system in human cancers: an overview of its prognostic and predictive role. *Thrombosis and haemostasis*, 118(12):2020–2036.

[96] Mahmood, N., Mihalcioiu, C., and Rabbani, S. A. (2018). Multifaceted role of the urokinase-type plasminogen activator (upa) and its receptor (upar): diagnostic, prognostic, and therapeutic applications. *Frontiers in oncology*, 8:24.

[97] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

[98] Meignan, M., Gallamini, A., Meignan, M., Gallamini, A., and Haioun, C. (2009). Report on the first international workshop on interim-pet scan in lymphoma. *Leukemia & lymphoma*, 50(8):1257–1260.

[99] Memarzadeh, S., Kozak, K. R., Chang, L., Natarajan, S., Shintaku, P., Reddy, S. T., and Farias-Eisner, R. (2002). Urokinase plasminogen activator receptor: prognostic biomarker for endometrial cancer. *Proceedings of the National Academy of Sciences*, 99(16):10647–10652.

[100] Menczer, J., Schejter, E., Geva, D., Ginath, S., and Zakut, H. (1998). Ovarian carcinoma associated thrombocytosis. correlation with prognostic factors and with survival. *European journal of gynaecological oncology*, 19(1):82–84.

[101] Moccia, M. A., Donaldson, J., Chhanabhai, M., Hoskins, P., Klasa, R., Savage, K. J., Shenkier, T., Skinnider, B., Gascoyne, R. D., Connors, J. M., et al. (2009). The international prognostic factor project score (ips) in advanced stage hodgkin lymphoma has limited utility in patients treated in the modern era.

[102] Mohanam, S., Gladson, C. L., Rao, C. N., and Rao, J. S. (1999). Biological significance of the expression of urokinase-type plasminogen activator receptors (upars) in brain tumors. *Front Biosci*, 4:D178–D187.

[103] Monreal, M., Fernandez-Llamazares, J., Piñol, M., Julian, J. F., Broggi, M., Abad, A., et al. (1998). Platelet count and survival in patients with colorectal cancer–a preliminary study. *Thrombosis and haemostasis*, 79(05):916–918.

[104] Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., Tamori, A., Kawada, N., Taguchi, Y., and Azuma, T. (2014). Comparison of hepatocellular carcinoma mirna expression profiling as evaluated by next generation sequencing and microarray. *PloS one*, 9(9):e106314.

[105] Nguyen, H. M., Cooper, E. W., and Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21.

[106] of Weights, I. B., Measures, Taylor, B. N., and Thompson, A. (2001). *The international system of units (SI) – Chapter*. US Department of Commerce, Technology Administration, National Institute of ....

[107] Ozsolak, F. and Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98.

[108] Pandey, G., Pandey, O. P., Rogers, A. J., Ahsen, M. E., Hoffman, G. E., Raby, B. A., Weiss, S. T., Schadt, E. E., and Bunyavanich, S. (2018). A nasal brush-based classifier of asthma identified by machine learning analysis of nasal rna sequence data. *Scientific reports*, 8(1):1–15.

[109] Parry, R., Jones, W., Stokes, T., Phan, J., Moffitt, R., Fang, H., Shi, L., Oberthuer, A., Fischer, M., Tong, W., et al. (2010). k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The pharmacogenomics journal*, 10(4):292–309.

[110] Pirooznia, M., Yang, J. Y., Yang, M. Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, 9(1):1–13.

[111] Plesner, T., Ralfkiær, E., Wittrup, M., Johnsen, H., Pyke, C., Pedersen, T. L., Hansen, N. E., and Danø, K. (1994). Expression of the receptor for urokinasetype plasminogen activator in normal and neoplastic blood cells and hematopoietic tissue. *American journal of clinical pathology*, 102(6):835–841.

[112] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[113] Ramroach, S., John, M., and Joshi, A. (2019). The efficacy of various machine learning models for multi-class classification of rna-seq expression data. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 918–928. Springer.

[114] Reis, P. P., Waldron, L., Goswami, R. S., Xu, W., Xuan, Y., Perez-Ordonez, B., Gullane, P., Irish, J., Jurisica, I., and Kamel-Reid, S. (2011). mrna transcript quantification in archival samples using multiplexed, color-coded probes. *BMC biotechnology*, 11(1):46.

[115] Rosen, H. R., Ausch, C., Reinerova, M., Zaspin, E., Renner, K., Rosen, A. C., Schiessel, R., and Moroz, C. (1998). Activated lymphocytes from breast cancer patients express the characteristics of type 2 helper cells–a possible role for breast cancer-associated p43. *Cancer letters*, 127(1-2):129–134.

[116] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.

[117] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

[118] Scott, D. W., Chan, F. C., Hong, F., Rogic, S., Tan, K. L., Meissner, B., Ben-Neriah, S., Boyle, M., Kridel, R., Telenius, A., et al. (2013). Gene expression–based model using formalin-fixed paraffin-embedded biopsies predicts overall survival in advanced-stage classical hodgkin lymphoma. *Journal of Clinical Oncology*, 31(6):692.

[119] Serrano, D., Ghiotto, F., Roncella, S., Airoldi, I., Cutrona, G., Truini, M., Burgio, V. L., Baroni, C. D., Ferrarini, M., and Pistoia, V. (1997). The patterns of il2, ifn-gamma, il4 and il5 gene expression in hodgkin's disease and reactive lymph nodes are similar. *Haematologica*, 82(5):542–549.

[120] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

[121] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[122] Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., et al. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74.

[123] Sierko, E. and Wojtukiewicz, M. Z. (2004). Platelets and angiogenesis in malignancy. In *Seminars in thrombosis and hemostasis*, volume 30, pages 95–108. Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New ....

[124] Somorjai, R. L., Dolenko, B., and Baumgartner, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19(12):1484–1491.

[125] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

[126] Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, 100(18):10393–10398.

[127] Steidl, C., Connors, J. M., and Gascoyne, R. D. (2011). Molecular pathogenesis of hodgkin's lymphoma: increasing evidence of the importance of the microenvironment. *Journal of Clinical Oncology*, 29(14):1812–1826.

[128] Steidl, C., Lee, T., Shah, S. P., Farinha, P., Han, G., Nayar, T., Delaney, A., Jones, S. J., Iqbal, J., Weisenburger, D. D., et al. (2010). Tumor-associated macrophages and survival in classic hodgkin's lymphoma. *New England Journal of Medicine*, 362(10):875–885.

[129] Su, Z., Li, Z., Chen, T., Li, Q.-Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R. G., et al. (2011). Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chemical research in toxicology*, 24(9):1486–1493.

[130] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[131] Swerdlow, S. H. (2008). Who classification of tumours of haematopoietic and lymphoid tissues. *WHO classification of tumours*, 22008:439.

[132] Symbas, N. P., Townsend, M. F., El-Galley, R., Keane, T. E., Graham, S., and Petros, J. (2000). Poor prognosis associated with thrombocytosis in patients with renal cell carcinoma. *BJU international*, 86(3):203–207.

[133] Tan, A. C. and Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.

[134] Tang, L. and Han, X. (2013). The urokinase plasminogen activator system in breast cancer invasion and metastasis. *Biomedicine & Pharmacotherapy*, 67(2):179–182.

[135] Tao, H., Lu, L., Xia, Y., Dai, F., Wang, Y., Bao, Y., Lundy, S. K., Ito, F., Pan, Q., Zhang, X., et al. (2015). Antitumor effector b cells directly kill tumor cells via the fas/fasl pathway and are regulated by il-10. *European journal of immunology*, 45(4):999–1009.

[136] Tesmer, L. A., Lundy, S. K., Sarkar, S., and Fox, D. A. (2008). Th17 cells in human disease. *Immunological reviews*, 223(1):87–113.

[137] Townsend, W. and Linch, D. (2012). Hodgkin's lymphoma in adults. *The Lancet*, 380(9844):836–847.

[138] Tubiana, M., Attie, E., Flamant, R., Gérard-Marchant, R., and Hayat, M. (1971). Prognostic factors in 454 cases of hodgkin's disease. *Cancer research*, 31(11):1801–1810.

[139] Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426.

[140] Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.

[141] Verboom, P., van Tinteren, H., Hoekstra, O. S., Smit, E. F., Van Den Bergh, J. H., Schreurs, A. J., Stallaert, R. A., van Velthoven, P. C., Comans, E. F., Diepenhorst, F. W., et al. (2003). Cost-effectiveness of fdg-pet in staging non-small cell lung cancer: the plus study. *European journal of nuclear medicine and molecular imaging*, 30(11):1444–1449.

[142] Wang, L., Xi, Y., Sung, S., and Qiao, H. (2018). Rna-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC genomics*, 19(1):546.

[143] Wang, L., Yi, T., Zhang, W., Pardoll, D. M., and Yu, H. (2010). Il-17 enhances tumor development in carcinogen-induced skin cancer. *Cancer research*, 70(24):10112–10120.

[144] Warburg, O. (1956). On the origin of cancer cells. *Science*, 123(3191):309–314.

[145] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[146] Weniger, M. A. and Küppers, R. (2016). Nf-κb deregulation in hodgkin lymphoma. In *Seminars in cancer biology*, volume 39, pages 32–39. Elsevier.

[147] Willenbrock, H., Salomon, J., Søkilde, R., Barken, K. B., Hansen, T. N., Nielsen, F. C., Møller, S., and Litman, T. (2009). Quantitative mirna expression analysis: comparing microarrays with next-generation sequencing. *Rna*, 15(11):2028–2034.

[148] Williams, A. and Halappanavar, S. (2015). Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. *Beilstein journal of nanotechnology*, 6(1):2438–2448.

[149] Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., and Staudt, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma. *Proceedings of the National Academy of Sciences*, 100(17):9991–9996.

[150] Yamamura, M., Modlin, R. L., Ohmen, J. D., Moy, R. L., et al. (1993). Local expression of antiinflammatory cytokines in cancer. *The Journal of clinical investigation*, 91(3):1005–1010.

[151] Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell*, 1(2):133–143.

Appendix

Table A.1: Parameters subjected to optimization

Model	Optimized Parameters
Naive Bayes[1]	NA
K-Nearest Neighbors[2]	Number of neighbors; Feature Weights;
Support Vector Machine[3]	C; Shrinking; Kernel; Gamma;
Decision Tree[4]	Split criterion; Strategy to split; Maximum features; Maximum depth; Minimum samples per split; Minimum samples per leaf; Minimal cost-complexity pruning alpha;
Random Forest[5]	Number of estimators; Split criterion; Maximum depth; Minimum samples per split; Minimum samples per leaf; Minimal cost-complexity pruning alpha;
XGBoost[6]	Learning rate; Booster; L1 regularization term; Number of estimators;

Further information about each parameter can be found at each model's corresponding link.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[6] https://xgboost.readthedocs.io/en/latest/python/python_api.html

Table A.2: Gene description

Gene Name	FC Value	Description
CD19	-1.33	Encodes a transmembrane protein expressed only on B cell lymphocytes. Associated diseases include Common Variable Immunodeficiency. GO annotations include obsolete signal transducer activity.
CD22	-1.04	Encodes a regulatory molecule that helps in the prevention of autoimmune diseases. ediates B-cell B-cell interactions. Associated diseases include Refractory Hematologic Cancer and Hairy Cell Leukemia. GO annotations include carbohydrate binding.
CD24	-0.97	Encodes a sialoglycoprotein expressed at the surface of B cells, differentiating neuroblasts and neutrophils. Associated with bile duct cancer and multiple sclerosis. GO annotations include protein kinase binding and carbohydrate binding.
CD80	-0.96	Encodes a membrane receptor that when activated by CD28 induces T-cell proliferation and cytokine production. Activation by CTLA-4 has the inverse effect. GO annotations include coreceptor activity.
CD9	-0.92	Encodes a cell surface glycoprotein. Involved in cell differentiation, adhesion, and signal transduction. Expression of this gene plays a critical role in the suppression of cancer cell motility and metastasis. GO annotations include integrin binding.
CR2	-0.71	Encodes a membrane protein that functions as a receptor for Epstein-Barr virus (EBV) binding on B and T lymphocytes. Associated with immunodeficiency. GO annotations include ransmembrane signaling receptor activity.
CSF3R	-0.21	Encodes a cytokine receptor for Colony Stimulating Factor 3. Involved in the proliferation, differientation and survival of cells along the neutrophilic lineage. GO annotations include cytokine recpetor activity.
FCER2	0.4	Encodes a B-cell specific antigen. Involved in B cell growth and differentiation, and the regulation of Immunoglobulin E (IgE) production. GO annotations include carbohydrate binding and IgE binding.
HLA-C	0.94	Encodes a class I heavy chain receptors, a molecule with an important role in reproduction and antiviral immunity. Associated diseases include include Psoriasis 1 and Human Immunodeficiency Virus Type 1. GO annotations include signaling receptor binding and TAP binding.
IL1R1	0.95	Encodes a cytokine receptor for interleukin-1. Involved in many cytokine-induced immune and inflammatory responses. Mediates interleukin-1-dependent activation of NF-kappa-B, MAPK and ERK signalling pathways.
IL2	0.99	Encodes a cytokine produced by activated CD4+ and CD8+ T lymphocytes, important for the proliferation of T and B lymphocytes. Involved in cell-mediated immunity. GO annotations include carbohydrate binding and growth factor activity.
IL6	1.04	Encodes a cytokine that functions in inflammation and the maturation of B cells. Required to drive naive CD4(+) T cells to the Th17 lineage. The functioning of this gene is implicated in a wide variety of inflammation-associated disease states, including suspectibility to diabetes mellitus and systemic juvenile rheumatoid arthritis. GO annotations include signaling receptor binding and growth factor activity.
ITGA1	1.23	Encodes a laminin and collagen receptor. Involved in cell-cell adhesion and may play a role in inflammation and fibrosis. GO annotations include signaling receptor binding and collagen binding.
ITGA5	1.25	Encodes a fibronectin and fibrinogen receptor. May promote tumor invasion, and higher expression of this gene may be correlated with shorter survival time in lung cancer patients. GO annotatiosn include integrin binding and epidermal growth factor receptor binding.
VEGFA	1.41	Encodes a growth factor active in angiogenesis, vasculogenesis and endothelial cell growth. It is upregulated in many known tumors and its expression is correlated with tumor stage and progression. GO annotations include protein homodimerization activity and protein heterodimerization activity.

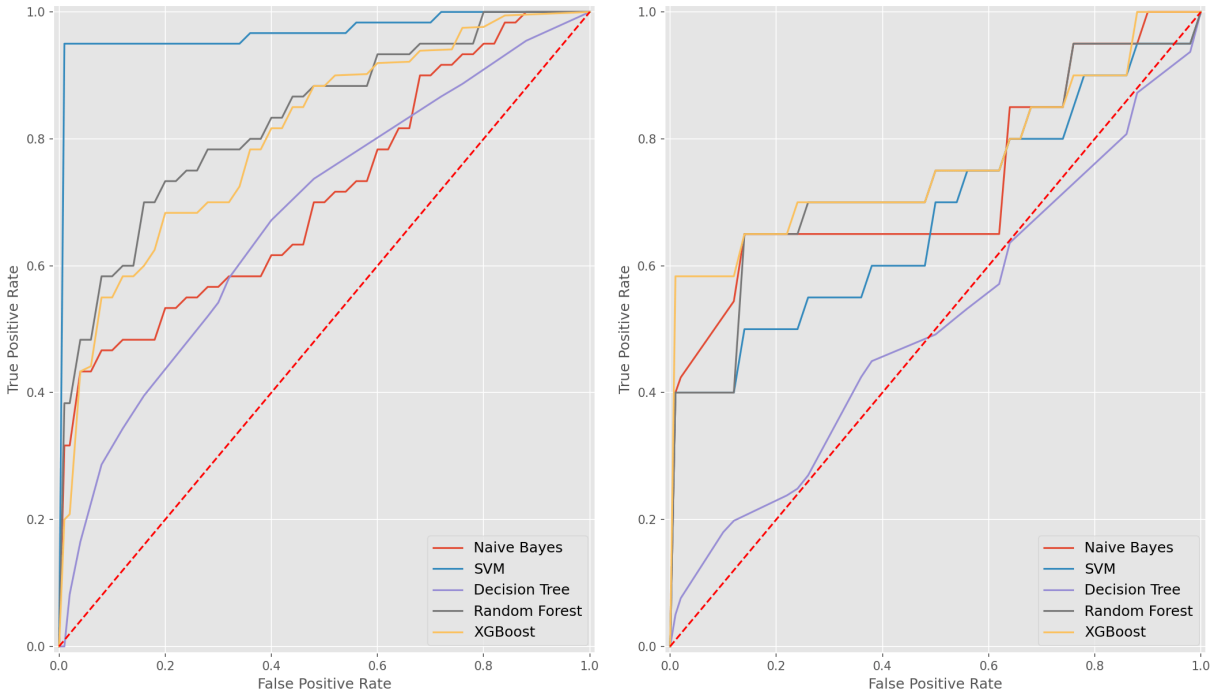

Figure A.1: ROC curves for configuration one (left) and two (right)

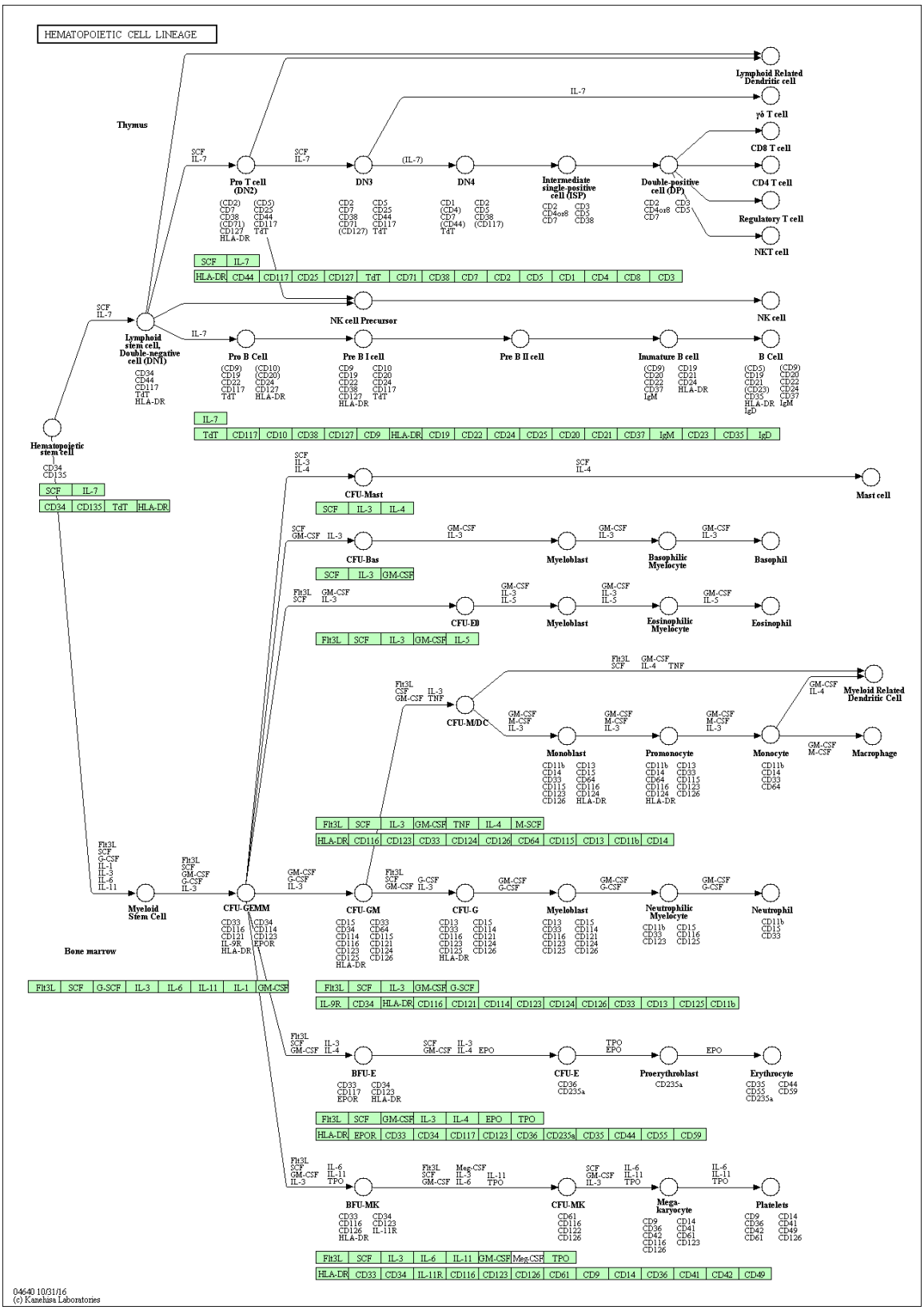

Figure A.2: Hematopoietic cell lineage pathway provided by KEGG PATHWAY Database